SIAM J. APPL. MATH.
Vol. 63, No. 1, pp. 1–41

ESTIMATE OF TURBULENT EDDY DIFFUSION BY EXACT RENORMALIZATION*

ALEXANDRA INDEIKINA† AND HSUEH-CHIA CHANG†

Abstract. By using a Lagrangian renormalization formulation, the effective diffusion equation is rigorously derived for a tracer in a homogeneous, isotropic, stationary, multidimensional and zero-mean Gaussian velocity field with a known two-point/two-time correlation tensor. The basic idea is to find the appropriate representation for the averaged small-scale solute distribution, to express it in terms of large-scale variables, and then to evaluate the limit of the infinite separation between the dissipation and the integral scales of turbulence. Key to the derivation is the validation of Corrsin's independence hypothesis for the selected velocity field at any diffusion time. The requirement of the nontrivial limiting behavior for the averaged solute distribution then results in the determination of the appropriate time scale for the averaged effective large-scale long-time spreading problem and in the evaluation of the effective transport coefficient. Unlike the simple shear flow case of Avellaneda and Majda, multidimensional velocity fluctuations ensure a constant eddy diffusivity in the limit of infinite time for any spectral parameters. By adjusting a single decorrelation time spectral parameter for a velocity field with Kolmogorov spectrum, the effective evolution equation is shown to produce the same time-evolution of the lateral mean-square displacement as a numerical simulation of planar flow and experimental heat-transfer data in turbulent pipe flow. The predicted constant asymptotic eddy diffusivity at infinite time, $E_D = 7.26 \times 10^{-3} \nu Re^{7/8}$, agrees with experimental data for eddy diffusivities in pipes and ducts over three decades of Reynolds numbers.

Key words. turbulence, mixing, eddy diffusion, renormalization methods, stochastic analysis

AMS subject classifications. 76F05, 76F25, 76F30, 60H30

PII. S0036139900379596

1. Introduction.

It is well known that the effective diffusivity of a turbulent flow can be a thousand times larger than the molecular transport coefficient. Hence, the computation of eddy diffusivity in a fully developed turbulence is an extremely important practical problem in both engineering and environmental science. Macroscopic description of the dispersion enhancement is a complex problem due to large fluctuations in the scalar field caused by the turbulent flow, where the velocity involves a continuous range of excited space and/or time scales and admits only a statistical description. The goal of eddy diffusivity theories is to assess the effects of the continuum of energetic smaller scales on the large scales through an effective equation without resolving this effect explicitly. Mathematically, it requires the statistical averaging of the small-scale fluctuations and then the determination of new appropriate time scales for the resulting effective transport equation such that the desired solution behaves nontrivially in the large-scale long-time limit.

Unfortunately, because of the presence of convective term $\mathbf{u} \cdot \boldsymbol{\nabla}(\dots)$ in transport equations, all statistical quantities are coupled up to infinite order. To find the average value $\langle C \rangle$, for example, one needs to know the correlation $\langle Cu \rangle$; if one tries to write down the equation for $\langle Cu \rangle$, terms like $\langle Cuu \rangle$ will appear in it, and so on. Hence, one needs to close somehow this infinite hierarchy of statistical moments. The fluctuations

*Received by the editors October 16, 2000; accepted for publication February 18, 2002; published electronically August 5, 2002.

http://www.siam.org/journals/siap/63-1/37959.html

†Department of Chemical Engineering, University of Notre Dame, Notre Dame, IN 46556-5637 (aindeiki@nd.edu, chang.2@nd.edu, http://www.nd.edu/~changlab/). The first author was supported by the Center for Applied Mathematics Fellowship. The second author was supported by the Bayer Chair Fund.

in turbulent flows are typically large, however, and the usual perturbation expansions, which simply neglect higher-order moments, will completely fail to predict something reasonable for problems of such kind. Instead, these problems have been attacked through a wide variety of renormalized perturbation theories that mimic ideas from field theory and the renormalization group theory from critical phenomena, both involving partial summation of the perturbation series.

The basic idea of the renormalized perturbation theories (RPT) is in the replacement, after ensemble averaging, of the zero-order terms in a formal widely divergent perturbation expansion by the exact values. It should be mentioned that convergence is not ensured for the renormalized perturbation series. However, even if it does not converge, it is much more accurate than widely divergent primitive perturbation series. The application of renormalized perturbation theories to turbulence has been pioneered by Kraichnan in a series of papers during the late 1950s, cumulating in the direct interaction approximation (DIA). (See Kraichnan [16], for example.) The "direct interaction principle" means that the strongest coupling occurs between the "nearest neighbors" in the wavenumber space. Consequently, only terms responsible for such interactions should be taken into account, summed, and averaged. As a result, the renormalized perturbation series is truncated at the second order.

Note that in its original formulation, DIA fails even to reproduce the Kolmogorov "$k^{-5/3}$" energy spectrum and results instead in a $k^{-3/2}$ decay for the inertial range of wavenumbers. Further development of this theory that includes reformulation in mixed Eulerian Lagrangian coordinates (see Kraichnan [17]) results in at least qualitative agreement with spectral measurements. Unfortunately, as the performance improves, the length and complexity of the final equations grow dramatically. Until now, different modifications of DIA remain the most popular among all renormalized perturbation theories. As an example of the application of DIA to scalar transport problems, one can mention the study of Koch and Brady [14]. They use DIA to predict the rate of growth of the variance of a tracer for a slowly decaying velocity covariance $\sim x^{-\gamma}$. Their analysis indicates that the spread will be nondiffusive with the mean-square displacement growing like $t^{4/(2+\gamma)}$ as $t \to \infty$ for $0 < \gamma \leq 2$, and this result is qualitatively consistent with numerical simulation. However, higher-order moments obtained from this approximation are incorrect.

In general, the strength of the renormalized perturbation theories lies in their generality and the absence of ad hoc assumptions or disposable constants. However, the unpredictable error of all RPT's (because of unknown mathematical properties of renormalized perturbation series) on one hand, and the enormous complexity when formulated for inhomogeneous turbulence on the other, restrict using these theories in both fundamental and engineering applications. A partial answer to the first of these problems lies within the renormalization group approach, which is quite distinct from RPT.

The renormalization group method (RNG) already has some success when applied to problems in critical phenomena. The pioneers in the development of the RNG for turbulence are Forster, Nelson, and Stephen [9]. Their theory is later generalized by De Dominicis and Martin [6]. The RNG involves iterative averaging over the small bands of modes and progressive scaling away from the highest wavenumbers (short waves), whose effect on the lowest wavenumbers (long waves) can be retained in an average form as a contribution to the transport coefficients. If the system becomes invariant under the mode elimination procedure, the scaling transformation is said to have reached a fixed point. One can then eliminate all fluctuations, and

the value of the new transport coefficient at the fixed point corresponds to enhanced diffusivity. Using the RNG, Yakhot and Orszag [23] obtain the turbulent viscosity and the relation between turbulent Prandtl number and turbulent viscosity for unbounded homogeneous turbulence and applied these results to heat transfer in a pipe, using an empirical model for the viscosity in a wall region (Yakhot, Orszag, and Yakhot [24]). It has been reported that the proposed formula gives good agreement with experimental data in a wide range of Prandtl number, $10^{-2} < Pr < 10^{6}$.

From both a physical and mathematical point of view, the RNG is more rigorous than the RPT, and at the same time, the RNG is simpler: at the end, the mathematical problem reduces to the study of a system of first-order ordinary differential equations. However, it should be mentioned that the RNG mode elimination procedure is valid to any order only if the neglected higher-order terms (they are truly negligible only when the first band of modes is eliminated) remain negligible after scaling transformations. In turbulent flow calculations, the above condition often requires the space dimension to be "slightly higher than 3," with the best performance, for example, in seven-dimensional space (Yakhot and Orszag [23]). Consequently, it is hard to expect the RNG to give the correct prediction for any case, Yakhot, Orszag, and Yakhot's claim notwithstanding.

For scalar transport problems, perhaps the most rigorous and accurate approach has been suggested by Avellaneda and Majda [1]. They consider an advection-diffusion of a passive scalar in a simple shear flow $\mathbf{u} = (0, u_2(x_1, t), 0)$, where $u_2(x_1, t)$ has stationary Gaussian statistics. This model problem admits an explicit representation of the solution through the Lagrangian formulation. Using the above representation and some properties of stochastic differential equations, Avellaneda and Majda develop a complete renormalization theory for this exactly solvable model with full mathematical rigor. They found several regimes of anomalous diffusion that depend on the parameters of the velocity spectrum, such as time-dependent diffusivities and even an effective equation with a random nonlocal diffusion coefficient.

With this exact result, the capability of the RNG, Lagrangian RPT, and DIA in predicting turbulent transport for the same model flow is also examined (Avellaneda and Majda [2]). It is found that all these approximate theories give incorrect predictions for some regions of renormalization, which depend on the parameters defining the velocity spectrum.

The RNG always predicts a simple local diffusion equation with constant diffusivity and often erroneously determines the appropriate time scale for the effective transport problem. More importantly, it has been found that the RNG is barely acceptable for velocity spectra pertinent to turbulent transport problems. The important example of the Kolmogorov velocity spectrum belongs to the boundary of applicability of the RNG, where this method can still give the correct long time scale for the effective diffusion problem. However, in this case RNG can only provide the infinite-time asymptotic value of effective diffusivity and cannot resolve the time-evolution of the mean-square displacement. Hence, the application of the results of Yakhot and Orszag [23] to nonsteady turbulent transport problems seems questionable.

In contrast, both RPT always reproduce the correct time scale for effective diffusion problems. However, the resulting effective equations can vary from the simple wave equation to some complicated integro-differential equation instead of the local diffusion equation with time-dependent diffusivity predicted by the exact renormalization theory of Avellaneda and Majda [1].

For the more general case of turbulent transport by isotropic homogeneous ran-

dom velocity fields, Avellaneda and Majda [3] investigate a part of the spectral parameter space in the vicinity of the Kolmogorov value. They find that the anomalous long-time scale for the effective transport process remains the same as for a simple shear flow model considered earlier [1], regardless of flow dimensions. The governing equation may, however, differ. If the temporal fluctuations are not irrelevant in the large-scale long-time limit, Avellaneda and Majda suggest a nonlocal diffusion equation but do not present details in their report [3].

Recently, Fannjiang [8] invokes variational principles to analyze scalar transport by the same three-dimensional turbulent flow for the entire range of spectral parameters. He also investigates the effect of the cut-off wavenumber on the resulting scaling laws. (In fact, this introduces a third dimension in the parameter space.) While this method produces a long time scale for anomalous diffusion, consistent with earlier results of Avellaneda and Majda [3] at identical values of the parameters, it provides only an upper bound for the effective diffusion coefficient. The limiting long time scale is found to be dependent on the wavenumber scaling. However, it should be noted that the actual wavenumber range, important for the scalar transport, is not arbitrary. Instead, it should be determined from physical arguments (as has been done by Avellaneda and Majda [1]) like the velocity spectrum, spatial scales of initial data, requirement of finite (not only bounded) eddy diffusivity, etc. Hence, the applicability of Fannjiang's results [8] to practical turbulent transport is quite limited.

It also should be noted that none of the earlier works relate the main scaling parameter (the ratio of dissipation to integral length scales of turbulence) to other typically reported physical quantities (like Reynolds number, friction velocity, pipe diameter, etc.). For the Kolmogorov spectrum, this relation is well known (see McComb [19], for example, or any other textbook on turbulence), but it is not obvious for other spectra. Also, there are almost no comparisons of suggested theories with experimental data. The only exceptions are the work of Yakhot, Orszag, and Yakhot [24] and in the fundamental book of McComb [19], where some of the predicted infinite-time asymptotic values of effective diffusivities are compared to data. If, however, the theory suggests time-dependent eddy diffusivity or nonlocal scalar transport, one needs to use data on the time-evolution of mean-square displacement to check the prediction adequately.

The investigation of full two-dimensional (2-D) and three-dimensional (3-D) turbulent transport using an extension of exact renormalization theory of Avellaneda and Majda is the subject of this report. With the usual idealization for the turbulent core, the zero-mean unbounded turbulent flow in the inertial range of length scales is assumed to be stationary, homogeneous, and isotropic. The spreading of the pulse of the solute in 2-D or 3-D isotropic random flow is examined as the simplest example, allowing the derivation of large-scale long-time effective transport equation and the associated effective diffusivity, which is time-dependent in general, in terms of parameters defining the velocity statistics.

The basic idea is to find the appropriate representation for the averaged small-scale solute distribution, to express it in terms of large-scale variables, and then to evaluate the limit of the infinite separation between the dissipation and the integral scales of turbulence. (In fact, this is the limit of infinite Reynolds number). The requirement of the nontrivial limiting behavior for the averaged solute distribution (it is never equal to zero nor the initial data) then results in the determination of the appropriate time scale for the averaged effective large-scale long-time spreading problem and in the evaluation of the effective transport coefficient.

Since the whole idea of renormalization in this case is based on the appropriate rescaling, the initial nondimensionalization of the small-scale transport problem and the statistics of the random velocity field are defined precisely in sections 2 and 3.

Section 4 describes the Lagrangian formulation of the advection-diffusion problem through Ito's stochastic differential equation. It provides the solution for the distribution of the evolved concentration field that, however, has to be averaged over the distribution of random velocity field and over the Brownian motion that represents the molecular diffusion effect in Ito's formulation.

This averaging is completed in section 5, and it is shown that the resulting averaged concentration distribution and the evolution equation for the effective diffusivity are exact for the selected stationary homogeneous isotropic Gaussian velocity field. With some physically plausible assumptions, the above result also can be applied to the more important case of non-Gaussian velocity statistics.

In section 6 the rescaling of the derived equations is carried out. The large-scale long-time limiting behavior of the averaged solute distribution is evaluated and the effective diffusion equation is derived, provided that the effective diffusivity can be properly renormalized.

The renormalization procedure for different parameters of the velocity spectrum is described in section 7. It includes the evaluation of the distinguishing limit for the evolution of rescaled enhanced diffusivity and hence the determination of the appropriate time-rescaling function for the effective large-scale long-time diffusion equation.

The summary and discussion of results of the renormalization analysis are offered in section 8. The Kolmogorov velocity spectrum is chosen to compare the predicted transport coefficient with experimental data on turbulent scalar transport and with earlier numerical simulation. Both quantitative and qualitative agreement seem to be satisfactory, taking into account the strong limitation of the assumed homogeneous and isotropic Gaussian statistics.

2. Initial nondimensionalization for the transport problem. Consider the dispersion of a solute/temperature field by a well-developed turbulent flow. Away from the boundaries, in a turbulent core, the mean velocity profile is often assumed to be flat. Consequently, such mean flow results only in the translation of the initial solute distribution. Hence, in a frame of reference moving with the constant mean flow, the spreading of the pulse is governed by the advection-diffusion equation

$$\frac{\partial C'}{\partial t'} + \mathbf{u}' \cdot \boldsymbol{\nabla}' C' = D_0' \boldsymbol{\nabla}'^2 C', \qquad C'(\mathbf{x}', 0) = C_0'(\mathbf{x}'), \tag{2.1}$$

where $\mathbf{u}'(\mathbf{x}', t')$ is the random velocity field with zero mean, D_0' is the molecular diffusivity and primes denote dimensional variables. Because only the turbulent core will be considered here, no boundary conditions are imposed. Since the goal of this study is to average the effect of all inertial range fluctuations of the velocity field, the problem should be made nondimensional with the smallest possible scale of turbulence initially. Hence, the dissipation length scale L_d for the fluctuation velocity field $\mathbf{u}'$ is taken as the characteristic length. The characteristic velocity U_d is then defined by the requirement that the dissipation Reynolds number $R_d = U_d L_d/\nu$ is equal to 1, such that $U_d = \nu/L_d$. This results in the characteristic time $\tau_d = L_d/U_d = {L_d}^2/\nu$, which corresponds to the viscous dissipation time scale. The inverse of the Schmidt number defines the nondimensional diffusivity $D_0 = D_0'/\nu = Sc^{-1}$.

For the renormalization procedure, it also is necessary to define the macroscopic scales for the problem. The macroscopic length scale L_0 is defined as the integral length scale of the turbulence. For the realistic systems, L_0 is of the order of the lateral dimension of the flow. Since the effect of constant mean flow in this idealized problem has been eliminated and only the inertial range of the turbulence will be considered, the large-scale Reynolds number $R_0 = U_0 L_0/\nu$ is defined through the root-mean-square velocity U_0 for the large-eddy motion.

The ratio between the dissipation and the integral length scales increases with R_0 and it is represented by the small parameter $\delta = L_d/L_0 \ll 1$. The actual dependence of δ on the large-scale Reynolds number R_0 depends on the spectrum of the velocity fluctuation and will be specified later.

The advection-diffusion problem (2.1) is linear in C' and, because there are no boundary conditions, the characteristic value of the concentration difference $\triangle C'$ used for nondimensionalization is not important. However, several restrictions should be imposed on the initial nondimensional solute distribution $C_0 = C_0'/\triangle C'$. Since (2.1) describes the solute dispersion in the unbounded domain, $C_0(\mathbf{x}')$ should decay to infinity fast enough such that the integral of the initial distribution over the entire n-dimensional space is finite,

$$0 < {L'}^{-n} \int_{-\infty}^{\infty} C_0(\mathbf{x}') d^n\mathbf{x}' < \infty,$$

where an appropriate length scale L' is added for dimensional reasons. Consequently, $C_0(\mathbf{x}')$ can be defined by its Fourier integral

$$C_0(\mathbf{x}') = (2\pi/L')^{-n} \int_{-\infty}^{\infty} \hat{C}_0(\mathbf{K}') \exp(\imath \mathbf{K}'\mathbf{x}') d^n\mathbf{K}', \tag{2.2}$$

where $\mathbf{K}'$ is the n-dimensional wavevector.

If one would like to describe a large-scale long-time behavior of the spreading process, it is necessary to assume large-scale initial data for the advection-diffusion problem (2.1). This implies that $C_0(\mathbf{x}')$ varies only over the integral length scale L_0. Consequently, one should set $L' = L_0$ in (2.2) such that $C_0(\mathbf{x}') = C_0(\mathbf{x}'/L_0)$ and $\hat{C}_0(\mathbf{K}') = \hat{C}_0(\mathbf{K})$, where $\mathbf{K} = \mathbf{K}' L_0$ is the large-scale nondimensional wavevector.

Hence, with the nondimensionalization on dissipation length and time scales L_d and τ_d, the governing equation (2.1) and the initial distribution (2.2) becomes

$$\frac{\partial C}{\partial t} + \mathbf{u} \cdot \boldsymbol{\nabla} C = D_0 \boldsymbol{\nabla}^2 C, \tag{2.3}$$

$$C(\mathbf{x}, 0) = C_0(\delta \mathbf{x}) = (2\pi)^{-n} \int_{-\infty}^{\infty} \hat{C}_0(\mathbf{K}) \exp(\imath \delta \mathbf{K}\mathbf{x}) d^n\mathbf{K}, \tag{2.4}$$

and all nondimensional variables and parameters introduced in this section are summarized below:

$$\begin{array}{llll} \mathbf{x} = \mathbf{x}'/L_d, & \mathbf{u} = \mathbf{u}'/U_d, & \mathbf{K} = \mathbf{K}' L_0, & t = t'/\tau_d, \\ \tau_d = L_d/U_d = {L_d}^2/\nu, & D_0 = D_0'/\nu = Sc^{-1}, & & \\ R_d = U_d L_d/\nu = 1, & R_0 = U_0 L_0/\nu \gg 1, & \delta = L_d/L_0 \ll 1. & \end{array}$$

3. Velocity statistics. Let us specify the random velocity field $\mathbf{u}(\mathbf{x},t)$. The fluid is assumed to be incompressible, such that the continuity equation $\boldsymbol{\nabla}\cdot\mathbf{u}'=0$ holds. With the usual idealization for the turbulent core, the zero-mean unbounded turbulent flow in the inertial range of length scales ($L_d \ll L' \ll L_0$ or, in nondimensional form, $1 \ll L' \ll 1/\delta$) is assumed to be stationary, homogeneous, and isotropic.

It also is assumed that the random flow field $\mathbf{u}(\mathbf{x},t)$ has Gaussian statistics. In general, this is a very strong and, moreover, nonphysical assumption for the turbulent flow. It is well known that real turbulence is never Gaussian. More importantly, the nonzero triple correlation $\langle u_i(\mathbf{x},t)u_j(\mathbf{x},t)u_k(\mathbf{y},t)\rangle$ is responsible for turbulent energy transfer. At the same time, the Gaussian statistics for zero-mean $\mathbf{u}(\mathbf{x},t)$ immediately leads to the vanishing of all odd-order moments. For the problem of a scalar transport, however, this is a rather common approximation (see Kimura and Kraichnan [13], Koch and Shaqfeh [15], or Avellaneda and Majda [2], for example) if one is not interested in how the random flow field has been created and sustained.

With the above assumptions, the m-dimensional Gaussian velocity field $\mathbf{u}(\mathbf{x},t)$ is specified by the spectral form of two-point two-time correlation tensor $\mathbf{R}$,

$$\begin{aligned}\langle u_i(\mathbf{x},t)u_j(\mathbf{y},s)\rangle &= R_{ij}(|\mathbf{x}-\mathbf{y}|,|t-s|)\\ &= (2\pi)^{-m}\int_{-\infty}^{\infty}\hat{Q}(k,|t-s|)\hat{P}_{ij}(\mathbf{k})\exp(\imath\mathbf{k}(\mathbf{x}-\mathbf{y}))d^m\mathbf{k},\end{aligned}\tag{3.1}$$

where $k=|\mathbf{k}|$, $\hat{P}_{ij}(\mathbf{k})=\delta_{ij}-k_ik_j/k^2$, and the spectral amplitude $\hat{Q}(k,|t-s|)$ is defined in the inertial range of wavenumbers

$$\begin{aligned}&\hat{Q}(k,|t-s|)=\alpha^2k^{1-m-\epsilon}\exp(-ak^z|t-s|), && \delta\le k\le 1,\\ &\hat{Q}(k,|t-s|)\equiv 0 && \text{otherwise.}\end{aligned}\tag{3.2}$$

One should mention that the correlation of the Fourier-component of the velocity field is given by

$$\begin{aligned}\langle\hat{u}_i(\mathbf{k},t)\hat{u}_j(\mathbf{q},s)\rangle &= \delta(\mathbf{k}+\mathbf{q})\hat{R}_{ij}(\mathbf{k},|t-s|)\\ &= \delta(\mathbf{k}+\mathbf{q})\hat{Q}(k,|t-s|)\hat{P}_{ij}(\mathbf{k}).\end{aligned}\tag{3.3}$$

Note that for the simple shear flow considered by Avellaneda and Majda [1] ($m=1$, $\mathbf{u}(\mathbf{x},t)=(0,u_2(x_1,t),0)$, $\mathbf{k}=(k_1,0,0)$) the spectral correlation tensor $\hat{\mathbf{R}}(\mathbf{k},|t-s|)$ reduces to a single component $\hat{R}_{22}=\hat{Q}(k,|t-s|)$ and in this case the requirement of the isotropy of the velocity field should be omitted.

For the realistic turbulent core, the spatial dimensions m and n for the velocity and concentration field should always be set equal to 3 because fluctuations are always 3-D. However, the general form of the isotropic spectrum (3.2) and the initial conditions (2.4) with $m=n=2$ or 3 will be considered, keeping in mind other possible applications and the comparison with earlier works.

If the experimental data on the two-point two-time correlation are available, all parameters in the model spectrum, as well as the value of the scale ratio δ, can be determined directly. However, most velocity measurements provide only the energy spectrum. Hence, it is useful to discuss the physical meaning of the parameters of the model spectrum (3.2) and to establish some relations between values of these parameters and more common measurable quantities, like the Reynolds number, for example. The parameters ϵ and α are related to the kinetic energy spectrum (per

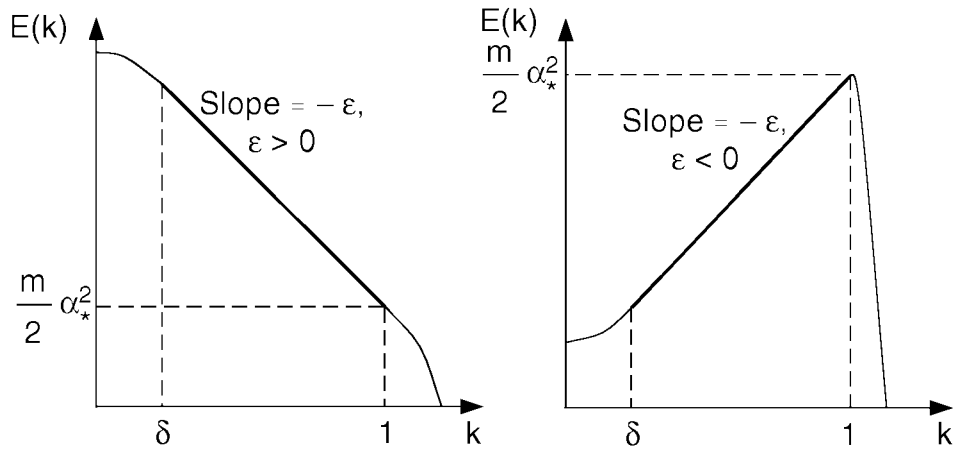


FIG. 1. *Schematic of the model energy spectrum (log-log plot).* (a) *For positive ϵ, the energy is supplied by the large-eddy motion, with $k < \delta$, and is dissipated in small-scale fluctuations $k > 1$, with the energy cascade along the spectrum.* (b) *In the case of negative ϵ, the spectrum should necessarily have a peak, where some additional stochastic forcing is applied.*

unit mass of liquid)

$$\hat{E}(k)dk = \frac{1}{2}\mathrm{tr}\hat{\mathbf{R}}(\mathbf{k},0)k^{m-1}\frac{S^{(m)}}{(2\pi)^m}dk, \tag{3.4}$$

where $S^{(m)} = 2\pi^{m/2}/\Gamma(m/2)$ is the area of the m-dimensional unit sphere. For 2-D or 3-D flow, the model energy spectrum (3.4) becomes

$$\hat{E}(k)dk = \frac{\alpha^2(m-1)}{4\pi^{m-1}}k^{-\epsilon}dk, \qquad \delta \le k \le 1, \quad m = 2, 3. \tag{3.5}$$

The energy spectrum (3.5) is shown schematically in Figure 1(a) for positive values of the exponent ϵ. Hence, ϵ defines the strength of the infra-red divergence of the kinetic energy at low wavenumbers, and α is the nondimensional amplitude parameters which are assumed to be independent on the separation of scales δ. The values of $\epsilon = 5/3$, $m = 3$, and $\alpha^2(m-1)/(4\pi^{m-1}) = \alpha_0$ in (3.5) provide the nondimensional version for the Kolmogorov energy spectrum

$$\hat{E}'_{(K)}(k')dk' = \alpha_0\varepsilon^{2/3}k'^{-5/3}dk', \tag{3.6}$$

where $\alpha_0 \approx 1.5$ is the Kolmogorov constant and ε is the energy dissipation rate.

With the nondimensionalization on the dissipation length scale, the kinetic energy of the small-scale fluctuations ($k, dk \sim O(1)$) should be of the order of $O(1)$. Consequently, the amplitude for the energy spectrum $\alpha^2(m-1)/(4\pi^{m-1}) \sim O(1)$. The universality of the spectrum in the inertial range of wavenumbers implies that (3.5) should remain invariant under the scaling transformations ($L_d \leftrightarrow L_0, U_d \leftrightarrow U_0$). Consequently, in the large-scale limit ($k, dk \sim O(\delta)$) expression (3.5) should provide the kinetic energy for the largest eddies included in consideration, without rescaling of the amplitude α. This requirement gives the relation between the integral and dissipation scales

$$\frac{{U_0}^2}{{U_d}^2} \sim \delta^{1-\epsilon} \tag{3.7}$$

or, in terms of the Reynolds number for the large-eddy motion,

$$\delta \sim {R_0}^{-2/(\epsilon+1)}. \tag{3.8}$$

With $\epsilon = 5/3$, (3.8) reduces to $\delta \sim {R_0}^{-3/4}$, i.e., provides the correct dependence for the scales ratio δ on the Reynolds number of the large-eddy motion for well-developed isotropic turbulence.

Several remarks should be made here. As is evident from (3.7), a physically meaningful spectrum for the flow with an energy cascade from the large to the small scales should have $\epsilon > 1$. It is unlikely for the speed of the large-eddy motion U_0 to be smaller than the root-mean-square velocity of the smallest scales U_d, unless some additional small-scale stochastic forcing is applied. In such a case, when the system presents both "slow" large eddies and "fast" small fluctuations, (3.8) provides another restriction: $\epsilon > -1$ because the separation of scales δ should increase with R_0. For negative values of ϵ, the spectrum should necessarily have a peaked shape, like in Figure 1(b), and the functional form of the decaying part of the spectrum may be important.

The integration over k in (3.5) results in the total kinetic energy of the inertial-range fluctuations

$$E = \int_0^\infty \hat{E}(k)dk = \frac{1}{2}\mathrm{tr}\mathbf{R}(0,0) = \frac{m}{2}\frac{U_{rms}^2}{U_d^2}$$

and provides the relation for the parameters of the model spectrum with the root-mean-square velocity U_{rms},

$$\frac{U_{rms}^2}{U_d^2} = \frac{\alpha^2(m-1)}{2m\pi^{m-1}} \begin{cases} \frac{1}{\epsilon-1}(\delta^{1-\epsilon}-1) & \epsilon \neq 1, \\ -\ln\delta & \epsilon = 1, \end{cases} \tag{3.9}$$

or, in terms of the root-mean-square Reynolds number $R_{rms} = U_{rms}L_0/\nu$,

$$R_{rms}^2 = \frac{\alpha^2(m-1)}{2m\pi^{m-1}} \begin{cases} \frac{1}{\epsilon-1}(\delta^{-(1+\epsilon)}-\delta^{-2}) & \epsilon \neq 1, \\ -\delta^{-2}\ln\delta & \epsilon = 1. \end{cases} \tag{3.10}$$

In the limit $\delta \to 0$, (3.8) and/or (3.10) allow us to relate the amplitude parameter α in the model and the ratio of scales δ with R_{rms} and R_0 for ϵ different from Kolmogorov's 5/3 value.

Now let us consider the time-dependent part of the two-point two-time correlation (3.2). The decorrelation time $T = 1/ak^z$ corresponds to the turnover time for different spatial modes. The parameter $z > 0$ represents the fact that low-wavenumber modes have larger turnover times than the short waves. Larger values of z increase the separation of the decorrelation times for long and short waves. The nonnegative parameter a defines the decorrelation time $1/a$ for the shortest waves with $k = 1$. The value of $a = 0$ corresponds to the spatially random steady flow $\mathbf{u} = \mathbf{u}(\mathbf{x})$. The limiting value $a \to \infty$ (provided that α^2 scales as a) corresponds to fluctuations that are completely decorrelated in time. In such a case, the exponential term should be replaced by a delta-function $\delta(k^z(t-s)) = k^{-z}\delta(t-s)$. Similar to the amplitude of the spectrum α, a is assumed to be independent of the scale ratio δ.

The physical association of T with the turnover time for different modes $\hat{\mathbf{u}}(\mathbf{k})$ implies that $1/T \sim ku(k)$. Since the speed of the mode $u(k)$ can be estimated as a

root-mean-square fluctuation velocity in the range of wavenumbers $q > k$,

$$u(k) \sim \left(\int_k^\infty \hat{E}(q) dq \right)^{1/2},$$

the kinetic energy spectrum (3.5) results in the estimate for $1/T$,

$$1/T = ak^z \sim \alpha k^{(3-\epsilon)/2}, \tag{3.11}$$

and, consequently,

$$a \sim \alpha, \qquad z = \frac{3-\epsilon}{2}. \tag{3.12}$$

The above estimate prescribes the particular relationship (3.12) between z and ϵ. Note that, according to the Kolmogorov similarity hypothesis, in the inertial range of wavenumbers, T can depend only on the energy dissipation rate ε and the wavenumber itself. Dimensional analysis then leads to the well-known result for the frequency-response function $\omega'(k') \sim 1/T'$ and turbulent viscosity $\nu_T(k')$:

$$\omega'(k') = \nu_T(k') k'^2 \sim 1/T' \sim \varepsilon^{1/3} k'^{2/3}. \tag{3.13}$$

As can be easily seen, for the Kolmogorov value of the exponent $\epsilon = 5/3$, (3.11) is just a nondimensional version of (3.13).

Measurement of the decorrelation time T from two-time correlations are rare. Hence, (3.12) provides a welcomed estimate of z from ϵ. However, the spectral parameter a remains unknown as $a \sim \alpha$ is only an order estimate. We shall determine a parameter a empirically in section 8.

Hence, the Kolmogorov energy spectrum (3.6) corresponds to $\epsilon = 5/3$, $z = 2/3$, $m = 3$, $\alpha^2(m-1)/(4\pi^{m-1}) = \alpha_0 \approx 1.5$, and $a \sim O(1)$ in the model spectrum (3.2). However, regardless of solute dispersion by a well-developed turbulent flow, different "velocity" statistics are possible. In a subsequent analysis, the general form of the two-point two-time correlation function (3.1)–(3.2) for the isotropic stationary Gaussian velocity field $\mathbf{u}(\mathbf{x}, t)$ will be used. The possible values of parameters in (3.2) are $m = \{2; 3\}$, $-1 < \epsilon < 3$, $z \geq 0$, $a \geq 0$, $\alpha \sim O(1)$.

4. Lagrangian formulation for statistical advection-diffusion problem. The advection-diffusion problem (2.3)–(2.4) for the concentration field $C(\mathbf{x}, t)$ is essentially the Fokker–Planck equation for the probability density function to find a fluid particle in a particular location $\mathbf{x}$ at time t for given initial condition $C(\mathbf{x}, 0) = C_0(\delta\mathbf{x})$. The equivalent representation for a Fokker–Planck equation is an Ito stochastic differential equation for the trajectory of the fluid particle

$$\begin{aligned} &d\mathbf{X}(s) = \mathbf{u}(\mathbf{X}(s), s)ds + \sqrt{2D_0} d\mathbf{W}(s), \\ &\mathbf{X}(s=0) = \mathbf{x}, \end{aligned} \tag{4.1}$$

where the Gaussian white noise $\mathbf{W}(s)$ represents the displacement of the fluid particle due to molecular diffusion. Hence, equation (4.1) is, in fact, the Lagrangian formulation of the advection-diffusion problem (2.3)–(2.4), where the molecular diffusion is replaced by a random walk $\sqrt{2D_0}\mathbf{W}(s)$ of a fluid particle. Note that the Lagrangian velocity $\mathbf{u}(\mathbf{X}(s), s)$ depends on $\mathbf{W}(s)$ while the realization of the Eulerian velocity field $\mathbf{u}(\mathbf{x}, s)$ is, of course, independent of molecular diffusion. The goal

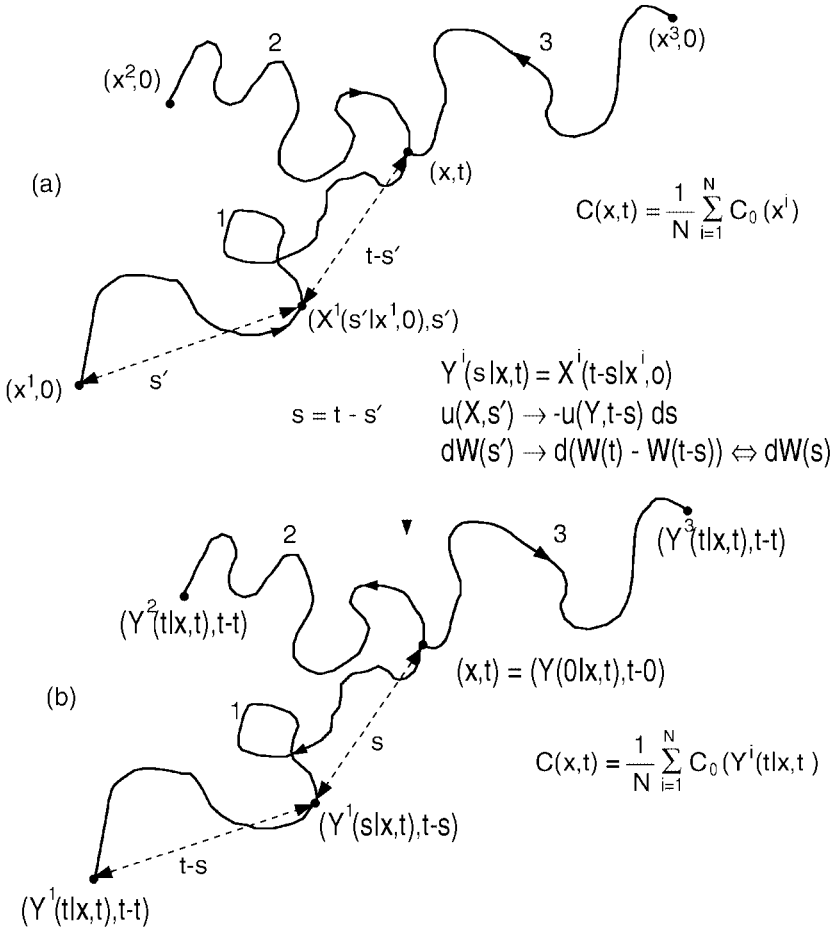


FIG. 2. *Schematic of the Lagrangian formulation for the advection-diffusion problem.* (a) *The concentration in point x at time t is determined by the initial solute distribution, averaged over all possible initial points $x^{(i)}$ of fluid particles trajectories $X^{(i)}(t) = x$.* (b) *Transition to the inverse problem allows us to find the distribution of the initial points $(x^{(i)}, 0)$ with respect to the final state (x, t).*

now is to find the representation for the average concentration distribution for any $\mathbf{x}$ and t. For each fixed realization (i) of the "composed" random "velocity field" $\mathbf{V} = \mathbf{u} + \sqrt{2D_0}\dot{\mathbf{W}}$ (of course, $\dot{\mathbf{W}}$ is only a formal notation because the trajectories of the Brownian motion are nondifferentiable), there exists a particular trajectory $\mathbf{X}^{(i)}(s|\mathbf{x}^{(i)}, 0)$ which begins at some initial point $\mathbf{x}^{(i)}$ at $s = 0$ and arrives at $\mathbf{x}$ at time $s = t$, such that $\mathbf{x} = \mathbf{X}^{(i)}(s|\mathbf{x}^{(i)}, 0)$, as shown schematically in Figure 2(a). Since in such formulation (2.3)–(2.4) is reduced to a pure convection problem, the concentration does not change along the trajectory $\mathbf{X}^{(i)}(s|\mathbf{x}^{(i)}, 0)$ and is equal to $C_0(\delta\mathbf{x}^{(i)})$. The corresponding realization of the random concentration field is then

$$C^{(i)}(\mathbf{x}, t) = C(\mathbf{X}^{(i)}(s|\mathbf{x}^{(i)}, 0)) = C_0(\delta\mathbf{x}^{(i)}),$$

and the averaging over all possible trajectories gives the solute distribution

$$\overline{C(\mathbf{x},t)} = \frac{1}{N}\lim_{N\to\infty}\sum_{i=1}^{N} C_0(\delta\mathbf{x}^{(i)}). \tag{4.2}$$

In order to compute (4.2) one needs to determine the evolution of the initial points $(\mathbf{x}^{(i)},0)$ with respect to the final state $(\mathbf{x},t) = (\mathbf{X}^{(i)}(t|\mathbf{x}^{(i)},0),t)$—in other words, to find the inverse trajectories $\mathbf{Y}^{(i)}(s) = \mathbf{Y}^{(i)}(s|\mathbf{x},t)$, which begin at $(\mathbf{x},t)$ at $s=0$ and come to $(\mathbf{x}^{(i)},0)$ at $s=t$. The schematic of the inverse problem is shown in Figure 2(b): for any fixed realization of $\mathbf{u}$ and $\mathbf{W}$ one can go on the trajectory $\mathbf{X}^{(i)}(s|\mathbf{x}^{(i)},0)$ in the opposite direction, such that the inverse trajectory $\mathbf{Y}^{(i)}(s|\mathbf{x},t) = \mathbf{X}^{(i)}(t-s|\mathbf{x}^{(i)},0)$, and, consequently,

$$\overline{C(\mathbf{x},t)} = \frac{1}{N}\lim_{N\to\infty}\sum_{i=1}^{N} C_0(\delta\mathbf{Y}^{(i)}(t|\mathbf{x},t)). \tag{4.3}$$

Hence, it is possible to write down the stochastic differential equation for the inverse problem:

$$\begin{aligned} &d\mathbf{Y}(s) = -\mathbf{u}(\mathbf{Y}(s),t-s)ds + \sqrt{2D_0}d\mathbf{W}(s),\\ &\mathbf{Y}(s=0) = \mathbf{x}. \end{aligned} \tag{4.4}$$

The solution for the average concentration distribution is then given by (4.3) and (2.4) with the replacement of the empirical averaging in (4.3) by the averaging over the distribution of trajectory $Y(t|\mathbf{x},t)$,

$$\begin{aligned} \overline{C(\mathbf{x},t)} &= (2\pi)^{-n}\int_{-\infty}^{\infty}\hat{C}_0(\mathbf{K})\langle\exp(\imath\delta\mathbf{K}\mathbf{Y}(t|\mathbf{x},t))\rangle_Y d^n\mathbf{K}\\ &= (2\pi)^{-n}\int_{-\infty}^{\infty}\hat{C}_0(\mathbf{K})\langle\langle\exp(\imath\delta\mathbf{K}\mathbf{Y}(t|\mathbf{x},t))\rangle\rangle_{u,W} d^n\mathbf{K}. \end{aligned} \tag{4.5}$$

Note that since $Y(s|\mathbf{x},t)$ is related to $\mathbf{u}$ and $\mathbf{W}$ by (4.4), the above averaging is equivalent to the averaging over independent distributions of the Eulerian velocity field $\mathbf{u}$ and $\mathbf{W}$, $\langle\dots\rangle_Y = \langle\langle\dots\rangle_{Y(u|W)}\rangle_W = \langle\langle\dots\rangle_{Y(W|u)}\rangle_u = \langle\langle\dots\rangle_u\rangle_W = \langle\langle\dots\rangle_W\rangle_u$.

The representation of the advection-diffusion problem (2.3)–(2.4) by (4.4)–(4.5) is exact regardless of whether the velocity field is stochastic or not. For example, in the case of time-periodic linear planar flow field ($\mathbf{u}_{\text{linear}} = \mathbf{A}\mathbf{x}\cos\omega t$), where the constant matrix $\mathbf{A}$ has zero trace), equation (4.4) describes a 2-D time-dependent Ornstein–Uhlenbeck process (see Gardiner [10], for example). The solution of (4.4) then becomes

$$\mathbf{Y}_{\text{linear}}(s=t) = \mathbf{B}(\mathbf{t})\mathbf{x} + \sqrt{(2/Pe)}\int_0^t \mathbf{B}(t-s)d\mathbf{W}(s),$$

where the matrix $\mathbf{B}(t) = \exp(-\mathbf{A}\sin\omega t/\omega)$. Since $\mathbf{Y}_{\text{linear}}$ is linear in $d\mathbf{W}$, and $\mathbf{B}(\mathbf{t})$ is nonstochastic, $\mathbf{Y}_{\text{linear}}$ is a Gaussian random process. The averaging over W in (4.5) then gives a solute distribution (equation (6) of Indeikina and Chang [12]) in terms of the Fourier transform.

5. Evaluation of the partially-averaged enhanced diffusivity tensor. Unfortunately, explicit solution of the stochastic differential equations can be easily done only for several special cases, mostly including linear equations or some simple explicit form of coefficients. In the present case, the random velocity field is specified by the spectral correlation function (3.1) and hence (4.4) is, in fact, a system of coupled stochastic integro-differential equations. Hence, one should make some possible simplifications.

Note that the "final" time t is only a parameter in (4.4) containing only $\mathbf{u}(\mathbf{Y}(s), t-s)$ and that $\mathbf{u}(\mathbf{Y}(0), t-0) = \mathbf{u}(\mathbf{x}, t)$. Hence, because of the stationarity of the random velocity field (everything depends only on $|(t-s)-(t-0)| = s$), the statistical properties of the random function $\mathbf{Y}(s|\mathbf{x}, t)$ also cannot be dependent on t. Consequently, the replacement of $\mathbf{u}(\mathbf{Y}(s), t-s)$ on $\mathbf{u}(\mathbf{Y}(s), s)$ in (4.4) changes nothing in the statistics of $\mathbf{Y}$. It can be easily shown that the absence of the mean flow and the requirement of the homogeneity of the velocity field leads to $\langle\langle \mathbf{Y}(s|\mathbf{x}, t)\rangle\rangle_{u,W} = \langle\langle \mathbf{Y}(s|\mathbf{x}, t)\rangle_u\rangle_W = \mathbf{x}$. The homogeneity of the velocity field then gives rise to the independence of all statistical properties of the zero-mean function $\mathbf{Y}(s|\mathbf{x}, t) - \mathbf{x}$ on $\mathbf{x}$. Also one can partially separate the influence of Brownian motion $\mathbf{W}$ and focus first on the averaging over the distribution of Eulerian velocity $\mathbf{u}$ in (4.5). Hence, (4.4) can be rewritten as

$$\mathbf{Y}(s|\mathbf{x}, t) = \mathbf{x} + \sqrt{2D_0}\mathbf{W}(s) + \mathbf{Z}(\mathbf{s}, \mathbf{W}(\mathbf{s})), \tag{5.1a}$$

$$d\mathbf{Z}(s) = -\mathbf{u}(\mathbf{x} + \sqrt{2D_0}\mathbf{W}(s) + \mathbf{Z}(\mathbf{s}), s)ds, \tag{5.1b}$$

$$\mathbf{Z}(s = 0) = 0, \tag{5.1c}$$

and, without loss of generality, one can set $x = 0$ in (5.1b) for $\mathbf{Z}$. It should be emphasized that all these changes are possible only for the stationary and homogeneous velocity field with zero mean. For example, the linear flow field of Indeikina and Chang [12] does not satisfy the homogeneity condition and one should use (4.4).

Invoking the representation of Eulerian velocity $\mathbf{u}(\mathbf{x}, t)$ by the spatial Fourier transform, one can formally integrate (5.1b)–(5.1c) to get

$$\mathbf{Z}(t) = (2\pi)^{-m/2} \int_0^t \int_{-\infty}^{\infty} \hat{\mathbf{u}}(\mathbf{k}, s) \exp(\imath\sqrt{2D_0}\mathbf{k}\mathbf{W}(s)) \exp(\imath\mathbf{k}\mathbf{Z}(s)) d^m\mathbf{k}ds. \tag{5.2}$$

With the representation of trajectory $\mathbf{Y}(s|\mathbf{x}, t)$ by (5.1a) and (5.2), the averaging in (4.5) also can be partially decomposed on the independent averaging over distributions of $\hat{\mathbf{u}}$ and $\mathbf{W}$:

$$\begin{aligned}\langle\langle \exp(\imath\delta\mathbf{K}\mathbf{Y}(t|\mathbf{x}, t))\rangle\rangle_{u,W} = \\ \exp(\imath\delta\mathbf{K}\mathbf{x})\langle \exp(\imath\delta\sqrt{2D_0}\mathbf{K}\mathbf{W}(t))\langle \exp(\imath\delta\mathbf{K}\mathbf{Z}(t, \mathbf{W}(t)))\rangle_{Z(u|W)}\rangle_W,\end{aligned} \tag{5.3}$$

where the averaging over $\hat{\mathbf{u}}$ should be made first.

For the simple shear flow $\mathbf{u}(\mathbf{x}, t) = (0, u_2(x_1, t), 0)$, analyzed by Avellaneda and Majda [1], integral (5.2) provides the exact solution. In this case, without coupling of the fluctuations of the velocity field, $\mathbf{Z}(t)$ is a Gaussian random variable in the sense of averaging over the distribution of $\hat{\mathbf{u}}$ because $\hat{\mathbf{u}}$ is assumed Gaussian. In several dimensions, the nonlinear coupling of the fluctuations in different directions may result in the deviation of the distribution of trajectories from Gaussian at intermediate times. However, according to the central limit theorem, in the long-time limit, $\mathbf{Z}(t)$ again approaches a Gaussian variable regardless of the distribution of $\hat{\mathbf{u}}$ as a result of the summation of a large number of random steps (McComb [19, pp. 446–447], Gardiner

[10, pp. 37–38]) and this fact is well supported by numerous experiments on turbulent transport of passive scalar (for example, Shlien and Corrsin [21], McComb and Rabie [18], Groenhof [11], and many others).

Hence, one can simply assume that the Gaussian distribution of trajectories $\mathbf{Z}(t,\mathbf{W}(t))$ for a fixed $\mathbf{W}(t)$, that is strictly valid in the long-time limit (and it is the topic of interest), is not in great error also at intermediate times. In such a case, the averaging over the distribution of the velocity field in (4.5) reduces to

$$\langle \exp(\imath\delta\mathbf{KZ})\rangle_{Z(u|W)} = \exp\left(-\frac{\delta^2}{2}\langle(\mathbf{KZ})^2\rangle_{Z(u|W)}\right) = \exp\left(-\frac{\delta^2}{2}K_i\tilde{Z^2_{ij}}K_j\right), \tag{5.4}$$
$$\tilde{Z^2_{ij}} = \tilde{Z^2_{ij}}(t,\mathbf{W}(t)) = \langle Z_i(t,\hat{\mathbf{u}},\mathbf{W}(t))Z_j(t,\hat{\mathbf{u}},\mathbf{W}(t))\rangle_{\hat{u}},$$

where the tensor of "convective" mean-square displacement $\tilde{Z^2_{ij}}$ is obviously symmetric.

Now it is necessary to determine the partially averaged tensor $\tilde{Z^2_{ij}}$. As is evident from (5.2), it requires the evaluation of the quantity

$$\langle \hat{u}_i(\mathbf{k},s_1)\hat{u}_j(\mathbf{q},s_2)\exp(\imath(\mathbf{kZ}(s_1)+\mathbf{qZ}(s_2)))\rangle_{\hat{u},Z|W} \tag{5.5}$$

if one cannot solve (5.1) explicitly. Unfortunately, $\mathbf{Z}$ and $\hat{\mathbf{u}}$ are dependent and, in general, one cannot separate the averaging in (5.5) into two independent averagings $\langle\hat{u}_i\hat{u}_j\rangle_{\hat{u}}\langle\exp(\cdots)\rangle_{Z|W}$, as is suggested by Corrsin's independence hypothesis for long-time turbulent diffusion (Corrsin [5]).

However, we show (details are given in Appendices A and B) that, for the selected *stationary homogeneous isotropic Gaussian* Eulerian velocity field $\mathbf{u}$, (5.4) provides the exact result. (We invoke the symmetry of the velocity field and the factorization property of higher-order moments of the Gaussian distribution to show the Gaussian distribution of trajectories $\mathbf{Z}(t,\mathbf{W}(t))$ at fixed $\mathbf{W}(t)$.) By introducing the "deviation variables" for trajectories and velocity and utilizing the stationarity of both Eulerian and Lagrangian Gaussian velocity fields, we also have shown (see Appendix C for details) that, despite the dependence of $\mathbf{Z}$ on $\hat{\mathbf{u}}$, averaging in (5.5) leads to the same expression for $\tilde{Z^2_{ij}}$ as if Corrsin's independence hypothesis holds for any diffusion time. Hence, using the above result and the two-point two-time correlation (3.1)–(3.3) yields the representation for $\tilde{Z^2_{ij}}$ in terms of the enhanced diffusivity tensor $\tilde{D_{ij}}(t) \equiv \tilde{D_{ij}}(t,\mathbf{W}(t))$. In the 3-D case ($m=3$ below) it is given by (see Appendix C)

$$\frac{1}{2}\tilde{Z^2_{ij}}(t,\mathbf{W}(t)) = \int_0^t \tilde{D_{ij}}(s)ds, \tag{5.6}$$

$$\tilde{D_{ij}}(s) = \frac{\alpha^2}{(2\pi)^m}\int_{-\pi}^{\pi}\int_0^{2\pi}\int_\delta^1 k^{-\epsilon}\tilde{F}(\mathbf{k},s)\left(\delta_{ij}-\frac{k_ik_j}{k^2}\right)dkd\phi d\theta, \tag{5.7a}$$

$$d\tilde{F}(s) = (1-[ak^z+k^2D_0+k_i\tilde{D_{ij}}(s)k_j]\tilde{F})ds - \tilde{G}\sqrt{2D_0}k_ndW_n(s), \tag{5.7b}$$

$$d\tilde{G}(s) = -[ak^z+k^2D_0+k_i\tilde{D_{ij}}(s)k_j]\tilde{G}ds + \tilde{F}\sqrt{2D_0}k_ndW_n(s), \tag{5.7c}$$

$$\tilde{D_{ij}}(0)=0, \qquad \tilde{F}(\mathbf{k},0)=0, \qquad \tilde{G}(\mathbf{k},0)=0, \tag{5.7d}$$

and for 2-D flow one should set $m=2$ and omit integration over θ in (5.7a).

Note that without molecular diffusion ($D_0=0$), the enhanced diffusivity tensor, as well as the concentration distribution, is already fully averaged. After the integration of angular dependence in (5.7a), $\tilde{D_{ij}}(s)$ reduces to an isotropic one, $\tilde{D_{ij}}(s) =$

$\delta_{ij}\overline{D(s)}$, and, consequently, $k_i\tilde{D_{ij}}(s)k_j = k^2\overline{D(s)}$, as it should be because of the isotropy of the velocity field. It is clear that, since the Brownian motion $\mathbf{W}(t)$ also is isotropic, there does not exist any interaction that can create some preferred direction. However, formally one cannot integrate out the angular dependence in (5.7a) and hence reduce the tensor to a single scalar before the molecular diffusion effect is averaged or neglected because of the presence of the last term in the right-hand side of (5.7b).

Hence, invoking (5.3) and (5.6), one obtains for the average concentration distribution (4.5):

$$\begin{aligned}\overline{C(\mathbf{x},t)} &= (2\pi)^{-n}\int_{-\infty}^{\infty}\hat{C}_0(\mathbf{K})\exp(\imath\delta\mathbf{K}\mathbf{x})\\ &\times\left\langle\exp\left(\imath\delta\mathbf{K}\mathbf{W}(t)\sqrt{2D_0}-\delta^2\int_0^t K_i\tilde{D_{ij}}(s)K_j ds\right)\right\rangle_W d^n\mathbf{K},\end{aligned}\tag{5.8}$$

where $\tilde{D_{ij}}$ is given by (5.7).

As also is evident from (5.7b) and (5.7c), in the limit of infinite time, the average value of F approaches fixed point value $[ak^z + k^2D_0 + k^2\tilde{D}_\infty]^{-1}$ while the average value of $G \to 0$. The dispersion of F, which is determined by $k^2D_0\langle G^2\rangle_W$, can remain finite in general or even large in comparison with $(\langle F\rangle_W)^2$. Such an effect has been found by Avellaneda and Majda [1] for the simple shear flow for some range of spectral parameters that are, however, very far from the region of interest of turbulent problems. Fortunately, this is not the case for isotropic 2- or 3-D flow. In several dimensions the coupling of fluctuations in different directions results, from one side, in a larger enhanced dissipation and, from the other side, in the necessary existence of the "random walk" limit with constant diffusivity. Hence, any nonlocal behavior due to the interaction of the flow with molecular diffusion, as in the Avellaneda and Majda [1] case, can only be a transient effect.

It has been specially checked that in the large-scale long-time limit the contribution of dW_m terms in $\tilde{D}$ and, consequently, the dispersion always remains negligible in comparison with the average for all values of spectral parameters. This means that the fluctuations of $\tilde{D}$ due to the interaction with the molecular diffusion are the higher-order effect and only the mean value affects the solute distribution (5.8). Hence, in order to simplify the presentation of the renormalization procedure, the averaging in (5.7) and (5.8) will be taken independently now since the cross-interaction terms always become irrelevant in the large-scale long-time limit. As a result, (5.7) and (5.8) become

$$\overline{D(s)} = \frac{\alpha^2(m-1)}{2m\pi^{m-1}}\int_\delta^1 k^{-\epsilon}\overline{F}(k,s)dk,\tag{5.9a}$$

$$\frac{\partial\overline{F}}{\partial s} = 1 - [ak^z + k^2D_0 + k^2\overline{D(s)}]\overline{F},\tag{5.9b}$$

$$\overline{D(0)} = 0, \qquad \overline{F(k,0)} = 0,\tag{5.9c}$$

$$\begin{aligned}\overline{C(\mathbf{x},t)} &= (2\pi)^{-n}\int_{-\infty}^{\infty}\hat{C}_0(\mathbf{K})\exp(-\imath\delta\mathbf{K}\mathbf{x})\\ &\times\exp\left[-\delta^2K^2\left(D_0t+\int_0^t\overline{D(s)}ds\right)\right]d^n\mathbf{K}.\end{aligned}\tag{5.10}$$

6. Large-scale long-time behavior of the averaged concentration distribution. It should be recalled that all variables in (5.9)–(5.10), except the initial data wavevector $\mathbf{K}$, are made dimensionless in the dissipation length scale, while the topic of interest is the integral-scale averaged solute distribution. Consequently, in order to find this large-scale long-time limiting behavior of the spreading process, one should rescale other variables in (5.9)–(5.10),

$$\mathbf{x} = \mathbf{x}^*/\delta, \qquad k = k^* g(\delta), \qquad t = t^*/\rho^2(\delta), \tag{6.1}$$

such that new nondimensional space and time variables become

$$\mathbf{x}^* = \mathbf{x}'/L_0, \quad k^* = k' L_0 \left[\frac{\delta}{g(\delta)}\right], \quad t^* = t'/\tau, \quad \text{where } \tau = \frac{L_0^2}{\nu}\left[\frac{\delta}{\rho(\delta)}\right]^2. \tag{6.2}$$

The scaling functions $\rho(\delta)$ and $g(\delta)$ then must be determined from the requirement that the averaged concentration distribution has nontrivial limiting behavior, namely, it is never equal to zero nor the initial data,

$$\overline{C^*(\mathbf{x}^*, t^*)} = \lim_{\delta\to 0} \overline{C\left(\mathbf{x}^*/\delta, t^*/\rho^2(\delta)\right)}; \tag{6.3a}$$

$$\overline{C^*(\mathbf{x}^*, t^*)} \neq 0, \qquad \overline{C^*(\mathbf{x}^*, t^*)} \neq C_0(\mathbf{x}^*). \tag{6.3b}$$

Note that the scaling $\rho(\delta) = \delta$ results in the diffusion time scale $\tau = L_0^2/\nu$ and scalings $\rho(\delta) = \delta^b$ with $b < 1$ correspond to shorter time scales. This means that the spreading occurs faster than pure diffusion motion.

The application of rescaling (6.1) to (5.9)–(5.10) then gives the representation for the averaged solute distribution

$$\overline{C^*(\mathbf{x}*, t^*)} = \int_{-\infty}^{\infty} \hat{C}_0(\mathbf{K}) \exp\left[-\imath \mathbf{K}\mathbf{x}^* - K^2 \int_0^{t^*} \overline{D_*(s^*)} ds^*\right] \frac{d^n\mathbf{K}}{(2\pi)^n}, \tag{6.4}$$

which corresponds to the large-scale effective diffusion equation

$$\frac{\partial \overline{C^*}}{\partial t^*} = \overline{D_*(t^*)} \boldsymbol{\nabla_*}^2 \overline{C^*}, \qquad C^*(x^*, 0) = C_0(x^*) \tag{6.5}$$

in nondimensional variables specified by (6.2). The conditions in (6.3b) imply that the effective diffusivity $\overline{D_*(t^*)}$ must satisfy $0 < \overline{D_*(t^*)} < \infty$ and it is then defined by the large-scale limit

$$\begin{aligned} \overline{D_*(t^*)} &= \lim_{\delta\to 0}\left[\frac{\delta}{\rho(\delta)}\right]^2 \left(D_0 + \overline{D\left(t^*/\rho^2(\delta), \delta, g(\delta)\right)}\right) \\ &= \lim_{\delta\to 0}\left[\frac{\delta}{\rho(\delta)}\right]^2 D_0 + \triangle D(t^*), \end{aligned} \tag{6.6}$$

where

$$\triangle D(t^*) = \lim_{\delta\to 0} \frac{\delta^2 g^{1-\epsilon}(\delta)}{\rho^4(\delta)} \alpha_*^2 \int_{\delta/g(\delta)}^{1/g(\delta)} k^{*-\epsilon} \overline{F}(k^*, t^*, \delta) dk^*, \tag{6.7a}$$

$$\frac{\partial \overline{F}}{\partial t^*} = 1 - \left(\frac{g^z}{\rho^2} a k^{*z} + \frac{g^2}{\rho^2} k^{*2} D_0 + \frac{g^2}{\delta^2} k^{*2} \triangle D(t^*)\right) \overline{F}, \qquad \overline{F}(k^*, 0) = 0, \tag{6.7b}$$

$$0 < \triangle D(t^*) < \infty, \tag{6.7c}$$

and the new amplitude constant

$$\alpha_*^2 = \frac{\alpha^2(m-1)}{2m\pi^{m-1}} \tag{6.8}$$

is introduced only to simplify the notation since the spectral amplitude α^2 is assumed to be independent of δ.

In the next section the limits (6.6) and (6.7) will be evaluated with an appropriate choice of scaling functions for all possible values of exponents $-1 < \epsilon < 3$ and $z \geq 0$ of the velocity spectrum (3.2).

7. Renormalization for different parameters of the velocity spectrum.

7.1. Mean field regime (region 1). This regime corresponds to the case when the standard diffusive scaling $\rho(\delta) = \delta$ describes a large-scale long-time behavior of the average concentration field. Hence, one can expect that the molecular diffusion also will make a contribution into the effective diffusion coefficient. With the choice $\rho(\delta) = \delta$ and $g(\delta) = 1$ in (6.6) and (6.7) one obtains

$$\overline{D_*} = \triangle D + D_0, \tag{7.1a}$$

$$\triangle D(t^*) = \lim_{\delta \to 0} \frac{1}{\delta^2} \alpha_*^2 \int_\delta^1 k^{*-\epsilon} \overline{F}(k^*, t^*, \delta) dk^*, \tag{7.1b}$$

$$\frac{\partial \overline{F}}{\partial t^*} = 1 - \left(a k^{*z} + k^{*2} D_0 + k^{*2} \triangle D(t^*) \right) \frac{\overline{F}}{\delta^2}, \qquad \overline{F}(k^*, 0) = 0, \tag{7.1c}$$

$$0 < \triangle D(t^*) < \infty. \tag{7.1d}$$

It is evident from (7.1c) that as $\delta \to 0$ the fixed point value $\overline{F}_\infty/\delta^2 = [a k^{*z} + k^{*2} D_0 + k^{*2} \triangle D(\infty)]^{-1}$ is reached exponentially fast. Hence, for any $t^* > 0$ one can set the enhanced diffusivity equal to the limiting value $\triangle D(t^*) = \triangle D(\infty)$. Consequently, (7.1b) and the fixed point value from (7.1c) provide the expression for $\triangle D(\infty)$:

$$\triangle D(\infty) = \alpha_*^2 \int_0^1 \frac{k^{*-\epsilon} dk^*}{a k^{*z} + k^{*2} D_0 + k^{*2} \triangle D(\infty)}. \tag{7.2}$$

The integral in (7.2) is finite for $\epsilon < 1 - z$ and $z < 2$ or for $\epsilon < -1$ and $z \geq 2$ which, together with the lowest allowable value of $\epsilon > -1$ define the boundary of the mean field regime: $-1 < \epsilon < 1 - z$. For such spectra, there is no infra-red divergence of the kinetic energy and the contributions of the individual fluctuations are, in fact, simply summed. Instead, because the upper limit of integration in (7.2) remains finite, the ultraviolet cut-off, which corresponds to the decaying part of the spectrum at large wavenumbers in Figure 1(b), is important. Consequently, the simple cut-off by setting the velocity correlation tensor equal to zero above some highest wavenumber may be inappropriate for peaked spectra.

7.2. Superdiffusive regimes. These regimes describe spreading that is faster than pure diffusion motion. As has been already mentioned, such dispersion requires the shorter time scale in (6.2),

$$\frac{\delta}{\rho(\delta)} \to 0 \quad \text{as} \quad \delta \to 0. \tag{7.3}$$

Physically, this corresponds to convection-dominant spreading and to spectra with infra-red divergence of kinetic energy of fluctuations. Hence, one can expect that

molecular diffusion will be negligible under such conditions. Indeed, the term containing D_0 in (6.6) vanishes in this limit and the effective diffusivity $\overline{D_*} = \triangle D(t^*)$. However, since for larger values of ϵ the infinite-time enhanced diffusivity (7.2) diverges at low wavenumbers, one needs to introduce a nontrivial wavenumber rescaling

$$g(\delta) \neq 1, \quad g(\delta) \to 0 \quad \text{as} \quad \delta \to 0 \tag{7.4}$$

in order to reach the convergence of the integral in (6.7a) and hence to satisfy the condition of the existence of the nontrivial solution (6.7c). Consequently, for some values of $g(\delta)$, the molecular diffusion coefficient still can enter into the expression for $\overline{D_*}$ through (6.7b).

However, let us consider first the natural choice of $g(\delta) = \delta$ which is stipulated by the infra-red cut-off of the spectrum. In this case, because of the requirement (7.3), the molecular diffusion term in (6.7b) vanishes and the governing equations for renormalization become

$$\triangle D(t^*) = \lim_{\delta \to 0} \frac{\delta^{3-\epsilon}}{\rho^4(\delta)} \alpha_*^2 \int_1^{1/\delta} k^{*-\epsilon} \overline{F}(k^*, t^*, \delta) dk^*, \tag{7.5a}$$

$$\frac{\partial \overline{F}}{\partial t^*} = 1 - \left(\frac{\delta^z}{\rho^2} a k^{*z} + k^{*2} \triangle D(t^*) \right) \overline{F}, \qquad \overline{F}(k^*, 0) = 0, \tag{7.5b}$$

$$0 < \triangle D(t^*) < \infty. \tag{7.5c}$$

Because the dependence on δ still remains in (7.5b), it is necessary to analyze three separate cases:

$$\frac{\delta^z}{\rho^2} \to \infty, \qquad \frac{\delta^z}{\rho^2} \to 0, \qquad \text{and} \quad \frac{\delta^z}{\rho^2} \equiv 1 \quad \text{as} \quad \delta \to 0$$

I. $\frac{\delta^z}{\rho^2} \to \infty$ (region 2). Under such conditions, the fixed point value in (7.5b) is again reached infinitely fast. Hence, similar to the mean field regime, the enhanced diffusivity is effectively time-independent and can be set equal to its limiting value for any $t^* > 0$. The substitution of the fixed point value $\overline{F}_\infty = \rho^2/(\delta^z a k^{*z})$ of (7.5b) into (7.5a) and the straightforward integration gives the renormalized enhanced diffusivity

$$\triangle D(t^* > 0) = \triangle D(\infty) = \frac{\alpha_*^2}{a[z + \epsilon - 1]} \lim_{\delta \to 0} \frac{\delta^{3-\epsilon-z}}{\rho^2(\delta)}. \tag{7.6}$$

The requirement of the nontriviality solution (7.5c) implies that $\triangle D(\infty)$ should be positive and that the limit in the right-hand side of (7.6) should be finite. Without loss of generality, one can set it equal to one, because any constant value can be included into ρ or $\triangle D$.

Hence, these restrictions together with the definition of the case define the time-rescaling function $\rho(\delta)$ and the boundaries for region 2 of renormalization

$$\rho(\delta) = \delta^{(3-\epsilon-z)/2} \qquad \text{for} \qquad 1 - z < \epsilon < 3 - 2z \tag{7.7}$$

with the effective diffusivity

$$\overline{D_*} = \triangle D(\infty) = \frac{\alpha_*^2}{a[z + \epsilon - 1]}. \tag{7.8}$$

II. $\frac{\delta^z}{\rho^2} \to 0$ (region 3). In this case (7.5b) is independent on δ and hence the time-rescaling function $\rho(\delta)$ can be determined directly from the limit in (7.5a) that gives $\rho(\delta) = \delta^{(3-\epsilon)/4}$, provided that the integral over k converges. Note that the integral should converge for any time, $0 \le t^* < \infty$, since (7.5) predicts the time-dependent effective diffusivity in this regime. The estimate of the limiting behavior of the enhanced diffusivity at small time, $t^* \to 0$, gives

$$\triangle D(t^* \to 0) \approx \frac{t^* \alpha_*^2}{\epsilon - 1} \left(1 - \delta^{\epsilon - 1}\right),$$

and the requirement (7.5c) then provides the lower bound for ϵ, $\epsilon > 1$. The restriction for the superdiffusive scaling (7.3) and the definition of the case give another bound: $3 - 2z < \epsilon \le 3$. Consequently, region 3 of the renormalization is finally defined by

$$\rho(\delta) = \delta^{(3-\epsilon)/4} \qquad \text{for} \qquad \max\{1, 3 - 2z\} < \epsilon \le 3, \tag{7.9}$$

$$\overline{D_*} = \triangle D(t^*) = \alpha_*^2 \int_1^\infty k^{*-\epsilon} \overline{F}(k^*, t^*) dk^*, \tag{7.10}$$

$$\frac{\partial \overline{F}}{\partial t^*} = 1 - k^{*2} \triangle D(t^*) \overline{F}, \qquad \overline{F}(k^*, 0) = 0, \tag{7.11}$$

with the following limiting behavior at short and long time:

$$\triangle D(t^* \to 0) \approx \frac{t^* \alpha_*^2}{\epsilon - 1}, \qquad \triangle D(t^* \to \infty) \to \frac{\alpha_*}{\sqrt{\epsilon + 1}}. \tag{7.12}$$

III. $\frac{\delta^z}{\rho^2} \equiv 1$ (Kolmogorov boundary). This case corresponds to the boundary between the regions 2 and 3 and it is defined by $\epsilon = 3 - 2z$, $z < 1$, where both scaling functions (7.7) and (7.9) collapse into $\rho(\delta) = \delta^{z/2}$. From the order-of-magnitude and dimensional analysis of section 3 one can conclude that the physically plausible velocity correlation, which can provide the energy cascade and remain invariant under the scaling transformations, should belong to this boundary. The Kolmogorov velocity spectrum with $\epsilon = 5/3$ and $z = 2/3$ also corresponds to this regime and that is why this region is called the "Kolmogorov boundary." The effective diffusivity $\overline{D_*}$ is hence given by (7.10), but (7.11) for $\overline{F}(k^*, t^*)$ becomes different,

$$\frac{\partial \overline{F}}{\partial t^*} = 1 - \left(a k^{*z} + k^{*2} \triangle D(t^*)\right) \overline{F}, \tag{7.13}$$

and in such a case one obtains the widest dependence on the parameters of the spectrum. The small-time asymptote in (7.12) does not change on this boundary, but in the large-time limit $t^* \to \infty$ the evaluation of the integral in (7.10) with $\epsilon = 3 - 2z$ and $\overline{F}_\infty = \left(a k^{*z} + k^{*2} \triangle D(\infty)\right)^{-1}$ results in the following implicit expression for $\triangle D(\infty)$:

$$\frac{a^2 (2 - z)}{\alpha_*^2} = \frac{a}{\triangle D(\infty)} - \ln\left(1 + \frac{a}{\triangle D(\infty)}\right). \tag{7.14}$$

As is evident from (7.14), for any fixed z (or ϵ), $\triangle D(\infty)/a$ depends, in fact, on one parameter, a^2/α_*^2. It should be noted that, as has been established in section 3 (equations (3.11)–(3.12)), in this regime $a \sim \alpha_*$ and the proportionality constant should not depend much on flow conditions, as is stipulated by the Kolmogorov similarity hypothesis. The parameters α_* and δ also are related to the root-mean-square

velocity and the large-scale Reynolds number through (3.7)–(3.10). Hence, in fact, in the dimensional version of (7.14), there are no free parameters to play with in order to fit any experimental data.

Intermediate wavenumber scaling $\frac{\delta}{g(\delta)} \to 0$ as $\delta \to 0$ (region 4). The diffusive ($g(\delta) = 1$) and convective ($g(\delta) = \delta$) wavenumber scalings have already been considered. However, it still remains the part of ϵ–z plane, $\max\{-1, 3-2z\} < \epsilon \leq 1$, where these scalings do not give successful renormalization. Consequently, one should select some intermediate wavenumber scaling, something like $g(\delta) = \delta^b$ with $b < 1$, such that $\delta/g(\delta) \to 0$ as $\delta \to 0$ slower than for pure diffusive scaling.

For simple shear flow (Avellaneda and Majda [1]), the requirement $\delta/g(\delta) \to 0$ splits this remaining part of the ϵ–z plane into two regions by the line $z = 2$. In both cases the integration over k in (6.7a) is extended from 0 to ∞ such that both cut-offs are negligible, and the analysis of Avellaneda and Majda [1] results in an effective diffusivity that grows as a noninteger power of time, $\triangle D \sim t^c$ with $0 < c < 1$. The coefficient of proportionality is determined by time-correlation of the velocity fluctuations (ak^{*z} term in (6.7b)) for $z < 2$. For $z > 2$, random nonlocal diffusivity has been obtained because of nonlinear interaction of the velocity field with Brownian motion $\sqrt{2D_0}dW(t)$, which describes the molecular diffusion.

In 2- or 3-D flow, however, the mathematical consequence of the coupling of the fluctuations is the term $\triangle D(t^*)k^{*2}g^2/\delta^2$ in (6.7b) for $\overline{F}$ and, consequently, for $\triangle D(t^*)$ because they also are related by (6.7a). It can be easily seen that, under the conditions of the superdiffusive time-scaling $\delta/\rho \to 0$ as $\delta \to 0$, for any choice of the wavenumber scale satisfying $\delta/g(\delta) \to 0$, this "feedback" term immediately leads to the achievement of infinite-time limit in (6.7b). Hence, it is quite possible that the effective diffusivity grows like a noninteger power of time at the beginning of the spreading process, but in two or three dimensions this is a transient effect, and for large-scale initial data the infinite-time limiting behavior will be seen for any $t^* > 0$. However, if the initial data vary in some intermediate scale between L_d and L_0, it also will require the rescaling of the initial data wavevector $\mathbf{K}$. This rescaling should reduce the "feedback" term, such that the above transient behavior can become visible from the point of view of large-scale motion. However, since such assumption effectively adds a third dimension to ϵ–z parameter space, this analysis will not be pursued here, mostly because this region of renormalization is far away from the turbulent transport problems.

Instead, without changing the assumption of the integral-scale initial data, the fixed-point value $\overline{F}_\infty = (\delta/g(\delta))^2[k^{*2}\triangle D(\infty)]^{-1}$ is utilized to determine the scaling function $\rho(\delta)$. The integration in (6.7a) then results in the same scaling law as for region 3 and the effective diffusivity is equal to the infinite-time asymptote in (7.12). Because there is no convergence problem on the boundary $\epsilon = 1$ with region 3 and the scaling law also is continuous, this boundary can be included in region 4. The remaining boundaries, however, should be considered separately.

Boundaries $\{\epsilon = 1 - z,\ z < 2\}$ and $\{\epsilon = 3 - 2z,\ 1 \leq z < 2\}$. Since all these boundaries separate regions with constant effective diffusivity, one can expect their behavior will not be different. In all these cases, the integral in (7.10) can be evaluated explicitly with $\overline{F}(k^*, t^*) = \overline{F}_\infty$ and the requirement of the large-scale long-time distinguishing limit provides expressions for the effective diffusivity and time-rescaling function.

For the continuation of the Kolmogorov boundary into the region $1 \leq z < 2$, the scaling function is continuous on the boundary between regions 2 and 4, and the

enhanced diffusivity corresponds to the infinite-time limit (7.14) of the Kolmogorov boundary.

For the boundary between regions 1 and 2 the analysis results in the logarithmic dependence for ρ^2,

$$\rho(\delta) = \delta(-\ln\delta)^{1/2}, \qquad \overline{D_*} = \triangle D(\infty) = \frac{\alpha_*^2}{a} \quad \text{for} \quad \epsilon = 1 - z,\ z < 2, \tag{7.15}$$

and, consequently, (7.15) can be used at finite scale ratio δ only if δ is so small that $(-\ln\delta)^{1/2}$ is a large number.

8. Comparison and discussion. From the renormalization analysis of the advection-diffusion problem (2.1) with Gaussian random velocity field defined by two-point two-time correlation (3.1)–(3.2), it has been found that in the large-scale long-time limit the averaged spreading process can be described by the effective diffusion equation (6.5),

$$\frac{\partial \overline{C^*}}{\partial t^*} = \overline{D_*(t^*)} \boldsymbol{\nabla_*}^2 \overline{C^*}, \qquad C^*(x^*, 0) = C_0(x^*),$$

where nondimensional variables are specified by (6.2):

$$\mathbf{x}^* = \frac{\mathbf{x}'}{L_0}, \qquad t^* = \frac{t'}{\tau}, \quad \tau = \frac{L_0^2}{\nu}\left[\frac{\delta}{\rho(\delta)}\right]^2 = \tau_{diff}\left[\frac{\delta}{\rho(\delta)}\right]^2.$$

The effective diffusivity $\overline{D_*(t^*)} = \overline{D_*}(t^*, \epsilon, z, a, \alpha_*)$ and the scaling function $\rho(\delta)$ have been determined for different spectral parameters. It has been found that the functional forms of $\overline{D_*}$ and $\rho(\delta)$ depend only on the exponents ϵ and z of two-point two-time correlation (3.2) and several different regions of renormalization in ϵ–z have been determined. The regions of renormalization are shown in Figure 3 and the values of $\overline{D_*}$, $\rho(\delta)$ and $\tau(\delta)$ are summarized below:

Region 1. $-1 < \epsilon < 1 - z$.

$$\rho(\delta) = \delta, \quad \tau = \tau_{diff}, \quad \overline{D_*} = D_0 + \triangle D, \quad \triangle D = \int_0^1 \frac{\alpha_*^2 k^{*-\epsilon} dk^*}{ak^{*z} + k^{*2}(D_0 + \triangle D)}.$$

Boundary 1–2. $\epsilon = 1 - z$, $0 \le z < 2$.

$$\rho(\delta) = \delta(-\ln\delta)^{1/2}, \quad \tau = \tau_{diff}/(-\ln\delta), \quad \overline{D_*} = \triangle D = \frac{\alpha_*^2}{a}.$$

Region 2. $1 - z < \epsilon < 3 - 2z$.

$$\rho(\delta) = \delta^{(3-\epsilon-z)/2}, \quad \tau = \tau_{diff}\delta^{\epsilon+z-1}, \quad \overline{D_*} = \triangle D = \frac{\alpha_*^2}{a[z + \epsilon - 1]}.$$

Kolmogorov boundary (regions 2–3). $\epsilon = 3 - 2z$, $\epsilon > 1$, $z < 1$.

$$\rho(\delta) = \delta^{(3-\epsilon)/4} = \delta^{z/2}, \quad \tau = \tau_{diff}\delta^{(1+\epsilon)/2}, \quad \overline{D_*} = \triangle D(t^*),$$

$$\triangle D(t^*) = \alpha_*^2 \int_1^\infty k^{*-\epsilon} \overline{F}(k^*, t^*) dk^*,$$

$$\frac{\partial \overline{F}}{\partial t^*} = 1 - \left(ak^{*z} + k^{*2}\triangle D(t^*)\right)\overline{F}, \quad \overline{F}(k^*, 0) = 0.$$

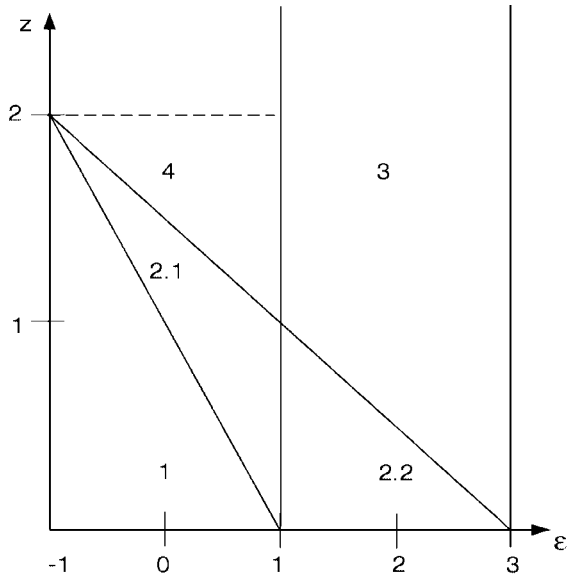


FIG. 3. *The regions of renormalization in the ϵ–z parameter plane. The dashed line corresponds to the split of region 4 in the 1-D case. The separation of region 2 occurs only in terms of Reynolds number because of the different dependence on the scale ratio δ for spectra with and without infra-red divergence of kinetic energy.*

Region 3. $\max\{1, 3-2z\} < \epsilon \le 3$.

$$\rho(\delta) = \delta^{(3-\epsilon)/4}, \quad \tau = \tau_{diff}\delta^{(1+\epsilon)/2}, \quad \overline{D_*} = \triangle D(t^*),$$
$$\triangle D(t^*) = \alpha_*^2 \int_1^\infty k^{*-\epsilon}\overline{F}(k^*, t^*)dk^*,$$
$$\frac{\partial \overline{F}}{\partial t^*} = 1 - k^{*2}\triangle D(t^*)\overline{F}, \quad \overline{F}(k^*, 0) = 0.$$

Region 4. $\max\{1, 3-2z\} < \epsilon \le 1$.

$$\rho(\delta) = \delta^{(3-\epsilon)/4}, \quad \tau = \tau_{diff}\delta^{(1+\epsilon)/2}, \quad \overline{D_*} = \triangle D = \frac{\alpha_*}{\sqrt{\epsilon+1}}.$$

Boundary 2–4. $\epsilon = 3-2z, \quad -1 < \epsilon \le 1, \quad 1 \le z < 2$.

$$\rho(\delta) = \delta^{(3-\epsilon)/4}, \quad \tau = \tau_{diff}\delta^{(1+\epsilon)/2}, \quad \overline{D_*} = \triangle D,$$
$$\frac{a^2(2-z)}{\alpha_*^2} = \frac{a}{\triangle D} - \ln\left(1 + \frac{a}{\triangle D}\right).$$

The case of spatially random steady flow $\mathbf{u} = \mathbf{u}(\mathbf{x})$ ($a = 0$) corresponds to the line $z = 2$ on the ϵ–z plane shown in Figure 1, because one should preserve the convergence of all integrals with only diffusive terms $\sim k^2$. There also is no need to set $a = 0$, because in regions 3 and 4 the effective diffusivity and the time-rescaling function are independent of both a and z.

If $\mathbf{u}$ is a white noise in time ($a \to \infty, \alpha^2 \sim a$), one should cross the ϵ–z diagram by the line $z = 0$ and evaluate the above limit for regions 1 and 2. Note that in

the present analysis the exponent of the spectral correlation function $\hat{Q} \sim k^{1-m-\epsilon}$, introduced by (3.2), is defined in a manner that depends on the space dimension m in such a way that the kinetic energy spectrum (per unit mass of liquid) depends only on ϵ. Hence, 2-D and 3-D cases have the same ϵ–z diagram, and only the amplitude constant α_*^2 of (6.8) depends on the space dimension. If one would like to work in terms of exponent of the spectral correlation function instead of kinetic energy, one simply has to move the ϵ–z diagram of Figure 3 to the right along the ϵ axis by one or two units and make corresponding changes of notation in the definitions of the time scale.

The ϵ–z diagram of Figure 3 is identical to that obtained by Avellaneda and Majda [1] for the simple shear flow with appropriate changes of notation ($\epsilon \leftrightarrow \tilde{\epsilon} - 1$, such that the left boundary of the mean field regime corresponds to $\tilde{\epsilon} = 0$) except for region 4. The effective diffusivities for two or three dimensions are different everywhere except in region 2. There the enhanced diffusivity is completely determined by strongly correlated (low z) high-wavenumber fluctuations of the velocity field. The above results also are consistent with earlier findings of Avellaneda and Majda [3] and Fannjiang [8] for multidimensional flows.

In general, the one-dimensional (1-D) analysis of Avellaneda and Majda [1] provides a small-time asymptote for the spreading process in 2-D or 3-D flow when enhanced diffusivity is small: a linear growth of effective diffusivity in time for region 3 and Kolmogorov boundary and noninteger power growth in time for two different parts of region 4. Later on, coupling of fluctuations in several dimensions necessarily results in the random walk limit, while this does not always happen in the 1-D simple shear flow. However, for region 3 and the Kolmogorov boundary, Avellaneda and Majda [3] suggest an effective nonlocal diffusion equation, while our analysis gives the usual diffusion equation with a time-dependent coefficient (which approaches a constant as $t \to \infty$). At the least, our infinite-time limit for the Kolmogorov spectrum (i.e., for usual turbulence) is well supported by numerous experimental data (see, for example, McComb [19, pp. 470–471] or Sherwood, Pigford, and Wilke [20, pp. 124–125] for data collections). Below, we also will compare our predicted time-evolution of mean-square displacement with available data, which will provide additional support for our results.

It has been shown in section 2 that the scale ratio δ and the spectral amplitude α_* are related to the large-eddy Reynolds number $R_0 = U_0 L_0/\nu$ and with the root-mean-square Reynolds number $R_{rms} = U_{rms} L_0/\nu$ by equations (3.7)–(3.10), and these dependencies are different for $\epsilon \ >, <, = \ 1$. It also is established in section 2 that, at least for the Kolmogorov boundary, a is of the order α. Hence, it also is useful to provide the expressions for the time scale τ and the effective diffusivity $\overline{D_*}$ in terms of the above Reynolds numbers. This summary is given below, where all time scales are defined with respect to the convective time scale of the fluctuations $\tau_{conv} = L_0/U_{rms}$ instead of the diffusion time scale $\tau_{diff} = L_0^2/\nu$, and the typically unknown spectral parameter a has been replaced by

$$\beta = a/\alpha_* \tag{8.1}$$

such that β is unit order coefficient.

Region 1. $-1 < \epsilon < 1 - z$.

$$\tau = \tau_{conv} R_0^{2/(\epsilon+1)}, \quad \overline{D_*} = \frac{D_0}{\tilde{\alpha}} + \triangle D, \quad \tilde{\alpha} = \frac{R_{rms}}{R_0^{-2/(\epsilon+1)}},$$

$$\triangle D = \int_0^1 \frac{(1-\epsilon) k^{*-\epsilon} dk^*}{\beta k^{*z} + k^{*2}(D_0/\tilde{\alpha} + \triangle D)}.$$

Region 2.1. $1 - z < \epsilon < 3 - 2z, \quad \epsilon < 1.$

$$\tau = \tau_{conv} R_0^{2(2-\epsilon-z)/(\epsilon+1)}, \quad \overline{D_*} = \frac{1-\epsilon}{\beta[z+\epsilon-1]}.$$

Region 2.2. $1 - z < \epsilon < 3 - 2z, \quad \epsilon > 1.$

$$\tau = \tau_{conv} R_0^{2(3-\epsilon-2z)/(\epsilon+1)}, \quad \overline{D_*} = \frac{\epsilon-1}{\beta[z+\epsilon-1]}.$$

Boundary 2.1–2.2. $\epsilon = 1, \quad 0 < z < 1.$

$$\tau = \tau_{conv} R_0^{1-z} \ln R_0, \quad \overline{D_*} = \frac{1}{\beta z}.$$

Kolmogorov boundary (regions 2–3). $\epsilon = 3 - 2z, \quad \epsilon > 1, \quad z < 1.$

$$\tau = \tau_{conv}, \quad \overline{D_*}(t^*) = (\epsilon - 1) \int_1^\infty k^{*-\epsilon} \overline{F}(k^*, t^*) dk^*,$$
$$\frac{\partial \overline{F}}{\partial t^*} = 1 - \left(\beta k^{*z} + k^{*2} \overline{D_*}(t^*) \right) \overline{F}, \quad \overline{F}(k^*, 0) = 0.$$

Region 3. $\max\{1, 3 - 2z\} < \epsilon \leq 3.$

$$\tau = \tau_{conv}, \quad \overline{D_*}(t^*) = (\epsilon - 1) \int_1^\infty k^{*-\epsilon} \overline{F}(k^*, t^*) dk^*,$$
$$\frac{\partial \overline{F}}{\partial t^*} = 1 - k^{*2} \overline{D_*}(t^*) \overline{F}, \quad \overline{F}(k^*, 0) = 0.$$

Region 4. $\max\{1, 3 - 2z\} < \epsilon < 1.$

$$\tau = \tau_{conv} R_0^{(1-\epsilon)/(1+\epsilon)}, \quad \overline{D_*} = \sqrt{\frac{1-\epsilon}{\epsilon+1}}.$$

Boundary 2–4. $\epsilon = 3 - 2z, \quad -1 < \epsilon < 1, \quad 1 < z < 2.$

$$\tau = \tau_{conv} R_0^{(1-\epsilon)/(1+\epsilon)}, \quad \frac{\beta^2(2-z)}{2(z-1)} = \frac{\beta}{\overline{D_*}} - \ln\left(1 + \frac{\beta}{\overline{D_*}}\right).$$

The boundary 1–2 with logarithmic scaling with respect to δ is not included in this summary because the accuracy of the logarithmically distinguished limit is already too low. It also should be mentioned that the above analysis is valid in the limit of infinite (i.e., very large, in practice) Reynolds number. This is especially true for spectra with weak infra-red divergence, when ϵ is close to 1. In such a case, the $|\epsilon - 1|$ factor in the expressions for the effective diffusivity should be replaced on $|(\epsilon - 1)/(1 - R_0^{(\epsilon-1)/(\epsilon+1)})|$, and this will make transitions across regime boundaries continuous.

Hence, depending on the data available, one can use any of these summaries to scale variables in the effective diffusion equation (6.5). It should be emphasized, however, that the integral scales L_0 and R_0 correspond to the largest possible eddies of the inertial range and are not the same as the mean flow parameters of the real flows. Because the lower limit of integration over k in the superdiffusive regimes has been set equal to 1, the integral length scale L_0 corresponds to the inverse of the maximal

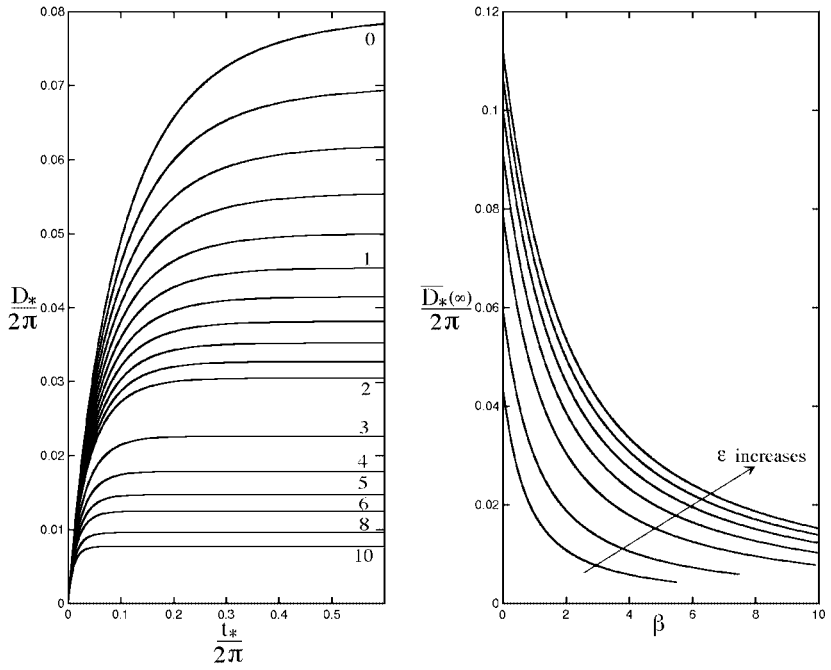


FIG. 4. *Effective diffusivity at the Kolmogorov boundary* $\epsilon = 3 - 2z$. (a) *Time-evolution of the effective diffusivity for the Kolmogorov spectrum* ($\epsilon = 5/3$) *with indicated values of* β. *In the range from* 0 *to* 2, β *increases with a* 0.2 *increment.* (b) *The dependence of* $\overline{D}_*(\infty)$ *on* β *for* $\epsilon = 7/6$; 4/3; 5/3; 2; 7/3; 8/3; 3.

possible wavenumber. Hence, by the usual convention for spectral methods, one can set $L_0 = L'/2\pi$ in the definition of the convective time scale, $\tau_{conv} = L_0/U_{rms} = L'/2\pi U_{rms}$, where L' is the conventional spatial dimension of the real flow (pipe diameter, for example).

The most interesting regions for turbulence are the Kolmogorov boundary and region 3, where the time scale τ does not depend on R_0, as it should be in the limit of the infinite Reynolds number. In other regions of renormalization, the physics that cause the lower infra-red divergence of the spectrum should be invoked to determine large-eddies characteristics.

The effective diffusivity depends on the molecular one only for the mean field regime (region 1), which corresponds to the slowest rate of spreading. In the superdiffusive regimes (regions 2–4), the transport process is faster than pure diffusion motion. The integral time scale for these cases takes the form $\tau = \tau_{diff}\delta^b$ or $\tau = \tau_{conv}R_0^c$, where $b > 0,\ 0 \le c < 1$. The shortest pure convective time scale belongs to region 3 and to the Kolmogorov boundary. These regions have the largest values of the parameter ϵ, which defines the strength of the infra-red divergence of the velocity spectrum. For these regions, time-dependent effective diffusivity is obtained.

It is well known, beginning from the classical work of Taylor [22], that the mean-square displacement of fluid particles $\langle X^2(t)\rangle$ by turbulent flow is defined by the convective scaling $\langle X^2(t)\rangle = \langle V^2\rangle t^2$ initially and approaches the random walk limit $\langle X^2(t)\rangle = 2\langle V^2\rangle T_L t$ for long diffusion time. Figure 4(a) shows the dependence of the effective diffusivity on time for the Kolmogorov spectrum for the different values

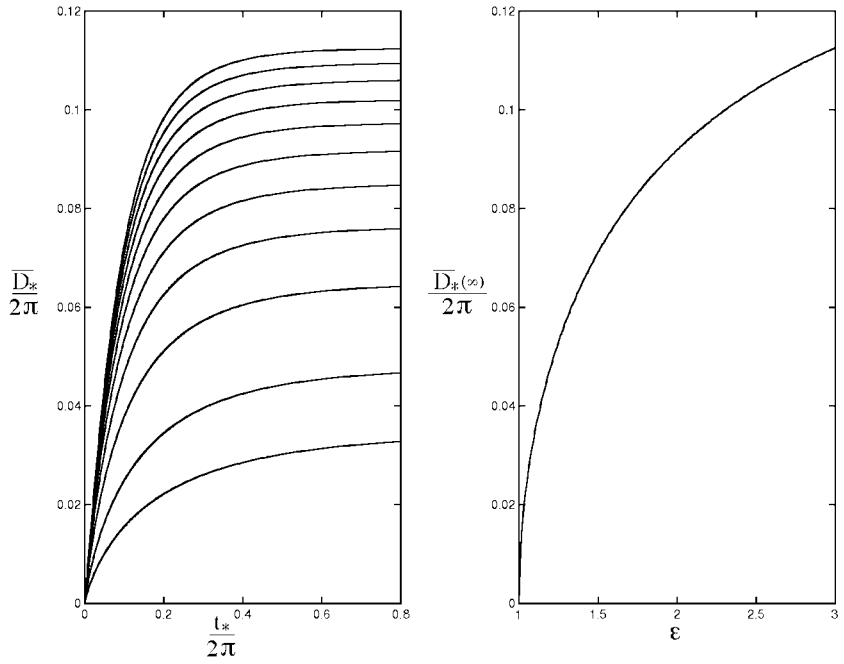


FIG. 5. *Effective diffusivity in region* 3. (a) *Time-evolution of the effective diffusivity. The lowest line corresponds to* $\epsilon = 1.1$ *and others correspond to the increasing values of* ϵ *from* 1.2 *to* 3 *with* 0.2 *increment.* (b) *The dependence of* $\overline{D}_*(\infty)$ *on* ϵ.

of the parameter β. Recalling the definition of the effective diffusivity through the "convective" mean-square displacement by (5.6),

$$\overline{Z}^2 = 2\int_0^t \overline{D_*(s)}ds,$$

one can conclude that the time-evolution of the effective diffusivity is qualitatively consistent with the classical result: linear growth of $\overline{D_*}(t)$ at $t \to 0$ and approach to the constant value $\overline{D_*}(\infty)$ at large times for all values of β. For other values of $\epsilon = 3 - 2z$ at the Kolmogorov boundary, the time-evolution of the effective diffusivity occurs in a similar manner, as shown in Figure 4(a).

Increasing the parameter β corresponds to decreasing the correlation time for the fluctuation of the velocity field. As one can see in Figure 4(a), the random walk limit is reached faster when the correlation time decreases. The largest diffusivity enhancement corresponds to the steady flow case $\beta = 0$, which produces the $D_*(t)$ dependence of region 3 for the corresponding value of ϵ. The dependence of the infinite-time asymptotic diffusivity $\overline{D_\infty} = \overline{D_*}(\infty)/2\pi$ on β for the Kolmogorov boundary $\epsilon = 3 - 2z$ is shown in Figure 4(b) for several values of ϵ.

The dependence $\overline{D_*}(t)$ for region 3 has a similar shape as that for the Kolmogorov boundary and is shown in Figure 5(a) for several values of ϵ. The stronger infra-red divergence of the spectrum (higher ϵ) results in larger effective diffusivity and shorter time to reach the random walk limit. The dependence of infinite-time asymptote $\overline{D_\infty}$ on ϵ in Figure 5(b) confirms that the boundary $\epsilon = 1$ does not belong to this region, since $\overline{D_\infty}(\epsilon = 1) = 0$, and a different, slower time scale should be imposed in this

case. Note that, for fixed β, the dependence of the effective diffusivity on ϵ for the spectra from the Kolmogorov boundary is similar to one shown in Figure 5.

The superdiffusive regimes in regions 2 and 4 correspond to intermediate time scales between diffusion and convection. In these time scales, the transition period for $\overline{D_*} = \overline{D_*}(t)$ is short and one immediately sees the limiting value $\overline{D_*}(\infty)$ in the integral time scale.

We shall compare our theoretical predictions to experimental data for the transverse (normal) diffusivity for turbulent channel flows. Due to the zero-mean velocity assumption, we cannot capture the downstream diffusivity of the most common turbulent flows. In order to make a comparison with earlier experimental and numerical works on turbulent transport, the Kolmogorov spectrum is chosen, which is defined by $\epsilon = 5/3$, $z = 2/3$. Unfortunately, experimental measurements of the two-point two-time correlation in well-developed isotropic turbulence has not been found. Different spectral theories suggest different relations between the value of β and the Kolmogorov constant α_0 (mostly like $\beta = c\alpha_0^2$, where the constant c varies from one theory to another) and, at the same time, predict the value of the Kolmogorov constant that is several factors off [19]. Hence, instead of selecting the appropriate value of β from existing theoretical predictions, we shall determine it empirically from the asymptotic diffusivity at infinite time.

Note that the values of $\overline{D} = \overline{D_*}/2\pi$ and $t = t_*/2\pi$ shown in Figure 4 are already rescaled according to the conventional choice $L_0 = L'/2\pi$, such that the dimensional effective diffusivity and time are

$$\overline{D}' = \overline{D} U_{rms} L', \qquad t' = \frac{tL'}{U_{rms}},$$

where L' is the appropriate spatial dimension for the real turbulent flow. For pipe flows, for example, the long-time effective diffusivity in the transverse direction scaled on the friction velocity and a pipe diameter is about $[3-4]\times 10^{-2}$ [11], [18], [19].

Hence, it is already evident in Figure 4(b) that the infinite time asymptotic $\overline{D_\infty}$ for the Kolmogorov spectrum provides a correct order-of-magnitude estimate for any $\beta < 10$. Because the root-mean-square velocity near the center of the pipe is typically about 0.8 of the friction velocity U_τ [19], one can expect a quantitative agreement with the choice of $\beta \approx 1-2$, and the selected value of β should not vary from one particular set of data to another. Figure 6 reproduces experimental data collection from the book of Sherwood, Pigford, and Wilke [20, Figure 4.11, p. 125]. The solid line $E_D = \overline{D}'_\infty$ is calculated using the value of $\beta = 1$ for the Kolmogorov spectrum, which gives $\overline{D}_\infty = 4.55\times 10^{-2}$ and the typical value 0.8 for the ratio U_{rms}/U_τ. Hence, in terms of the friction velocity, the long-time effective diffusivity becomes

$$\overline{D}'_\infty = 0.8\overline{D}_\infty(\epsilon = 5/3, z = 2/3, \beta = 1) U_\tau d = 3.64\times 10^{-2} U_\tau d,$$

where d is the pipe diameter. To express the friction velocity in terms of the average velocity U_{av}, the well-known expression for the friction factor in a pipe flow has been invoked,

$$U_\tau = U_{av}\sqrt{f/2}, \qquad f = 0.067\left(\frac{U_{av}d}{\nu}\right)^{-1/4},$$

because most of the data in the high Reynolds number part of the figure has been taken for air flow in pipes. Hence, in variables of Figure 6, the line $E_D = \overline{D}'_\infty(U_{av}d)$

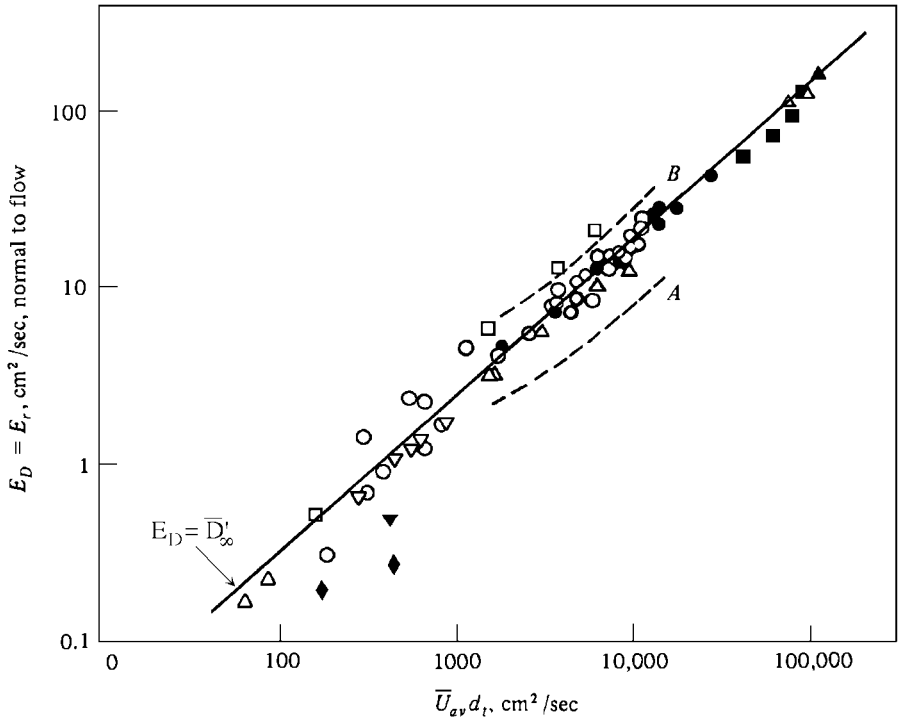


FIG. 6. *Eddy diffusion coefficient for transport normal to the flow direction in pipes and flat ducts. The data are taken from Sherwood, Pigford, and Wilke* [20, *Figure* 4.11, *p.* 125]. *The solid line indicated* $E_D = \overline{D}'_\infty$ *uses the Kolmogorov velocity spectrum with* $\beta = 1$.

is given by

$$E_D = 7.26 \times 10^{-3} \nu Re^{7/8} = 7.26 \times 10^{-3} \nu^{1/8} (U_{av} d)^{7/8} = 5.73 \times 10^{-3} (U_{av} d)^{7/8},$$

where in the last equality $\nu = 0.15$ cm^2/sec for air has been used and, consequently, $U_{av}d$ should be taken in the same units.

It is evident from Figure 6 that the choice of $\beta = 1$ gives a reasonable representation for the long-time effective diffusivity over a wide range of Reynolds numbers. Lower values of β quickly lift up the calculated line, while a slightly higher number, say $\beta = 1.2$, still gives a good representation. For $\beta = 2$, however, the $E_D = \overline{D}'_\infty$ line is already outside most of data points in Figure 6. These data hence allow us to select a very specific coefficient β for the decorrelation time spectral parameter. We shall use this value of $\beta = 1$ in all subsequent comparisons, without further adjustment to the particular flow conditions.

The usual measurable time-dependent quantity in turbulent transport experiments is the solute/temperature profile. For diffusion from a point or line source, it is well represented by a Gaussian shape. This allows the determination of the mean-square displacement $Z^2(t)$, and the slope of the long-time path of the curve $Z^2(t)/2$ gives the long-time value of effective diffusivity. Unfortunately, the mean-square displacement data are not reported so often as the long-time diffusivity and the resolution of the small-time region are typically low.

The solid line in Figure 7 shows the small-time evolution of the nondimensional mean-square displacement for the Kolmogorov spectrum with $\beta = 1$, and the dashed

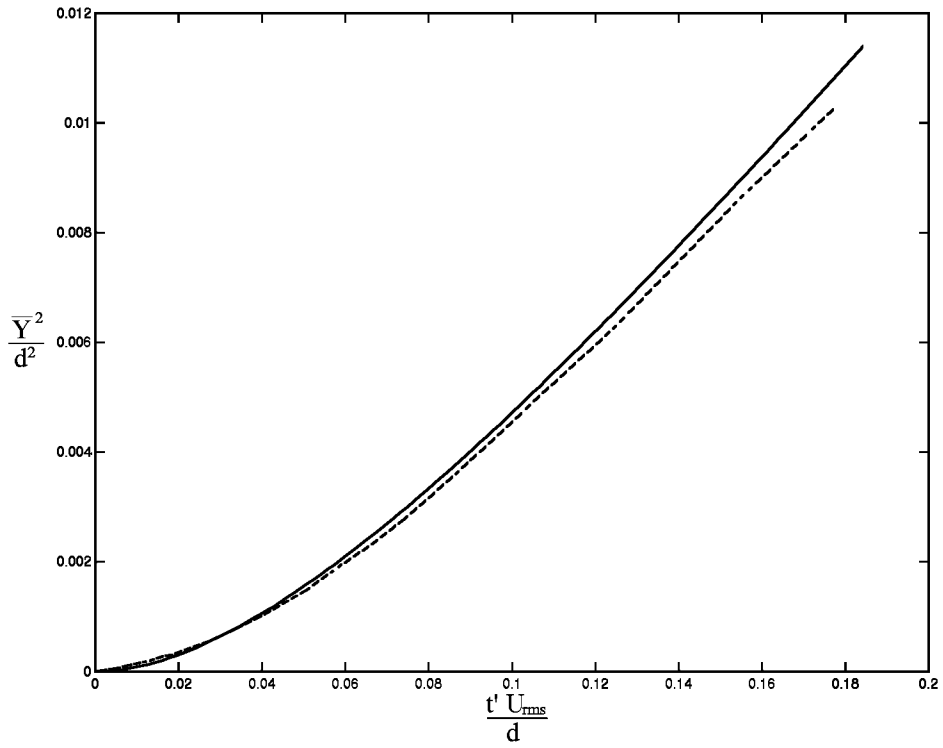


FIG. 7. *Mean-square displacement of fluid particles by turbulent flow.* (—) *Current theory with Kolmogorov velocity spectrum and* $\beta = 1$. (- -) *Experimental heat-transfer data of Baldwin and Walsh* [4] *in turbulent pipe flow, scaled by the reported values of the root-mean-square velocity and the pipe diameter.*

line corresponds to one of the data point from Figure 6 (the lowest black square in the high-Re range [4]). The experimental curve is scaled according to the suggestion of the present theory, on the reported values of the root-mean-square velocity ($U_{rms} \approx 0.035U$ with $U = 22$ m/sec), and the pipe diameter (20 cm). It is evident that small-time evolution of the mean-square displacement also is reproduced by the current theory with reasonable accuracy.

Near the axis of the pipe, turbulence is approximately isotropic. In the numerical experiment of Deardorff and Peskin [7], the trajectories of the fluid particles, released at 1/4 of the channel height and well within the region of mean shear for a turbulent plane air flow, have been calculated. The different statistical quantities, like mean-square displacements in different directions, correlation functions, mean-square particles separation, and so on, have been obtained by averaging the results for different sets of particles. All results are reported in nondimensional form, scaled on the friction velocity and channel height.

Since diffusion starts from a region of appreciable shear, the effective ratio of the root-mean square to friction velocity can vary significantly with time, even if spreading in a spatially homogeneous lateral direction is considered. This is because of the cross-influence of the fluctuations in different directions, included in the present theory and also apparent in reported results of Deardorff and Peskin [7] for the mean-square displacements in different directions for different sets of particles, before the

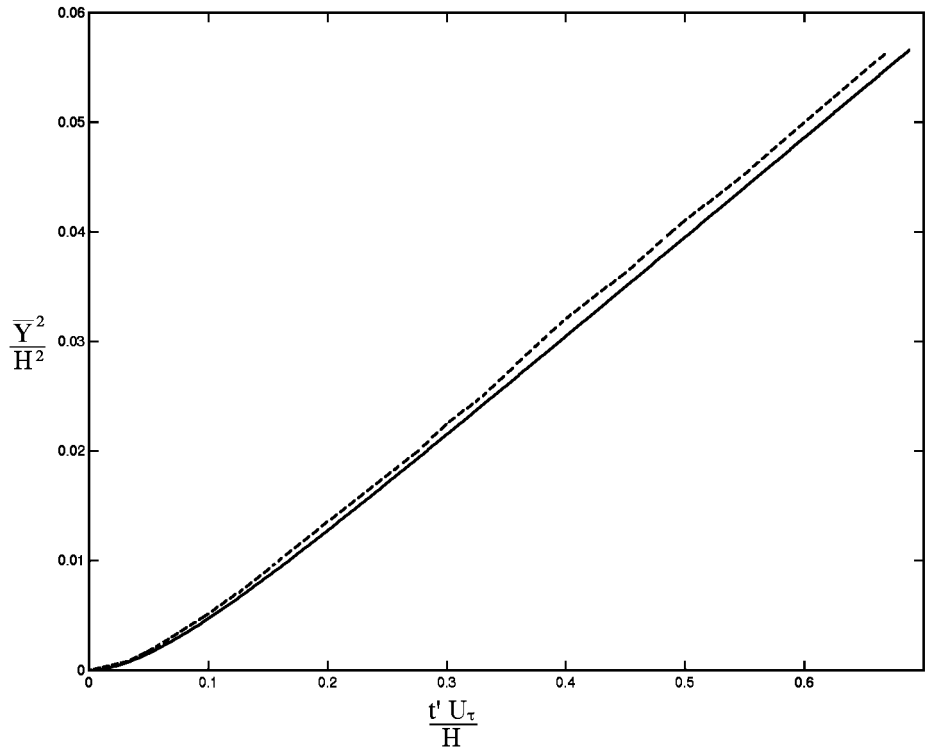


FIG. 8. *Mean-square displacement of the fluid particles by turbulent flow.* (—) *Current theory with Kolmogorov velocity spectrum and* $\beta = 1$. (- -) *Numerical experiment of Deardorff and Peskin* [7]*, mean-square displacement in the lateral direction of a shear flow in a plane channel.*

final averaging.

In Figure 8 the solid line again corresponds to the prediction of the present theory, and the dashed line now represents the Deardorff and Peskin [7] numerical result for the mean-square displacement in the lateral direction with no additional rescaling for the adjustment of the root-mean-square to friction velocity. The agreement again is quite good. Hence, one can expect that the above renormalization theory to yield satisfactory prediction from the spectral data for the transport properties of nearly isotropic random flows.

Appendix A. Distribution of convective trajectories (5.4). First of all, let us establish some statistical properties for the Lagrangian velocity field $\mathbf{U}(t) = -\mathbf{u}(\sqrt{2D_0}\mathbf{W}(t) + \mathbf{Z}(\mathbf{t}), t)$. Invoking the representation of the Eulerian velocity $\mathbf{u}(\mathbf{x}, t)$ by its spatial Fourier transform, one can represent $\mathbf{U}(t)$ by

$$\mathbf{U}(t) = -(2\pi)^{-m/2} \int_{-\infty}^{\infty} \hat{\mathbf{u}}(\mathbf{k}, t) g_w(\mathbf{k}, t) f(\mathbf{k}, t) d^m\mathbf{k}, \tag{A.1}$$

where

$$g_w(\mathbf{k}, t) = \exp(\imath\sqrt{2D_0}\mathbf{k}\mathbf{W}(t)),$$
$$f(\mathbf{k}, t) = \exp(\imath\mathbf{k}\mathbf{Z}(t)) = \exp\left(\imath\mathbf{k}\int_0^t \mathbf{U}(s)ds\right),$$

and the value of $\langle f(\delta\mathbf{K},t)\rangle_{U|W}$ gives the characteristic function of the distribution of "convective" trajectories $\mathbf{Z}(t)$. The product of any N number of the components of $\mathbf{U}(t)$ is then given by

$$\prod_{n=1}^{N} U_{i_n}(t) = \frac{(-1)^N}{(2\pi)^{\frac{mN}{2}}} \int g_w\left(\sum_{n=1}^{N}\mathbf{k}^n,t\right) f\left(\sum_{n=1}^{N}\mathbf{k}^n,t\right) \prod_{n=1}^{N} \hat{u}_{i_n}(\mathbf{k}^n,t)d^m\mathbf{k}^n. \tag{A.2}$$

Hence, in order to express any analytic function of the Lagrangian velocity $\mathbf{U}(t)$ in terms of known random functions $\hat{\mathbf{u}}$ and $\mathbf{W}(t)$, one needs to find such representation for $f(\mathbf{k},t)$.

Time-evolution of $f(\mathbf{k},t)$ can be described by the equation

$$df = \imath\mathbf{k}\mathbf{U}(t)f(\mathbf{k},t)dt, \quad f(\mathbf{k},0)=1,$$

and formal integration with $\mathbf{U}(t)$ given by (A.1) yields

$$f(\mathbf{k},t) = 1 - \int_0^t \int_{-\infty}^{\infty} \imath\mathbf{k}\hat{\mathbf{u}}(\mathbf{q},s)f(\mathbf{k}+\mathbf{q},s)g_w(\mathbf{q},s)\frac{d^m\mathbf{q}}{(2\pi)^{m/2}}ds. \tag{A.3}$$

Equation (A.3) is a linear integral equation for f with random but continuous and, in the mean-square sense, bounded kernel. It should be recalled that $\hat{\mathbf{u}}(\mathbf{q},s)$ is defined on the domain $\delta < q < 1$ and the dispersion of $\hat{\mathbf{u}}$, as is given by (3.3), is bounded for any finite δ. The probability for $\hat{\mathbf{u}}(\mathbf{q},s)$ to exceed essentially $\langle\hat{\mathbf{u}}\hat{\mathbf{u}}\rangle^{1/2}$ is really small because a Gaussian distribution has an exponentially small tail. Consequently, one can expect that the analogue of the Neumann expansion for (A.3) will converge, at least for some finite values of t and δ.

Hence, by iterative substitution of $f(\mathbf{k}+\mathbf{q},s)$ into (A.3) one can write the series solution for f in terms of $\hat{u}$ and $\mathbf{W}$:

$$\begin{aligned} f(\mathbf{k},t) &= 1 - \int_0^t \int_{-\infty}^{\infty} \imath k_{j_1}\hat{u}_{j_1}(\mathbf{q}^1,s_1)g_w(\mathbf{q}^1,s_1)\frac{d^m\mathbf{q}^1}{(2\pi)^{\frac{m}{2}}}ds_1 \\ &- \int_0^t ds_1 \int_0^{s_1} ds_2 \int_{-\infty}^{\infty} k_{j_1}\hat{u}_{j_1}(\mathbf{q}^1,s_1)(k+q^1)_{j_2}\hat{u}_{j_2}(\mathbf{q}^2,s_2) \\ &\qquad \times g_w(\mathbf{q}^1,s_1)g_w(\mathbf{q}^2,s_2)\frac{d^m\mathbf{q}^1 d^m\mathbf{q}^2}{(2\pi)^m} + \cdots \\ &+ \frac{(-\imath)^p}{(2\pi)^{\frac{mp}{2}}} \int_{-\infty}^{\infty} \prod_{k=1}^{p} \int_0^{s_{k-1}} \left(\sum_{l=0}^{k-1} q_{j_k}^l\right) \hat{u}_{j_k}(\mathbf{q}^k,s_k)g_w(\mathbf{q}^k,s_k)d^m\mathbf{q}^k ds_k + \cdots, \end{aligned} \tag{A.4}$$

where $s_0 = t$ and $\mathbf{q}^0 = \mathbf{k}$ in the representation of general terms in (A.4). The convergence of this Neumann series is established in Appendix B.

The substitution of (A.4) into (A.2) and averaging over $\hat{u}$ immediately gives that all terms in the series, except the first one (which is equal to 1), vanish because of the homogeneity of the Eulerian velocity field $\mathbf{u}$ and the continuity equation. Indeed, with $\mathbf{q}^0 = \mathbf{k} = \sum_{m=1}^{N}\mathbf{k}^m$ in (A.4), the general "pth" term in the product of "N" U-components in (A.2) becomes

$$\cdots \prod_{n=1}^{N} \hat{u}_{i_n}(\mathbf{k}^n,t) \prod_{k=1}^{p} \left(\sum_{m=1}^{N} k_{j_k}^m + \sum_{l=1}^{k-1} q_{j_k}^l\right) \hat{u}_{j_k}(\mathbf{q}^k,s_k)\cdots, \tag{A.5}$$

where only dependencies on $\hat{u}$ and wavevectors are shown. The homogeneity requirement implies that only terms with

$$\sum_{m=1}^{N} \mathbf{k}^m + \sum_{l=1}^{p} \mathbf{q}^l = 0$$

can give nonzero average (compare with the $\delta(\mathbf{k}+\mathbf{q})$ term in the spectral form of velocity correlation (3.3)) and, consequently, one can set

$$\sum_{m=1}^{N} \mathbf{k}^m + \sum_{l=1}^{p-1} \mathbf{q}^l = -\mathbf{q}^p$$

in (A.5). The last two factors in the product (A.5) then become

$$\left(\sum_{m=1}^{N} k^m + \sum_{l=1}^{p-1} q^l \right)_{j_p} \hat{u}_{j_p}(\mathbf{q}^p, s_p) = -q^p_{j_p} \hat{u}_{j_p}(\mathbf{q}^p, s_p) \equiv 0$$

because of the continuity equation. Hence, only $f_0(\mathbf{k}, t) = 1$ contributes to the average of (A.2), and the homogeneity restriction for the remaining term, $\sum_{n=1}^{N} \mathbf{k}^n = 0$, also leads to the disappearance of the dependence on $\mathbf{W}(t)$,

(A.6)
$$\begin{aligned}
\left\langle \prod_{n=1}^{N} U_{i_n}(t) \right\rangle &= \frac{(-1)^N}{(2\pi)^{\frac{mN}{2}}} \int \exp\left(\imath \sqrt{2D_0} \sum_{n=1}^{N} \mathbf{k}^n \mathbf{W}(t) \right) \left\langle \prod_{n=1}^{N} \hat{u}_{i_n}(\mathbf{k}^n, t) \right\rangle d^m\mathbf{k}^n \\
&= \frac{(-1)^N}{(2\pi)^{\frac{mN}{2}}} \int \left\langle \prod_{n=1}^{N} \hat{u}_{i_n}(\mathbf{k}^n, t) \right\rangle d^m\mathbf{k}^n = (-1)^N \left\langle \prod_{n=1}^{N} u_{i_n}(0, t) \right\rangle \\
&= (-1)^N \left\langle \prod_{n=1}^{N} u_{i_n}(\mathbf{x}, T-t) \right\rangle = (-1)^N \left\langle \prod_{n=1}^{N} u_{i_n}(0, 0) \right\rangle.
\end{aligned}$$

The last two identities follow from the original definition of the inverse trajectory and from the stationarity of the Eulerian velocity field $\mathbf{u}(\mathbf{x}, t)$.

Because any analytic function of the Lagrangian velocity can be represented by a power series in the velocity components, it follows from (A.6) that

$$\langle \phi(\mathbf{U}(s)) \rangle_U = \langle \phi(-\mathbf{u}(\mathbf{x}, t-s)) \rangle_u$$

for any arbitrary analytic function $\phi(\mathbf{U}(s))$. This implies that $\mathbf{U}(s)$ and $-\mathbf{u}(\mathbf{x}, t-s)$ have the same distribution. Because $\mathbf{u}$ is assumed Gaussian, $\mathbf{U}(t)$ and the convective trajectory $\mathbf{Z}(t)$, which is linear in $\mathbf{U}$, also are Gaussian. Consequently, equation (5.4),

(A.7)
$$\begin{aligned}
&\langle \exp(\imath\delta\mathbf{K}\mathbf{Z}) \rangle_{Z(u|W)} = \exp\left(-\frac{\delta^2}{2} \langle (\mathbf{K}\mathbf{Z})^2 \rangle_{Z(u|W)} \right) = \exp\left(-\frac{\delta^2}{2} K_i \tilde{Z^2_{ij}} K_j \right), \\
&\tilde{Z^2_{ij}} = \tilde{Z^2_{ij}}(t, \mathbf{W}(t)) = \langle Z_i(t) Z_j(t) \rangle_{\hat{u}} = \int_0^t \int_0^t \langle \mathbf{U}_i(s) \mathbf{U}_j(s') \rangle_U ds ds',
\end{aligned}$$

provides the exact representation for the characteristic function $\langle \exp(\imath\delta\mathbf{K}\mathbf{Z}) \rangle$ at any time t for the selected stationary homogeneous isotropic Gaussian Eulerian velocity field $\mathbf{u}$.

Appendix B. Convergence of Neumann series (A.4) for the characteristic function. For all the results in Appendix A to be correct, one needs to establish the convergence of the series (A.4) for $f(\mathbf{k},t)$. In fact, it is enough to estimate the average of a general even term, $|\langle f(\mathbf{k},t)_{2p}\rangle|$. The absolute convergence of the series $\langle f(\mathbf{k},t)_{2p}\rangle$ will then guarantee the convergence of (A.4) in the mean-square sense because of the structure of even and odd terms and because of the previously discussed fast decay of the tail of Gaussian distribution. Hence, let us consider the detailed structure of $|\langle f(\mathbf{k},t)_{2p}\rangle|$,

$$
\begin{aligned}
|\langle f(\mathbf{k},t)_{2p}\rangle| = \Bigg| & \frac{(-1)^p}{(2\pi)^{mp}} \int_0^t ds_1 \int_0^{s_1} ds_2 \ldots \int_0^{s_{2p-1}} ds_{2p} \int_{-\infty}^{\infty} d^m\mathbf{q}^1 d^m\mathbf{q}^2 \ldots d^m\mathbf{q}^{2p} \\
& \times \exp\left(\imath\sqrt{2D_0}\sum_{i=1}^{2p}\mathbf{q}^i\mathbf{W}(s_i)\right) \\
& \times k_{j_1}(k+q^1)_{j_2}(k+q^1+q^2)_{j_3}\ldots\left(k+\sum_{i=1}^{2p-1}q^i\right)_{j_{2p}} \\
& \times \langle \hat{u}_{j_1}(\mathbf{q}^1,s_1)\hat{u}_{j_2}(\mathbf{q}^2,s_2)\hat{u}_{j_3}(\mathbf{q}^3,s_3)\ldots\hat{u}_{j_{2p}}(\mathbf{q}^{2p},s_{2p})\rangle\Bigg|.
\end{aligned}
\tag{B.1}
$$

To calculate the average in the last line of (B.1), one can invoke the factorization property of high-order moments of Gaussian distribution

$$
\underbrace{\langle \hat{u}_i\hat{u}_j\hat{u}_k\hat{u}_m\ldots\rangle}_{2p} = \frac{2p!}{p!2^p}\left(\underbrace{\langle\hat{u}_i\hat{u}_j\rangle\langle\hat{u}_k\hat{u}_m\rangle\ldots}_{p\text{ 2-nd moments}}\right)_{\text{sym}},
\tag{B.2}
$$

where the subscript "sym" denotes the arithmetic mean of all symmetrized products of the $\hat{u}_i\hat{u}_j$. For example, the fourth-order moment is

$$
\langle \hat{u}_i\hat{u}_j\hat{u}_k\hat{u}_m\rangle = \frac{4!}{2!2^2}\left(\frac{1}{3}\left[\langle\hat{u}_i\hat{u}_j\rangle\langle\hat{u}_k\hat{u}_m\rangle + \langle\hat{u}_i\hat{u}_k\rangle\langle\hat{u}_j\hat{u}_m\rangle + \langle\hat{u}_i\hat{u}_m\rangle\langle\hat{u}_j\hat{u}_k\rangle\right]\right).
\tag{B.3}
$$

With the factorization (B.2) and the correlation of the Fourier-component of the velocity field given by (3.3),

$$
\begin{aligned}
&\langle \hat{u}_{j_i}(\mathbf{q}^i,s_i)\hat{u}_{j_n}(\mathbf{q}^n,s_n)\rangle = \\
&\qquad \delta(\mathbf{q}^i+\mathbf{q}^n)\left[\delta_{j_ij_n} - \frac{q^i_{j_i}q^i_{j_n}}{(q^i)^2}\right]\alpha^2(q^i)^{1-m-\epsilon}\exp(-a(q^i)^z|s_i-s_n|)
\end{aligned}
\tag{B.4}
$$

for $\delta \le q^i \equiv |\mathbf{q}^i| = |\mathbf{q}^n| \le 1$; $\langle \hat{u}_{j_i}(\mathbf{q}^i,s_i)\hat{u}_{j_n}(\mathbf{q}^n,s_n)\rangle \equiv 0$ otherwise, (B.1) becomes

$$|\langle f(\mathbf{k},t)_{2p}\rangle| = \frac{\alpha^{2p}}{(2\pi)^{mp}}\frac{2p!}{p!2^p}\prod_{k=1}^{p}\int_1^{\delta}(\tilde{q}^k)^{-\epsilon}d\tilde{q}^k\int_0^t ds_1\int_0^{s_1} ds_2\ldots\int_0^{s_{2p-1}} ds_{2p}$$

$$\times\Bigg|\Bigg\{\exp\left(-a\sum_{i=1}^{p}(\tilde{q}^i)^z\triangle_i s\right)\int_{\text{angles}}\cdots\int\exp\left(\imath\sqrt{2D_0}\sum_{i=1}^{2p}\tilde{\mathbf{q}}^i\triangle_i\mathbf{W}\right)$$

$$\times k_{j_1}(k+q^1)_{j_2}(k+q^1+q^2)_{j_3}\ldots\left(k+\sum_{i=1}^{2p-1}q^i\right)_{j_{2p}} \tag{B.5}$$

$$\times\underbrace{\left[\ldots\delta(\mathbf{q}^l+\mathbf{q}^n)\left(\delta_{j_l j_n}-\frac{q^l_{j_l}q^l_{j_n}}{(q^l)^2}\right)\ldots\right]}_{p\text{ factors}}\Bigg\}_{\text{sym}}\Bigg|,$$

where the symbolic notations $\triangle_i s$, $\triangle_i\mathbf{W}$, and $\tilde{\mathbf{q}}^i$ correspond to the ordering of terms in the symmetrized product

$$\tilde{\mathbf{q}}^1=\mathbf{q}^1,\quad \triangle_1 s=s_1-s_k,\quad \triangle_1\mathbf{W}=\mathbf{W}(s_1)-\mathbf{W}(s_k),$$

$$\tilde{\mathbf{q}}^2=\begin{cases}\mathbf{q}^2, & \triangle_2 s=s_2-s_i,\quad \triangle_2\mathbf{W}=\mathbf{W}(s_2)-\mathbf{W}(s_i) \quad \text{for } k\neq 2,\\ \mathbf{q}^3, & \triangle_3 s=s_3-s_i,\quad \triangle_3\mathbf{W}=\mathbf{W}(s_3)-\mathbf{W}(s_i) \quad \text{for } k=2,\end{cases}$$

and so on.

It is evident that the homogeneity requirement and the fluid incompressibility (product of p delta-functions and $[\delta_{j_l j_n}-q^l_{j_l}q^l_{j_n}/(q^l)^2]$ factors, respectively, in the last line of (B.5)) effectively eliminate most of the "q" terms in the third line of (B.5). The symmetry condition of the velocity field, however, cannot be invoked completely until Brownian motion effect is averaged or neglected. By explicit consideration of angular integration for the fourth and sixth moments (they are relatively short but already contain all representative combinations of even and odd powers of $\tilde{q}^i$), it can be shown that interaction with nonaveraged effect of molecular diffusion leads to factors $|\tilde{g}^i_w|<1$, like

$$\tilde{g}^i_w=\frac{\sin(\sqrt{2D_0}\tilde{q}^i|\triangle_i\mathbf{W}|)}{\sqrt{2D_0}\tilde{q}^i|\triangle_i\mathbf{W}|}\quad\text{and /or}\quad \tilde{g}^i_w\frac{\triangle_i\mathbf{W}}{|\triangle_i\mathbf{W}|},\quad \tilde{g}^m_w\tilde{g}^i_w\frac{\triangle_i\mathbf{W}+\triangle_m\mathbf{W}}{|\triangle_i\mathbf{W}+\triangle_m\mathbf{W}|}$$

in the 3-D case, for example. More importantly, the first one with a nonzero average always corresponds to terms with even powers in $\tilde{q}^i$ in the third line of (B.5). The others, which contain an odd power of $\mathbf{W}$ and hence will vanish after averaging, always appear only if odd powers in $\tilde{q}^i$ are present, i.e., they correspond to terms like $k_j\tilde{q}^i_j$ and $\tilde{q}^i_j\tilde{q}^m_j$ with $i\neq m$ and their odd powers in (B.5). Hence, the interaction of random symmetric convection with molecular diffusion does not induce net coupling between nonsymmetric terms $\tilde{q}^i_j\tilde{q}^m_j$ with $i\neq m$ in (B.5) and it cannot lead to net nonsymmetric effects.

Consequently, one can set $|\exp(\imath\sqrt{2D_0}\dots)| = 1$ in (B.5) before the angular integration. As a result, the absolute value of the integral over angles in (B.5) yields

$$\left| \int_{\text{angles}} \cdots \int \dots \right| \le C^p k^{2p}, \tag{B.6}$$

where the numerical constant C depends on the space dimension but is independent on p. Now, because $\exp\left(-a\sum_{i=1}^{p}(\tilde{q}^i)^z \triangle_i s\right)$ is always less than 1, one can replace it by 1 in the second line of (B.5) and complete the integration over s and $\tilde{q}^i$. The resulting estimate for $|\langle f(\mathbf{k},t)_{2p}\rangle|$ is

$$\begin{aligned} |\langle f(\mathbf{k},t)_{2p}\rangle| &\le \frac{\alpha^{2p}}{(2\pi)^{mp}} \frac{2p!}{p!2^p} \left(\frac{\delta^{1-\epsilon}-1}{\epsilon-1}\right)^p \frac{t^{2p}}{2p!} C^p k^{2p} \\ &= \frac{1}{p!}\left[\frac{C(kt\alpha)^2}{2(2\pi)^m}\left(\frac{\delta^{1-\epsilon}-1}{\epsilon-1}\right)\right]^p. \end{aligned} \tag{B.7}$$

With the Stirling's formula, $p! \approx p^p e^{-p}\sqrt{2\pi p}$, it is evident that for any fixed $t < \infty$, $k < \infty$, and $\delta > 0$

$$|\langle f(\mathbf{k},t)_{2p}\rangle|^{1/p} \le \frac{e}{p}\left[\frac{C(kt\alpha)^2}{2(2\pi)^m}\left(\frac{\delta^{1-\epsilon}-1}{\epsilon-1}\right)\right] \to 0 \text{ as } p \to \infty. \tag{B.8}$$

Hence, according to the Cauchy criteria, the average of series (A.4), $\langle f(\mathbf{k},t)\rangle = \langle f(\mathbf{k},t)_{2p}\rangle$, converges absolutely, regardless of the value of the exponent ϵ. Because of the exponentially small tail of the Gaussian distribution, this leads to the convergence of the unaveraged series (A.4).

It should be mentioned that (B.7) strongly overestimates $|\langle f(\mathbf{k},t)_{2p}\rangle|$ at large values of t because $\exp\left(-a\sum_{i=1}^{p}(\tilde{q}^i)^z \triangle_i s\right)$ in (B.5) has been set equal to 1, which corresponds to small-time behavior. Hence, it is not surprising that larger and larger values of p are needed in order for the right-hand side of (B.7) to decay in the large-scale, long-time limit for arbitrary values of ϵ. However, this does not yet mean that the actual convergence of (A.4) is not as rapid at large time. In fact, one of the goals of the subsequent renormalization analysis is to reach the uniform convergence of (A.4) in the above limit. In any case, (B.8) is valid for any finite $t < \infty$, $k < \infty$, and $\delta > 0$, and it is valid to replace the characteristic function with series (A.4).

Appendix C. Effective diffusivity tensor (5.7). It should be emphasized that establishing that the trajectories are Gaussian distributed tells us only that the characteristic function can be expressed through the mean-square displacement tensor alone as it is given by (A.7) (which is the same as (5.4)), but it does not yet provide the expression for $\tilde{Z}_{ij}^2$. It is evident that (A.7) involves the two-time correlation $\langle \mathbf{U}_i(s)\mathbf{U}_j(s')\rangle_U$, and only relations among the single-time moments $\langle \prod_{n=1}^{N} U_{i_n}(t)\rangle$ and $\langle \prod_{n=1}^{N} u_{i_n}(0,t)\rangle$ have been established by (A.6).

While estimate (B.7) for $|\langle f(\mathbf{k},t)_{2p}\rangle|$ is quite straightforward, it is difficult or even can be impossible to obtain the exact expression for $\tilde{Z}_{ij}^2$ by directly averaging and summing series (A.4) for $f(\delta\mathbf{K},t)$. Using the averaged version of (A.4) instead of (A.6) for any practical purposes also is difficult since (B.7) and (B.8) strongly overestimate the number of terms that is required for the convergence of the series (A.4) in the long-time limit. Hence, another approach will be used to obtain the

evolution equation (5.7) for the effective diffusivity $\tilde{D_{ij}}(s)$, which is related to $\tilde{Z_{ij}^2}$ by definition (5.6),

$$\frac{1}{2}\tilde{Z_{ij}^2}(t,\mathbf{W}(t)) = \int_0^t \tilde{D_{ij}}(s)ds. \tag{C.1}$$

Note that $\mathbf{Z}(t)$ is Gaussian, $\mathbf{Z}(0) = 0$, and $\langle \mathbf{Z}(t)\rangle_U = 0$. Then one can introduce the appropriate m-dimensional white noise $\mathbf{z}(t)$ with the usual autocorrelation properties, $dz_n(t)dz_k(t) = \delta_{nk}dt$ and $dz_n(t)dz_k(s) = 0$ for $t \neq s$, such that $\mathbf{Z}(t)$ can be represented by Ito's stochastic integral

$$Z_i(t) = \int_0^t U_i(s')ds' = \int_0^t A_{in}(s,\mathbf{W}(s))dz_n(s), \tag{C.2}$$

where the matrix $A_{in}(s,\mathbf{W}(s))$ does not depend explicitly on $\mathbf{z}(t)$. It, however, depends on $\hat{\mathbf{u}}$. The Gaussian random processes $\mathbf{z}(t)$ and $\mathbf{W}(t)$, as well as $\mathbf{z}(t)$ and $\int \ldots \hat{\mathbf{u}}(\mathbf{k},t)d^m\mathbf{k}$, also are dependent in general. For example, the dependence of $\mathbf{A}$ on $\hat{\mathbf{u}}$ is evident from relation (A.6) among the statistical properties of Eulerian and Lagrangian velocities that suggests $\mathbf{A} \sim \langle \mathbf{u}(0,0)\mathbf{u}(0,0)\rangle^{1/2}$. The general dependence, of course, is not limited by this simple relation.

However, according to the definition of Ito's stochastic integral, $\mathbf{z}(s)$, $\mathbf{W}(s)$, and $\hat{\mathbf{u}}(\mathbf{k},s)$ are statistically independent of $d\mathbf{z}(t)$ for $t \geq s$ (this also is true for $d\mathbf{W}(t)$). It should be emphasized that only independence on $d\mathbf{z}(t)$ is implied, but not on $\mathbf{z}(t)$; for example, $\langle \hat{u}_i(\mathbf{k},s)dz_j(t)\rangle = 0$ for $t \geq s$, but $\langle \hat{u}_i(\mathbf{k},s)z_j(t)\rangle \neq 0$ in general, regardless of the relative values of t and s. Physically, this means that all these physical quantities at time s "do not know what happens in the future," i.e., how the white noise $\mathbf{z}$ will change at the next instant in time. Hence, from definitions (C.1) and (C.2) and the above properties of $d\mathbf{z}$, it follows that because

$$\begin{aligned}\tilde{Z_{ij}^2}(t) = \langle Z_i(t)Z_j(t)\rangle_{\hat{u}} &= \left\langle \int_0^t A_{in}(s,\mathbf{W}(s))dz_n(s) \int_0^t A_{kj}(s',\mathbf{W}(s'))dz_k(s') \right\rangle_{\hat{u}} \\ &= \int_0^t \langle A_{in}(s,\mathbf{W}(s))A_{nj}(s,\mathbf{W}(s))\rangle_{\hat{u}}ds = 2\int_0^t \tilde{D_{ij}}(s)ds,\end{aligned}$$

the two-time correlation $\langle Z_i(t_1)Z_j(t_2)\rangle_{\hat{u}}$ becomes

$$\begin{aligned}\langle Z_i(t_1)Z_j(t_2)\rangle_{\hat{u}} &= \int_0^{\min\{t_1,t_2\}} \langle A_{in}(s,\mathbf{W}(s))A_{nj}(s,\mathbf{W}(s))\rangle_{\hat{u}}ds \\ &= 2\int_0^{\min\{t_1,t_2\}} \tilde{D_{ij}}(s)ds = \tilde{Z_{ij}^2}(\min\{t_1,t_2\}).\end{aligned} \tag{C.3}$$

Another important consequence of (C.2) and these properties of $d\mathbf{z}$ is the statistical independence of both $\mathbf{Z}(s)$ and $\hat{\mathbf{u}}(\mathbf{q},s)$ on

$$\begin{aligned}\triangle^Z(\mathbf{k}, t \geq t') = k_m(Z_m(t) - Z_m(t')) &= k_m \int_{t'}^t A_{mn}(s',\mathbf{W}(s'))dz_n(s') \\ &= k_m \int_{t'}^t U_m(s'')ds''\end{aligned} \tag{C.4}$$

for any $\mathbf{k}$, $\mathbf{q}$, and $t \geq t' \geq s$.

It also should be noted that the two-point two-time Eulerian velocity correlation (3.3) (same as (B.4)) yields

$$\begin{aligned}\langle \hat{u}_i(\mathbf{k},t)\hat{u}_j(\mathbf{q},s)\rangle &= \exp(-ak^z|t-s|)\langle \hat{u}_i(\mathbf{k},s)\hat{u}_j(\mathbf{q},s)\rangle \\ &= \exp(-ak^z|t-s|)\langle \hat{u}_i(\mathbf{k},0)\hat{u}_j(\mathbf{q},0)\rangle\end{aligned} \tag{C.5}$$

such that

$$\langle [\hat{u}_i(\mathbf{k},t) - \exp(-ak^z(t-s))\hat{u}_i(\mathbf{k},s)]\,\hat{u}_j(\mathbf{q},s)\rangle \equiv 0 \tag{C.6}$$

for any pairs of $\{\mathbf{k},\mathbf{q}\}$, $\{i,j\}$, and $t \geq s$. For further analysis, it is convenient to introduce the notation

$$\triangle_i^u(\mathbf{k}, t \geq t') = \hat{u}_i(\mathbf{k},t) - \exp(-ak^z(t-t'))\hat{u}_i(\mathbf{k},t'), \tag{C.7}$$

where, similar to (C.4), t should be $\geq t'$. Using (C.5) and definition (C.7), it can be easily shown that

$$\langle \triangle_i^u(\mathbf{k}, t \geq t')\hat{u}_j(\mathbf{q},s)\rangle \equiv 0 \tag{C.8}$$

for any $\{\mathbf{k},\mathbf{q}\}$, $\{i,j\}$, and $t \geq t' \geq s$.

This means that $\hat{\mathbf{u}}(\mathbf{q},s)$ is statistically independent of $\triangle^{\mathbf{u}}(\mathbf{k}, t \geq t')$ for $t' \geq s$. According to (C.2), $\mathbf{Z}(s)$ also can be defined through the Lagrangian velocity $\mathbf{U}(s')$ with $s' \leq s$, which, in turn, depends only on $\hat{\mathbf{u}}(\mathbf{q}^l, s_l)$ at different instants of time $s_l \leq s'$ because of the series solution (A.4). Consequently, $\mathbf{Z}(s)$ also is statistically independent of $\triangle^{\mathbf{u}}(\mathbf{k}, t \geq t')$ for $t' \geq s$. It is hence useful to summarize the established independence properties:

$$\langle \triangle_i^u(\mathbf{k}, t \geq t')\hat{u}_j(\mathbf{q},s)\rangle \equiv 0, \quad \langle \triangle_i^u(\mathbf{k}, t \geq t')Z_j(s)\rangle \equiv 0, \tag{C.9}$$

$$\begin{aligned}&\langle \triangle^Z(\mathbf{k}, t \geq t')\hat{u}_j(\mathbf{q},s)\rangle \equiv 0, \quad \langle \triangle^Z(\mathbf{k}, t \geq t')Z_j(s)\rangle \equiv 0 \\ &\text{for any} \quad \{\mathbf{k},\mathbf{q}\}, \quad \{i,j\}, \quad \text{and} \quad t \geq t' \geq s,\end{aligned} \tag{C.10}$$

where $\triangle^Z$ and $\triangle_i^u$ are defined by (C.4) and (C.7), respectively.

Now let us consider the two-time correlation for the Lagrangian velocity components

$$\begin{aligned}\langle U_i(t)U_j(s)\rangle_{\hat{u}} = (2\pi)^{-m} \iint \exp\left[\imath\sqrt{2D_0}(\mathbf{k}\mathbf{W}(t) + \mathbf{q}\mathbf{W}(s))\right] \\ \times \langle \hat{u}_i(\mathbf{k},t)\hat{u}_j(\mathbf{q},s)\exp\left[\imath(\mathbf{k}\mathbf{Z}(t) + \mathbf{q}\mathbf{Z}(s))\right]\rangle_{\hat{u}} d^m\mathbf{k}d^m\mathbf{q},\end{aligned} \tag{C.11}$$

where, without loss of generality, one can set $t \geq s$. The next step is to find the proper decomposition for the average in the last line of (C.11) in order to complete the averaging without obtaining the explicit solution for $\mathbf{Z}$. Using definitions (C.4) and (C.7) and the independence properties (C.9)–(C.10), one obtains

$$\begin{aligned}&\langle \hat{u}_i(\mathbf{k},t)\hat{u}_j(\mathbf{q},s)\exp\left[\imath(\mathbf{k}\mathbf{Z}(t) + \mathbf{q}\mathbf{Z}(s))\right]\rangle \\ &= \langle \{\triangle_i^u(\mathbf{k}, t \geq s) + \exp(-ak^z(t-s))\hat{u}_i(\mathbf{k},s)\}\,\hat{u}_j(\mathbf{q},s) \\ &\quad \times \exp\left[\imath\left\{\triangle^Z(\mathbf{k}, t \geq s) + (\mathbf{k}+\mathbf{q})\mathbf{Z}(s)\right\}\right]\rangle \\ &= \langle \triangle_i^u(\mathbf{k}, t \geq s)\exp\left[\imath\triangle^Z(\mathbf{k}, t \geq s)\right]\rangle\langle \hat{u}_j(\mathbf{q},s)\exp\left[\imath(\mathbf{k}+\mathbf{q})\mathbf{Z}(s)\right]\rangle \\ &+ \langle \exp\left[\imath\triangle^Z(\mathbf{k}, t \geq s)\right]\rangle \exp(-ak^z(t-s))\langle \hat{u}_i(\mathbf{k},s)\hat{u}_j(\mathbf{q},s)\exp\left[\imath(\mathbf{k}+\mathbf{q})\mathbf{Z}(s)\right]\rangle \\ &\equiv \{1\} + \{2\}.\end{aligned} \tag{C.12}$$

The second term $\{2\}$ in (C.12) can now be averaged explicitly. One can use two-time correlation (C.3) for $\mathbf{Z}$ in the first average, which involves only $\triangle^Z(\mathbf{k}, t \geq s) = \mathbf{k}(\mathbf{Z}(t) - \mathbf{Z}(s))$. The second $\langle \dots \rangle$ in $\{2\}$, which contains only s-dependent quantities, is exactly the same as $\langle \dots \rangle$ in the average of (A.2) for the single-time second moment of Lagrangian velocity $\langle U_i(s) U_j(s) \rangle_u$. Consequently, (A.6) and the Eulerian velocity correlation (B.4) allow the completion of this averaging. The result is

$$\begin{aligned}\{2\} &= \exp\left[-\frac{1}{2} k_m \left\{\tilde{Z}^2_{mn}(t) - \tilde{Z}^2_{mn}(s)\right\} k_n\right] \exp(-ak^z(t-s)) \langle \hat{u}_i(\mathbf{k}, 0) \hat{u}_j(\mathbf{q}, 0) \rangle \\ &= \delta(\mathbf{k}+\mathbf{q}) \alpha^2 k^{1-m-\epsilon} \left(\delta_{ij} - \frac{k_i k_j}{k^2}\right) \exp\left[-ak^z(t-s) - k_m \int_s^t \tilde{D}_{mn}(s') ds' k_n\right],\end{aligned} \tag{C.13}$$

$$k = |\mathbf{k}|, \quad \delta < k < 1,$$

where the definition of $\tilde{D}_{mn}$ in (C.1) also has been used. Note that (C.13) gives exactly the same result as that from invoking Corrsin's independence hypothesis, namely, the independent averaging of $\langle \hat{u}\hat{u} \rangle$ and $\langle \exp[\dots] \rangle$ in (C.11). The decomposition (C.12) for the averaging in (C.11), however, also contains the term $\{1\}$, which will be considered below.

The second average in $\{1\}$, which contains all $\mathbf{q}$-dependent factors and quantities, and depends only on s but not t, is not of great interest. Including $\exp[\imath \mathbf{q} \mathbf{W}(s)]$ from the first line of (C.11) (which remains invariant under the averaging over $\hat{u}$) into this $\langle \dots \rangle$ and integrating over $\mathbf{q}$ reduce it to $\langle U_j(s) \exp[\imath \mathbf{k} \mathbf{Z}(s)] \rangle_U$, which is some nonzero function of s and $\mathbf{k}$ in general.

Now let us examine the first average in $\{1\}$, which depends only on wavevector $\mathbf{k}$ and "differences" $\triangle^u_i$ and $\triangle^Z$, $\langle \triangle^u_i(\mathbf{k}, t \geq s) \exp\left[\imath \triangle^Z(\mathbf{k}, t \geq s)\right] \rangle$. Using the definition of $\triangle^Z$ by (C.4), one can write the stochastic differential equations (with respect to variable t) for the exponential factor

$$\begin{aligned} f^{\triangle}(\mathbf{k}, t \geq s) &= \exp\left[\imath \triangle^Z(\mathbf{k}, t \geq s)\right] \\ &= \exp\left[\imath \mathbf{k}(\mathbf{Z}(t) - \mathbf{Z}(s))\right] = \exp\left[\imath \mathbf{k} \int_s^t \mathbf{U}(s') ds'\right] \end{aligned} \tag{C.14}$$

both in terms of Lagrangian velocity and the new white noise $z(t)$:

$$df^{\triangle}(\mathbf{k}, t \geq s) = \imath \mathbf{k} \mathbf{U}(t) f^{\triangle}(\mathbf{k}, t \geq s) dt, \tag{C.15}$$

$$df^{\triangle}(\mathbf{k}, t \geq s) = \imath \mathbf{k} \mathbf{A}(t) f^{\triangle}(\mathbf{k}, t \geq s) d\mathbf{z}(t) - \mathbf{k} \tilde{\mathbf{D}}(t) \mathbf{k} f^{\triangle}(\mathbf{k}, t \geq s) dt. \tag{C.16}$$

Now, similar to (A.3), let us formally integrate (C.15), multiply it by $\triangle^{\mathbf{u}}(\mathbf{k}, t \geq s)$, and average the result:

$$\begin{aligned} \langle \triangle^{\mathbf{u}}(\mathbf{k}, t \geq s) f^{\triangle}(\mathbf{k}, t \geq s) \rangle &= \langle \triangle^{\mathbf{u}}(\mathbf{k}, t \geq s) \rangle \\ &\quad + \left\langle \triangle^{\mathbf{u}}(\mathbf{k}, t \geq s) \int_s^t \imath \mathbf{k} \mathbf{U}(t') f^{\triangle}(\mathbf{k}, t' \geq s) dt' \right\rangle \\ &= \left\langle \triangle^{\mathbf{u}}(\mathbf{k}, t \geq t') \int_s^t \imath \mathbf{k} \mathbf{U}(t') f^{\triangle}(\mathbf{k}, t' \geq s) dt' \right\rangle \\ &\quad + \left\langle \int_s^t \exp(-a|\mathbf{k}|^z (t-t') \triangle^{\mathbf{u}}(\mathbf{k}, t' \geq s)) \imath \mathbf{k} \mathbf{U}(t') f^{\triangle}(\mathbf{k}, t' \geq s) dt' \right\rangle, \end{aligned} \tag{C.17}$$

where the useful property of $\triangle^{\mathbf{u}}$,

$$\triangle^{\mathbf{u}}(\mathbf{k}, t \geq s) = \triangle^{\mathbf{u}}(\mathbf{k}, t \geq t') + \exp(-a|\mathbf{k}|^z(t-t'))\triangle^{\mathbf{u}}(\mathbf{k}, t' \geq s)$$
$$\text{for any} \qquad t \geq t' \geq s,$$

which can be easily obtained from its definition (C.7), has been invoked. The first term in the right-hand side of (C.17) vanishes because, as is evident from (A.3), both $\mathbf{U}(t')$ and $f^{\triangle}(\mathbf{k}, t' \geq s)$ contain only $\hat{\mathbf{u}}(\mathbf{q}^{\mathbf{l}}, t_l \leq t')$, and because of the independence properties (C.9). In order to average the second term, one can invoke the equivalent representation of $df^{\triangle}(\mathbf{k}, t \geq s) = \imath\mathbf{k}\mathbf{U}(t)f^{\triangle}(\mathbf{k}, t \geq s)dt$ by (C.16):

$$\begin{aligned}
&\exp(a|\mathbf{k}|^z t)\langle \triangle^{\mathbf{u}}(\mathbf{k}, t \geq s) f^{\triangle}(\mathbf{k}, t \geq s)\rangle \\
&= \left\langle \int_s^t \exp(a|\mathbf{k}|^z t')\imath\mathbf{k}\mathbf{A}(t')\triangle^{\mathbf{u}}(\mathbf{k}, t' \geq s) f^{\triangle}(\mathbf{k}, t' \geq s) d\mathbf{z}(t') \right\rangle \\
&\qquad - \int_s^t \mathbf{k}\tilde{\mathbf{D}}(t')\mathbf{k}\langle \triangle^{\mathbf{u}}(\mathbf{k}, t' \geq s) f^{\triangle}(\mathbf{k}, t' \geq s)\rangle dt' \\
&= -\int_s^t \exp(a|\mathbf{k}|^z t')\mathbf{k}\tilde{\mathbf{D}}(t')\mathbf{k}\langle \triangle^{\mathbf{u}}(\mathbf{k}, t' \geq s) f^{\triangle}(\mathbf{k}, t' \geq s)\rangle dt',
\end{aligned} \tag{C.18}$$

where both sides also have been multiplied by $\exp(a|\mathbf{k}|^z t)$. The first term in (C.18) again vanishes because it is linear in $dz(t')$. Consequently, the time-evolution of the average value $\langle \triangle^{\mathbf{u}} f^{\triangle}\rangle$ is described by the simple equation

$$\frac{d\langle \triangle^{\mathbf{u}} f^{\triangle}\rangle}{dt} = -\left[a|\mathbf{k}|^z + \mathbf{k}\tilde{\mathbf{D}}(t)\mathbf{k}\right]\langle \triangle^{\mathbf{u}} f^{\triangle}\rangle, \qquad \langle \triangle^{\mathbf{u}} f^{\triangle}\rangle(t \leq s) \equiv 0, \tag{C.19}$$

which has the only trivial solution $\langle \triangle^{\mathbf{u}} f^{\triangle}\rangle = \langle \triangle^{\mathbf{u}}(\mathbf{k}, t \geq s) f^{\triangle}(\mathbf{k}, t \geq s)\rangle \equiv 0$.

Hence, term {1} in (C.12) vanishes exactly, and term {2} gives the same average as if the Corrsin's independence hypothesis had been invoked during the averaging of (C.11). This is essentially due to Gaussian distribution of the trajectories and stationarity of both Eulerian and Lagrangian Gaussian velocity fields. All these properties lead to the statistical independence of "changes" in the Eulerian velocity field, $\triangle^{\mathbf{u}}(\mathbf{k}, t \geq s)$, and in the distribution of "differences in trajectories," $f^{\triangle}(\mathbf{k}, t \geq s)$, despite the interdependence of $\mathbf{Z}$ and $\mathbf{u}$.

The substitution of expression (C.13) for the "effectively independent" term {2} into (C.11) hence gives the two-time Lagrangian velocity correlation for $t \geq s$:

$$\begin{aligned}
\langle U_i(t)U_j(s)\rangle_{\hat{u}} &= RL_{ij}(t,s) \\
&= \frac{\alpha^2}{(2\pi)^m}\int_{\delta<k<1} k^{1-m-\epsilon}\exp(-ak^z(t-s))\left(\delta_{ij} - \frac{k_i k_j}{k^2}\right) \\
&\qquad \times \exp\left[-k_m \int_s^t \tilde{D}_{mn}(s')ds' k_n\right] \\
&\qquad \times \exp\left[\imath\sqrt{2D_0}\mathbf{k}(\mathbf{W}(t) - \mathbf{W}(s))\right] d^m\mathbf{k}.
\end{aligned} \tag{C.20}$$

The derivation of the evolution equation for the effective diffusivity now becomes straightforward. According to the definition of effective diffusivity by (C.1),

$$\frac{1}{2}d\tilde{Z}_{ij}(t) = \tilde{D}_{ij}(t)dt = \left\{\int_0^t \left[RL_{ij}(t,s) + RL_{ij}(s,t)\right] ds\right\} dt, \tag{C.21}$$

where (C.20) provides the expressions for $RL_{ij}(t,s)$ and $RL_{ij}(s,t)$ under the condition $t \geq s$. The resulting expression for $\tilde{D}_{ij}(t)$ is

$$\tilde{D}_{ij}(t) = \frac{\alpha^2}{(2\pi)^m} \int_{\delta<k<1} k^{1-m-\epsilon} \tilde{F}(\mathbf{k},t,\mathbf{W}(t)) \left(\delta_{ij} - \frac{k_i k_j}{k^2}\right) d^m\mathbf{k}, \tag{C.22a}$$

$$\tilde{F}(\mathbf{k},t,\mathbf{W}(t)) = \int_0^t \exp\left(-ak^z(t-s) - k_m \int_s^t \tilde{D}_{mn}(s')ds' k_n\right) \times \cos\left[\sqrt{2D_0} k_l (W_l(t) - W_l(s))\right] ds. \tag{C.22b}$$

The averaging of all quantities involving $\tilde{D}_{ij}(t)$ over $\mathbf{W}(t)$ and the analysis of their time-evolution becomes easier if one can write down the governing differential equation for the real-valued function $\tilde{F}(\mathbf{k},t,\mathbf{W}(t))$. Invoking Ito's stochastic differentiation,

$$df\,(t,\mathbf{W}(t)) = \left[\frac{\partial f}{\partial t} + \frac{1}{2}\nabla_W^2 f\right] dt + \frac{\partial f}{\partial W_i} dW_i(t),$$

one can write the evolution equation for $\tilde{F}(\mathbf{k},t,\mathbf{W}(t))$ defined in (C.22b):

$$\begin{aligned} d\tilde{F} &= \left[1 - \left(ak^z + k^2 D_0 + k_m \tilde{D}_{mn}(t) k_n\right) \tilde{F}\right] dt - \tilde{G}\sqrt{2D_0} k_l dW_l(t), \\ d\tilde{G} &= -\left[ak^z + k^2 D_0 + k_m \tilde{D}_{mn}(s) k_n\right] \tilde{G} dt + \tilde{F}\sqrt{2D_0} k_l dW_l(t), \\ \tilde{F}(\mathbf{k},0,\mathbf{W}(0)) &= 0, \qquad \tilde{G}(\mathbf{k},0,\mathbf{W}(0)) = 0, \end{aligned} \tag{C.23}$$

where

$$\begin{aligned} \tilde{G} = \tilde{G}(\mathbf{k},t,\mathbf{W}(t)) &= -\frac{\partial \tilde{F}}{\partial (k_l W_l(t))} \\ &= \int_0^t \exp\left(-ak^z(t-s) - k_m \int_s^t \tilde{D}_{mn}(s')ds' k_n\right) \\ &\qquad \times \sin\left[\sqrt{2D_0} k_l (W_l(t) - W_l(s))\right] ds. \end{aligned}$$

It is evident that (C.22a) and (C.23) are exactly the same as equations (5.7a)–(5.7d).

REFERENCES

[1] M. Avellaneda and A. Majda, *Mathematical models with exact renormalization for turbulent transport*, Comm. Math. Phys., 131 (1990), pp. 381–429.

[2] M. Avellaneda and A. Majda, *Approximate and exact renormalization theories for a model for turbulent transport*, Phys. Fluids A, 4 (1992), pp. 41–57.

[3] M. Avellaneda and A. Majda, *Renormalization theory for eddy diffusivity in turbulent transport*, Phys. Rev. Lett., 68 (1992), pp. 3028–3031.

[4] L. W. Baldwin and T. J. Walsh, *Turbulent diffusion in the core of fully developed pipe flow*, AIChE J., 7 (1961), pp. 53–61.

[5] S. Corrsin, *Theories of turbulent transport*, in Mecanique de la Turbulence, Centre Nat. Recherche Sci., Paris, 1962, pp. 27–52.

[6] C. De Dominicis and P. Martin, *Energy spectra of certain randomly-stirred fluids*, Phys. Rev. A (3), 19 (1979), pp. 419–422.

[7] J. W. Deardorff and R. L. Peskin, *Lagrangian statistics from numerically integrated turbulent flow*, Phys. Fluids, 13 (1970), pp. 584–595.

[8] A. C. Fannjiang, *Phase diagram for turbulent transport: Sampling drift, eddy diffusivity and variational principles*, Phys. D, 136 (2000), pp. 145–174.

[9] D. Forster, D. Nelson, and M. Stephen, *Large-distance and long-time properties of randomly stirred fluid*, Phys. Rev. A, 16 (1977), pp. 732–749.
[10] C. W. Gardiner, *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences*, 2nd ed., Springer-Verlag, Berlin, 1990.
[11] H. C. Groenhof, *Eddy diffusion in the central region of turbulent flows in pipes and between parallel planes*, Chem. Eng. Sci., 25 (1970), pp. 1005–1014.
[12] A. Indeikina and H.-C. Chang, *Effective diffusion in time-periodic linear planar flow*, Phys. Fluids A, 5 (1993), pp. 2563–2566.
[13] Y. Kimura and R. H. Kraichnan, *Statistics of an advected passive scalar*, Phys. Fluids A, 5 (1993), pp. 2264–2277.
[14] D. Koch and J. Brady, *Anomalous diffusion due to long-range velocity fluctuations in the absence of mean flow*, Phys. Fluids A, 1 (1989), pp. 47–51.
[15] D. Koch and E. Shaqfeh, *Averaged-equation and diagrammatic approximations to the average concentration of tracer dispersed by a Gaussian random velocity field*, Phys. Fluids A, 4 (1992), pp. 887–894.
[16] R. H. Kraichnan, *The structure of isotropic turbulence at very high Reynolds numbers*, J. Fluid Mech., 5 (1959), pp. 497–543.
[17] R. H. Kraichnan, *Lagrangian-history closure approximation for turbulence*, Phys. Fluids, 8 (1965), pp. 575–598.
[18] W. D. McComb and L. H. Rabie, *The turbulent diffusion of drug-reduction polymer solutions from a point source in flow through a pipe*, Chem. Eng. Sci., 37 (1982), pp. 1759–1770.
[19] W. D. McComb, *The Physics of Fluid Turbulence*, Clarendon, Oxford, 1990.
[20] T. K. Sherwood, R. L. Pigford, and C. R. Wilke, *Mass Transfer*, McGraw-Hill, New York, 1975.
[21] D. J. Shlien and S. Corrsin, *A measurement of Lagrangian velocity autocorrelation in approximately isotropic turbulence*, J. Fluid Mech., 62 (1974), pp. 255–271.
[22] G. I. Taylor, *Diffusion by continuous movements*, Proc. London Math. Soc., 20 (1921), pp. 196–212.
[23] V. Yakhot and S. Orszag, *Renormalization group analysis of turbulence.* I. *Basic theory*, J. Sci. Comput., 1 (1986), pp. 3–51.
[24] V. Yakhot, S. Orszag, and A. Yakhot, *Heat transfer in turbulent fluids – I. Pipe flow*, Phys. Fluids, 30 (1987), pp. 15–22.

SIAM J. APPL. MATH.
Vol. 63, No. 1, pp. 42–61

SELF-SIMILAR SOLUTIONS FOR WEAK SHOCK REFLECTION*

ALLEN M. TESDALL† AND JOHN K. HUNTER‡

Abstract. We present numerical solutions of a two-dimensional Riemann problem for the unsteady transonic small disturbance equations that provides an asymptotic description of the Mach reflection of weak shock waves. We develop a new numerical scheme to solve the equations in self-similar coordinates and use local grid refinement to resolve the solution in the reflection region. The solutions contain a remarkably complex structure: there is a sequence of triple points and tiny supersonic patches immediately behind the leading triple point that is formed by the reflection of weak shocks and expansion waves between the sonic line and the Mach shock. An expansion fan originates at each triple point, thus resolving the von Neumann paradox of weak shock reflection. These numerical solutions raise the question of whether there is an infinite sequence of triple points in an inviscid weak shock Mach reflection.

Key words. weak shock reflection, self-similar solutions, unsteady transonic small disturbance equation, two-dimensional Riemann problems, von Neumann paradox

AMS subject classifications. 65M06, 35L65, 76L05

PII. S0036139901383826

1. Introduction.

Experimental observations of the Mach reflection of weak shock waves off a wedge show a pattern that closely resembles a single Mach reflection, in which the incident, reflected, and Mach shocks meet at a triple point. The von Neumann theory of shock reflection [10, 16] shows that a standard triple point configuration, consisting of three shocks and a contact discontinuity, is impossible for sufficiently weak shocks. This apparent conflict between theory and experiment for weak shock reflection has been a long-standing puzzle and is often referred to as the triple point, or von Neumann, "paradox" (see section I.17 of [2], for example).

Guderley [8, 9] proposed that there is a supersonic region behind the triple point in a steady weak shock Mach reflection, in which case there is an additional expansion fan at the triple point, resolving the apparent paradox. There was, however, no evidence of a supersonic region or an expansion fan in experimental observations [3, 18, 19] or numerical solutions [4, 5, 20] of weak shock reflections off a wedge, until Hunter and Brio [12] obtained a numerical solution of a shock reflection problem for the unsteady transonic small disturbance equation that contained a supersonic region behind the triple point. The region is extremely small, which is why it had not been detected previously. Subsequently, Zakharian et al. [24] found a supersonic region in a numerical solution of a shock reflection problem for the full Euler equations, using local grid refinement near the triple point, for a set of parameter values corresponding to those in [12].

The solutions in [12, 24] are for a single set of parameter values, and they are not sufficiently well resolved to show an expansion fan at the triple point directly, or to show the structure of the flow inside the supersonic region. In this paper, we present

*Received by the editors January 17, 2001; accepted for publication (in revised form) February 25, 2002; published electronically August 5, 2002.
http://www.siam.org/journals/siap/63-1/38382.html

†Department of Mechanical and Aeronautical Engineering and Center for Computational Fluid Dynamics, University of California at Davis, Davis, CA 95616 (amtesdal@ucdavis.edu).

‡Department of Mathematics and Institute of Theoretical Dynamics, University of California at Davis, Davis, CA 95616 (jkhunter@ucdavis.edu). The research of this author was partially supported by the NSF under grant DMS-0072343.

high-resolution numerical solutions of the shock reflection problem for the unsteady transonic small disturbance equations for a range of parameter values. There is a supersonic region behind the triple point in all of the numerical solutions obtained here. This region consists of a sequence of supersonic patches formed by a sequence of expansion fans and shock waves that are reflected between the sonic line and the Mach shock (see Figures 5 and 6, for example). Each of the reflected shocks intersects the Mach shock, resulting in a sequence of triple points, rather than a single triple point. The numerical results raise the question of whether there is an infinite sequence of triple points in an inviscid weak shock Mach reflection.

The total size of the repeating structure of supersonic patches is approximately the same as that of the supersonic region in the solution obtained in [12], at the same parameter value, by a different numerical scheme. Other important quantities, including the strength of the reflected shock and the location of the triple point, agree closely with this solution, providing an independent check on the self-similar solutions presented here.

There are, at the moment, no experimental observations of a supersonic region behind the triple point in a weak shock Mach reflection. As we discuss in section 5, the small size of the region and the effect of viscosity may make it very difficult to detect experimentally. A structure similar to the one in the solutions presented here has been observed in shock-boundary layer interactions in transonic flows over an airfoil [1, 13] (see Figures 245 and 247 in [6]). The shock reflects off a laminar boundary layer as an expansion wave, leading to a sequence of reflected shock and expansion waves inside the supersonic bubble on the airfoil.

The numerical solutions of weak shock reflection in [5, 12, 20, 24] were obtained by solving an initial-value problem for the unsteady equations. The problem of inviscid shock reflection off a wedge is self-similar, and there are a number of advantages to solving the problem in self-similar, rather than unsteady, form. In the unsteady formulation the equations are time-marched, and any waves present move through the computational domain, complicating algorithms for local grid refinement near the triple point. By contrast, a solution of the self-similar equations is stationary, making local grid refinement algorithms much easier to implement. Moreover, a global grid refinement strategy is possible, in which a partially converged solution on a coarse grid is interpolated onto a fine grid, and then converged on the fine grid. This process may be repeated recursively until the desired resolution is obtained.

In this paper, we present numerical solutions of the shock reflection problem for the unsteady transonic small disturbance equations computed in self-similar coordinates. Samtaney [17] developed a scheme for the solution of the Euler equations in self-similar coordinates, but his scheme does not apply to the unsteady transonic small disturbance equations, and a different approach is required. In our approach, we introduce special self-similar variables in which the self-similar transonic small disturbance equations have the form of the usual transonic small disturbance equations modified by lower-order terms. What makes the use of the unsteady transonic small disturbance equations worthwhile is the fact that, with the same computational resources, we can obtain a much more finely resolved solution than for the Euler equations.

This paper is organized as follows. In section 2, we describe the shock reflection problem for the unsteady transonic small disturbance equation, and in section 3 we give the details of our numerical method. In section 4, we present our numerical solutions. In section 5, we discuss some of the questions raised by these solutions and consider the effect of physical viscosity on the inviscid solutions. We summarize our

conclusions in section 6.

2. The asymptotic shock reflection problem. The asymptotic shock reflection problem [11, 12, 14, 20] consists of the unsteady transonic small disturbance equation

$$\begin{aligned} &u_t + \left(\frac{1}{2}u^2\right)_x + v_y = 0, \\ &u_y - v_x = 0 \end{aligned} \tag{2.1}$$

in the half space $y > 0$ with the initial and boundary conditions

$$u(x, y, 0) = \begin{cases} 0 & \text{if } x > ay, \\ 1 & \text{if } x < ay, \end{cases} \tag{2.2}$$

$$v(x, y, t) = 0 \qquad \text{if } x > \sigma(y, t), \tag{2.3}$$

$$v(x, 0, t) = 0. \tag{2.4}$$

Here, $x = \sigma(y, t)$ is the location of the incident and Mach shocks. The location of the incident shock is given by

$$x = ay + \left(\frac{1}{2} + a^2\right) t. \tag{2.5}$$

The incident shock strength, as measured by the jump in u, is normalized to one. This problem depends on a single parameter a, the inverse slope of the incident shock.

These equations may be derived by a systematic asymptotic expansion of the shock reflection problem for the full Euler equations for weak shock reflection off thin wedges [12]. The variables $u(x, y, t)$, $v(x, y, t)$ are proportional to the x, y fluid velocity components, respectively, and pressure perturbations are proportional to u. The flow is irrotational and isentropic to leading order in the shock strength.

If the Mach number of the incident shock is M, and the wedge angle in radians is θ_w, then (2.1)–(2.4) is obtained in the limit $M \to 1$ and $\theta_w \to 0$, with

$$a = \frac{\theta_w}{2\sqrt{M - 1}} \tag{2.6}$$

fixed. Because of transonic similarity, the asymptotic problem depends on a single combination of the incident shock strength and the wedge angle. A regularly reflected solution of (2.1)–(2.4) is impossible when $a < \sqrt{2}$, and triple point solutions of (2.1), in which three plane shocks separated by constant states meet at a point, do not exist.

The problem (2.1)–(2.4) is self-similar, so the solution depends only on the similarity variables

$$\xi = \frac{x}{t}, \qquad \eta = \frac{y}{t}.$$

Writing (2.1) in terms of ξ, η, and a pseudo-time variable $\tau = \log t$, we get

$$\begin{aligned} &u_\tau - \xi u_\xi - \eta u_\eta + \left(\frac{1}{2}u^2\right)_\xi + v_\eta = 0, \\ &u_\eta - v_\xi = 0. \end{aligned} \tag{2.7}$$

As $\tau \to +\infty$, solutions of (2.7) converge to a pseudo-steady, self-similar solution that satisfies

$$\begin{aligned} &-\xi u_\xi - \eta u_\eta + \left(\frac{1}{2}u^2\right)_\xi + v_\eta = 0, \\ &u_\eta - v_\xi = 0. \end{aligned} \tag{2.8}$$

Equation (2.8) is hyperbolic when $u < \xi + \eta^2/4$, corresponding to supersonic flow in a self-similar coordinate frame, and is elliptic when $u > \xi + \eta^2/4$, corresponding to subsonic flow. The equation changes type across the sonic line given by

$$\xi + \frac{\eta^2}{4} = u(\xi, \eta). \tag{2.9}$$

3. The numerical method. In order to solve (2.7) numerically, we write it in terms of parabolic coordinates

$$\begin{aligned} &r = \xi + \frac{1}{4}\eta^2, \qquad \theta = \eta, \\ &\widetilde{u} = u - \left(\xi + \frac{1}{4}\eta^2\right), \qquad \widetilde{v} = v - \frac{1}{2}\eta u, \end{aligned} \tag{3.1}$$

which gives

$$\begin{aligned} &\widetilde{u}_\tau + \left(\frac{1}{2}\widetilde{u}^2\right)_r + \widetilde{v}_\theta + \frac{3}{2}\widetilde{u} + \frac{1}{2}r = 0, \\ &\widetilde{u}_\theta - \widetilde{v}_r = 0. \end{aligned} \tag{3.2}$$

With respect to these variables, the self-similar equations have the form of the usual transonic small disturbance equations modified by lower-order terms, and they can be solved by a standard numerical scheme. We introduce a potential $\varphi(r, \theta, \tau)$ such that

$$\widetilde{u} = \varphi_r, \qquad \widetilde{v} = \varphi_\theta, \tag{3.3}$$

and we write (3.2) in the potential form

$$\varphi_{r\tau} + \left(\frac{1}{2}\varphi_r^2\right)_r + \varphi_{\theta\theta} + \frac{3}{2}\varphi_r + \frac{1}{2}r = 0. \tag{3.4}$$

We define a nonuniform grid r_i in the r direction and θ_j in the θ direction, where $r_{i+1} = r_i + \Delta r_{i+1/2}$ and $\theta_{j+1} = \theta_j + \Delta\theta_{j+1/2}$. We also define $(r_{i-1/2}, r_{i+1/2})$ as the neighborhood of the point r_i, with length $\Delta r_i = \frac{1}{2}(\Delta r_{i-1/2} + \Delta r_{i+1/2})$, where $r_{i+1/2} = \frac{1}{2}(r_{i+1} + r_i)$. Similar definitions apply for the nonuniform grid θ_j. We denote an approximate solution of (3.4) by

$$\varphi_{i,j}^n \approx \varphi(r_i, \theta_j, n\Delta\tau),$$

where $\Delta\tau$ is a fixed time step, and we discretize (3.4) in time τ using

$$\frac{\varphi_r^{n+1} - \varphi_r^n}{\Delta\tau} + \varphi_{\theta\theta}^{n+1} + f(\varphi_r)_r^n + \frac{3}{2}\varphi_r^{n+1} + \frac{1}{2}r = 0, \tag{3.5}$$

where the flux function f is defined by

$$f(\widetilde{u}) = \frac{1}{2}\widetilde{u}^2. \tag{3.6}$$

We solve (3.5) by sweeping from right to left in r, using the spatial discretization

(3.7)

$$\begin{aligned}
&\varphi_{i,j}^{n+1} - \Delta r_{i+1/2}\Delta\tau \left(\frac{\frac{\varphi_{i,j+1}-\varphi_{i,j}}{\Delta\theta_{j+1/2}} - \frac{\varphi_{i,j}-\varphi_{i,j-1}}{\Delta\theta_{j-1/2}}}{\Delta\theta_j} \right)^{n+1} + \frac{3}{2}\Delta\tau\varphi_{i,j}^{n+1} \\
&\quad = \varphi_{i+1,j}^{n+1} - \varphi_{i+1,j}^{n} + \varphi_{i,j}^{n} + \Delta\tau \left(F(\widetilde{u}_{i+1/2,j}, \widetilde{u}_{i+3/2,j})^n - F(\widetilde{u}_{i-1/2,j}, \widetilde{u}_{i+1/2,j})^n \right) \\
&\qquad + \frac{3}{2}\Delta\tau\varphi_{i+1,j}^{n+1} + \frac{1}{2}\Delta\tau\Delta r_{i+1/2}\,\hat{r}_{i+1/2}.
\end{aligned}$$

Here, F is a numerical flux function, and

$$\widetilde{u}_{i-1/2,j} = \frac{\varphi_{i,j} - \varphi_{i-1,j}}{\Delta r_{i-1/2}}.$$

The variable $\hat{r}_{i+1/2}$ is the value of r at which the source term $\frac{1}{2}r$ is evaluated, and in most of the calculations we used the definition $\hat{r}_{i+1/2} = r_i$. We tried a number of different treatments of the source term and obtained similar results with them all. See [21] for a detailed discussion.

In most of the computations, we used an Engquist–Osher numerical flux function. Dropping the θ-subscript j, which is constant in the following definitions, the Engquist–Osher flux for (3.6) is

$$F^{EO}(\widetilde{u}_{i-1/2}, \widetilde{u}_{i+1/2}) = \frac{1}{2}\max(\widetilde{u}_{i-1/2}, 0)^2 + \frac{1}{2}\min(\widetilde{u}_{i+1/2}, 0)^2.$$

In our highest resolution computations for $a = 0.5$, we used a second-order, flux-limiter scheme [23], with a Lax–Wendroff flux as the higher-order flux, and an Engquist–Osher flux as the lower-order flux. The numerical flux function for this scheme is given by

$$F(\widetilde{u}_{i-1/2}, \widetilde{u}_{i+1/2}) = \psi(\varrho)F^{LW}(\widetilde{u}_{i-1/2}, \widetilde{u}_{i+1/2}) + (1 - \psi(\varrho))F^{EO}(\widetilde{u}_{i-1/2}, \widetilde{u}_{i+1/2}),$$

$$\varrho = \begin{cases}
\dfrac{\left(\left|\frac{\widetilde{u}_{i-3/2}+\widetilde{u}_{i-1/2}}{2}\right| - \frac{\Delta\tau}{\Delta r_i}\left(\frac{\widetilde{u}_{i-3/2}+\widetilde{u}_{i-1/2}}{2}\right)^2\right)(\widetilde{u}_{i-1/2} - \widetilde{u}_{i-3/2})}{\left(\left|\frac{\widetilde{u}_{i-1/2}+\widetilde{u}_{i+1/2}}{2}\right| - \frac{\Delta\tau}{\Delta r_i}\left(\frac{\widetilde{u}_{i-1/2}+\widetilde{u}_{i+1/2}}{2}\right)^2\right)(\widetilde{u}_{i+1/2} - \widetilde{u}_{i-1/2})}, & \frac{\widetilde{u}_{i-1/2}+\widetilde{u}_{i+1/2}}{2} \geq 0, \\[2ex]
\dfrac{\left(\left|\frac{\widetilde{u}_{i+1/2}+\widetilde{u}_{i+3/2}}{2}\right| - \frac{\Delta\tau}{\Delta r_i}\left(\frac{\widetilde{u}_{i+1/2}+\widetilde{u}_{i+3/2}}{2}\right)^2\right)(\widetilde{u}_{i+3/2} - \widetilde{u}_{i+1/2})}{\left(\left|\frac{\widetilde{u}_{i-1/2}+\widetilde{u}_{i+1/2}}{2}\right| - \frac{\Delta\tau}{\Delta r_i}\left(\frac{\widetilde{u}_{i-1/2}+\widetilde{u}_{i+1/2}}{2}\right)^2\right)(\widetilde{u}_{i+1/2} - \widetilde{u}_{i-1/2})}, & \frac{\widetilde{u}_{i-1/2}+\widetilde{u}_{i+1/2}}{2} < 0,
\end{cases}$$

where ψ is a minmod flux-limiter,

$$\psi(\varrho) = \begin{cases} 0, & \varrho \leq 0, \\ \varrho, & 0 < \varrho < 1, \\ 1, & \varrho \geq 1. \end{cases}$$

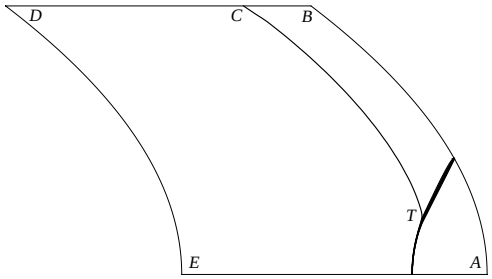


FIG. 1. *A schematic diagram of the computational domain. EA is the wall and ABDE is the numerical boundary. The incident shock enters the computational domain through AB. The incident, reflected, and Mach shocks meet at the triple point T.*

The Lax–Wendroff flux for (3.6) is given by

$$F^{LW}(\widetilde{u}_{i-1/2},\widetilde{u}_{i+1/2}) = \frac{1}{4}(\widetilde{u}^2_{i-1/2}+\widetilde{u}^2_{i+1/2}) - \frac{1}{2}\frac{\Delta\tau}{\Delta r_i}\left(\frac{\widetilde{u}_{i-1/2}+\widetilde{u}_{i+1/2}}{2}\right)^2(\widetilde{u}_{i+1/2}-\widetilde{u}_{i-1/2}).$$

We evolve the solution of (3.7) forward in time τ until it converges to a steady state, using line relaxation. The direction of sweep, from right to left in r, is consistent with the direction of propagation of the characteristics for (2.8), which is in the $-r$ direction.

3.1. Boundary conditions. We computed solutions of the half-space problem (2.1)–(2.4) in the finite computational domain

$$r^L \le r \le r^R, \qquad 0 \le \theta \le \theta^T,$$

shown schematically in Figure 1. The left and right boundaries of the computational domain are parabolic because of the use of the coordinates in (3.1). We use a nonuniform grid that has a locally refined area of uniform grid very close to the triple point, and is stretched exponentially away from the triple point toward the outer numerical boundaries and the wall. In the solutions shown below, the nonuniform grids are stretched by amounts between 0.5% and 1.5%, and the total number of grid points in our largest grid is approximately 3×10^6.

We impose the physical no-flow condition (2.4), which implies that $\varphi_\theta = 0$, on the wall EA. In addition, we require numerical boundary conditions on the outer computational boundaries.

On the right boundary AB, we impose Dirichlet data corresponding to the incident shock solution in (2.2)–(2.3). Using (3.1) in (2.5), we find that the incident shock location with respect to the transformed self-similar coordinates is given by

$$r = a\theta + \frac{1}{4}\theta^2 + \frac{1}{2} + a^2.$$

Thus, the incident shock location is a parabola with respect to the transformed coordinates, instead of a straight line. Ahead of the incident shock we have $(u,v) = (0,0)$, and behind the incident shock we have $(u,v) = (1,-a)$. Hence, using (3.1), (3.3), and

the requirement that the potential is continuous across the shock, we find that the potential for the incident shock solution is given by

$$\varphi(r,\theta) = \begin{cases} -\frac{1}{2}r^2, & r > a\theta + \frac{1}{4}\theta^2 + \frac{1}{2} + a^2, \\ r - a\theta - \frac{1}{4}\theta^2 - \frac{1}{2}r^2 - \frac{1}{2} - a^2, & r < a\theta + \frac{1}{4}\theta^2 + \frac{1}{2} + a^2. \end{cases} \tag{3.8}$$

We impose (3.8) as a boundary condition for (3.4) on AB.

The asymptotic behavior of the solution of the shock reflection problem at large distances from the reflection point is given by the solution of the linearized shock reflection problem [12]. We use this result to formulate a numerical boundary condition on the subsonic boundary CDE. In self-similar variables, the linearized solution for φ_r behind the reflected wavefront $r = 1$ is

$$\varphi_r = 1 - r + \frac{1}{\pi}\tan^{-1}\left(\frac{2a\sqrt{1-r}}{1 - r + \frac{1}{4}\theta^2 - a^2}\right), \quad r < 1. \tag{3.9}$$

We impose (3.9) as a Neumann condition on the left boundary DE. Writing (3.9) as $\varphi_r = f(r,\theta)$, we discretize it as

$$\frac{\varphi_{i+1,j} - \varphi_{i,j}}{\Delta r_{i+1/2}} = f(r_{i+1/2}, \theta_j).$$

On the top boundary BD, we impose the Dirichlet condition (3.8) when $r > 1$, corresponding to the segment BC, and the condition (3.9) when $r < 1$, corresponding to the segment CD. The exact location of the reflected shock is slightly different from the point $r = 1$, where we switch the numerical boundary conditions, and the exact solution differs slightly from the linearized solution, but we found that the disturbance originating from the top boundary was small provided that the boundary was far enough away from the triple point (see Figure 9). We tried a number of other numerical boundary conditions, but (3.8)–(3.9) gave the most satisfactory results.

4. Numerical results. We computed numerical solutions of (2.1)–(2.4) for a equal to 0.3, 0.4, 0.5, 0.6, 0.65, 0.7, 0.75, and 0.8. In the following figures, we present solutions for the values 0.3, 0.5, 0.6, and 0.8. The solutions for the other values of a are similar to the ones presented here. Figure 2 shows u-contour plots of the global solutions as a function of $(x/t, y/t)$. From (2.6), increasing a corresponds to increasing the wedge angle while fixing the Mach number of the incident shock, or decreasing the Mach number while fixing the wedge angle. Hence, the sequence of plots in Figure 2(a)–(d) is a numerical representation of a series of shock reflection experiments in which the wedge angle is increased, while the Mach number is held constant at a value near one.

The numerical solutions closely resemble a single Mach reflection. The Mach shock becomes shorter and stronger as a increases, and the strength of the reflected shock near the triple point, which is very weak for smaller values of a, also increases with a (see Table 4.1). For a fixed value of a, the strength of the Mach shock increases as it moves away from the triple point, reaching a maximum at the wall $y = 0$. The strength of the reflected shock increases initially as it moves away from the triple point, then decreases, approaching zero as $y \to +\infty$. The thickening of the incident shock as it moves away from the triple point in Figure 2(a)–(d) is caused by the use of a stretched grid.

In Figure 3, we show the u-contours and the numerically computed location of the sonic line (2.9) near the triple point for the values of a shown in Figure 2. All of

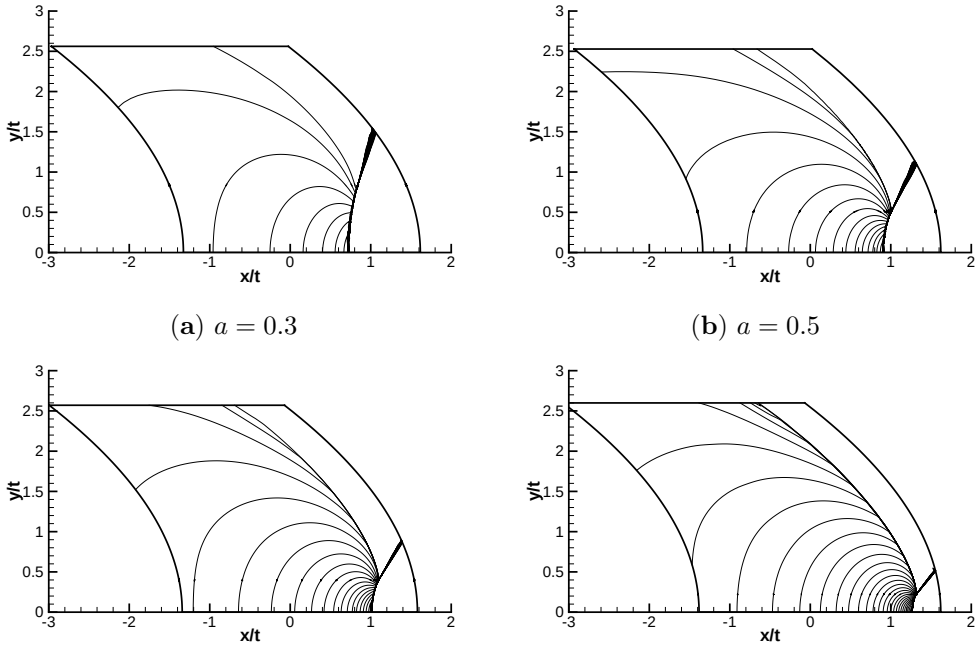


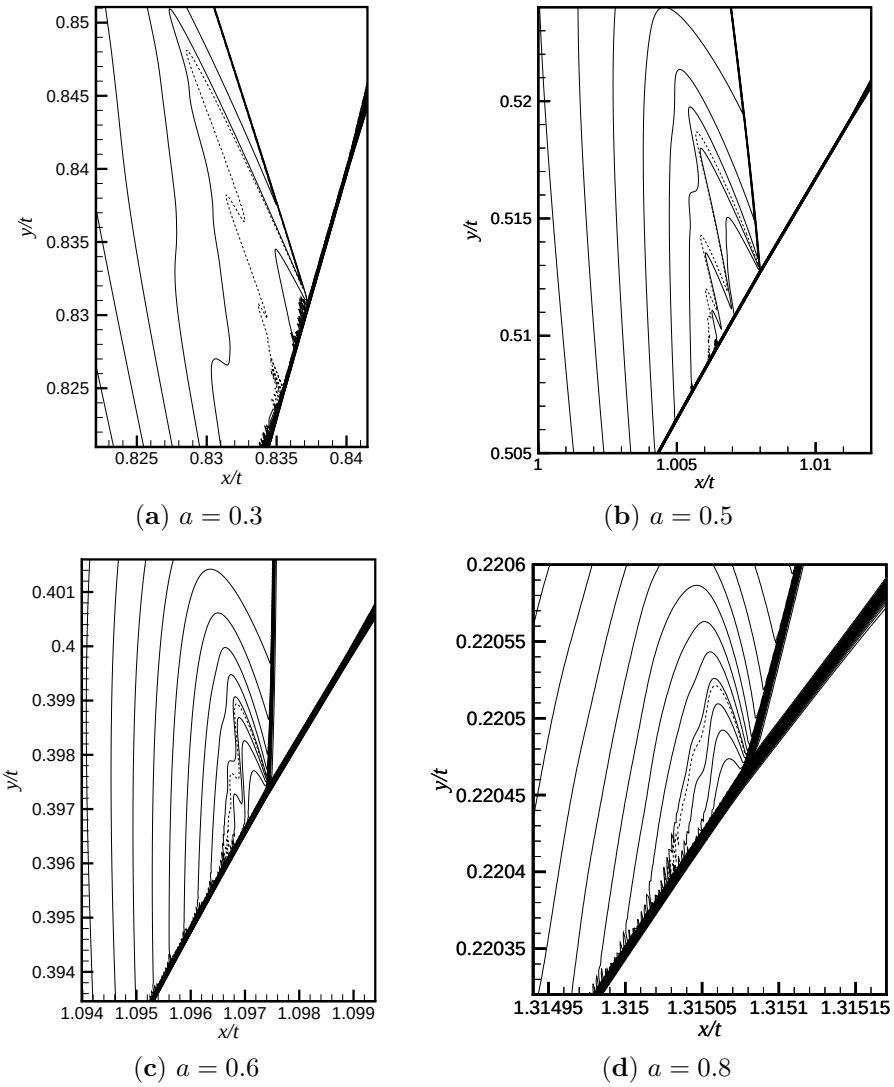
FIG. 2. *Contour plots of u for increasing values of a, showing the full numerical domain. The u-contour spacing is* 0.05.

TABLE 4.1

Numerically computed values of the size of the supersonic region at the triple point, the triple point location, and the strength $[u]_r$ *of the reflected shock at the sonic point. The shock strength is measured by the jump* $[u]$ *in u.*

a	$\Delta(x/t)$	$\Delta(y/t)$	$(x/t)_{t.p.}$	$(y/t)_{t.p.}$	$[u]_r$
0.3	0.0030	0.023	0.837	0.831	0.01
0.4	0.0023	0.019	0.924	0.665	0.03
0.5	0.0012	0.0096	1.008	0.513	0.07
0.6	0.0006	0.0030	1.098	0.398	0.13
0.65	0.0004	0.0014	1.148	0.349	0.17
0.7	0.00016	0.00074	1.200	0.302	0.22
0.75	0.00008	0.00027	1.255	0.258	0.27
0.8	0.00004	0.00011	1.315	0.220	0.33

the solutions contain a small region of supersonic flow behind the triple point, the size of which decreases rapidly with increasing a. Table 4.1 gives the size of the supersonic region in the numerical solution for each value of a. The height $\Delta(y/t)$ is a numerical estimate of the difference between the maximum value of y/t on the sonic line and the minimum value of y/t at the rear sonic point on the Mach shock. The width $\Delta(x/t)$ is an estimate of the width of the supersonic region at the value of y/t corresponding to the triple point. In detailed plots of our most refined solution with $a = 0.5$ (see Figures 5 and 6, for example), the expansion fan generated by the collision of the reflected shock with the incident shock at the triple point can be clearly seen. Behind the leading triple point, there is a sequence of shocks and expansion fans. These shocks are less apparent in the less resolved solutions, such as Figure 3(c), and in



(**a**) $a = 0.3$ (**b**) $a = 0.5$

(**c**) $a = 0.6$ (**d**) $a = 0.8$

FIG. 3. *Contour plots of u near the triple point for increasing values of a. The u-contour spacing is* 0.005 *in* (a)*, and* 0.01 *in* (b)–(d)*. The dotted line is the sonic line. The regions shown contain the refined uniform grids, which have the following numbers of grid points:* (a) 620×480*;* (b) 768×608*;* (c) 346×260*;* (d) 245×150.

Figure 3(d) they cannot be seen at all.

The area covered by the most refined uniform grid fits inside the region shown in Figure 3(a)–(d); the actual refined grid area would appear as a sheared rectangle because the equations are discretized with respect to the parabolic coordinates in (3.1). The figure caption gives the number of grid points in the most refined area of the grid. The small numerical oscillations immediately behind the Mach shock (see Figure 3(a) and (d), for example) seem to be caused by the lack of alignment of the shock with the grid.

We found that, for a given value of a, a certain minimum grid resolution was required to resolve the supersonic region behind the triple point. As we refined the

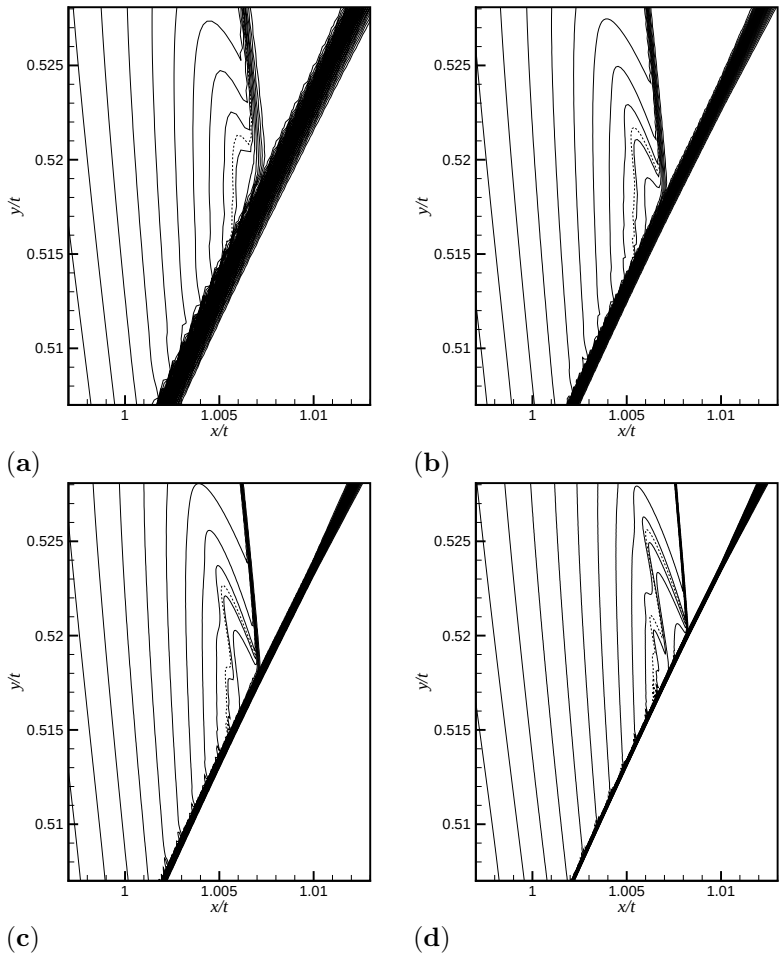


FIG. 4. *A sequence of contour plots illustrating the effect of increasing grid resolution on the numerical solution. The solutions plotted here are for* $a = 0.5$. *The figures show the* u*-contours in the refined grid area near the triple point, with a* u*-contour spacing of* 0.01. *Each grid is refined by a factor of two in both* x/t *and* y/t *in relation to the previous grid. The region shown includes the refined uniform grid area. The dotted line is the sonic line. In* (a), *the refined uniform grid contains* 64×42 *grid points. A supersonic region is visible as a bump in the sonic line, but it is poorly resolved. In* (b), *the refined uniform grid area contains* 128×84 *grid points. The supersonic region appears to be smooth. In* (c), *the refined uniform grid area contains* 256×168 *grid points. There is an indication of a shock wave behind the leading triple point. The refined uniform grid in* (d) *contains* 512×336 *grid points. Two shock waves are visible behind the leading triple point.*

grid beyond this minimum level, a detailed flowfield structure in the region emerged. Figure 4 shows the u-contours and the sonic line near the triple point for a sequence of solutions for $a = 0.5$ computed on successively refined grids. In this sequence, we refined each grid by a factor of two in x/t and y/t in relation to the previous grid. The resolution of the locally refined areas is indicated on the plots. In Figure 4(a)–(b), the sonic line appears fairly smooth. The supersonic region in Figure 4(b) is similar in size, shape, and resolution to the one obtained in [12]. At the next level of refinement, shown in Figure 4(c), there is an indication of the coalescence of u-contours at the rear of the supersonic region and evidence of a second reflected shock there. Finally,

in Figure 4(d), the second reflected shock is better defined, with an indication of a third, weaker shock following it. Further shocks appear in our most refined solution in Figure 3(b).

Returning to Figure 3, we can explain the qualitative differences between the solutions for different values of a in terms of their numerical resolution. As shown in Table 4.1, the size of the supersonic region decreases with increasing a. We therefore had to use more refined grids for higher values of a. For example, the solution shown in Figure 3(d) for $a = 0.8$ was computed using a grid that was a factor of 16 times more refined in x/t and y/t than the grid used in the solution for $a = 0.5$ shown in Figure 3(b). However, the supersonic region in Figure 3(d) is smaller than the one in Figure 3(b) by a linear factor of about 90, resulting in a lower relative resolution. Consequently, the detailed flowfield near the triple point is not visible in Figure 3(d), similar to the under-resolved solutions shown in Figure 4(a)–(b). By contrast, the solutions for $a = 0.3, 0.5, 0.6$ in Figure 3(a)–(c) contain a sequence of shocks and expansions, evident from the pronounced bumps in the sonic line.

There is a small discrepancy between the numerically computed location of the triple point in these figures and the theoretical location of the incident shock in (2.5). The reason for this discrepancy is that the numerical boundary conditions did not give an incident shock that was of exactly constant strength and exactly straight in $(x/t, y/t)$-coordinates. The deviation of the numerical solution for the incident shock from the exact uniform solution was, however, very small. For example, in our numerical solution for $a = 0.5$, the nonuniformity in u and v in the state behind the incident shock is less than 0.4%, and the numerically computed value of the x/t-coordinate of the triple point differs by 0.15% from the theoretical value obtained from (2.5) using the numerically computed value of y/t. We tried a number of different implementations of the numerical scheme and boundary conditions, but none of them gave an exactly straight incident shock. Nevertheless, we saw a supersonic region and the same structure of reflected shocks and expansion fans inside it for all of the implementations.

In Figure 5, we plot closely spaced u-contours, and more widely spaced v-contours, to give a detailed picture of the sequence of shock and expansion waves for $a = 0.5$. Figure 6 is an enlargement of the solution shown in Figure 5 over a very small area close to the leading triple point, which shows the expansion wave that originates at the triple point. The expansion wave is in the family opposite to the shock waves, and it reflects off the sonic line as a compression wave (cf. the discussion in [9]). This compression wave forms a shock that hits the Mach shock and reflects as the next expansion wave. The result is a sequence of triple points, rather than a single triple point. The variables u and v decrease smoothly across the expansion wave at the front of a patch from sonic to supersonic values, moving from right to left in the downstream direction; then u and v jump from supersonic to subsonic values across the shock at the rear of a patch. A very weak wave is visible behind the incident shock in Figure 5(b). This wave is a numerical artifact that is generated when the incident shock crosses from the stretched grid into the uniform grid.

Each shock-expansion pair in the sequence is smaller and weaker than the one preceding it. Four reflected shocks appear to be visible in Figures 5–6. From the numerical data, their approximate strengths, beginning with the leading reflected shock, are

$$[u]_1 \approx 0.08, \qquad [u]_2 \approx 0.02, \qquad [u]_3 \approx 0.01, \qquad [u]_4 \approx 0.003.$$

Here, the jump $[u]$ in u across a reflected shock is measured at the point where the flow

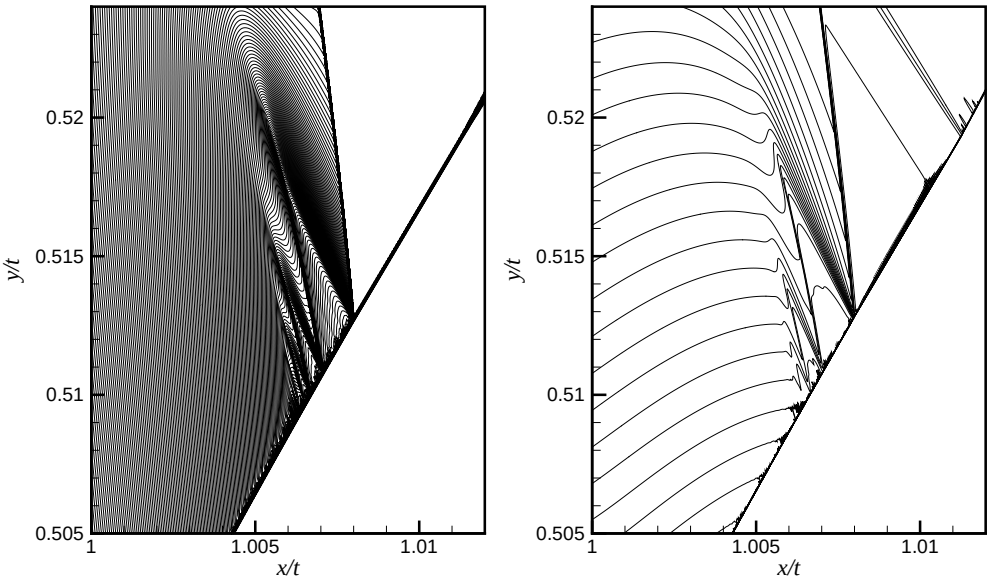


FIG. 5. *A detailed contour plot of* (a) u *and* (b) v *near the triple point for* $a = 0.5$. *The* u*-contour spacing is* 0.0005 *and the* v*-contour spacing is* 0.001. *The sonic line is plotted in Figure* 3(b) *and Figure* 6. *The figure shows a sequence of shock and expansion waves. Each expansion wave is centered at a triple shock intersection and reflects off the sonic line into a compression wave. The compression wave forms a shock wave that intersects the Mach shock, resulting in a sequence of triple points. Three shock-expansion wave pairs and triple points are visible in the plots, with indications of a fourth. The region shown contains the refined uniform grid, which has* 768×608 *grid points.*

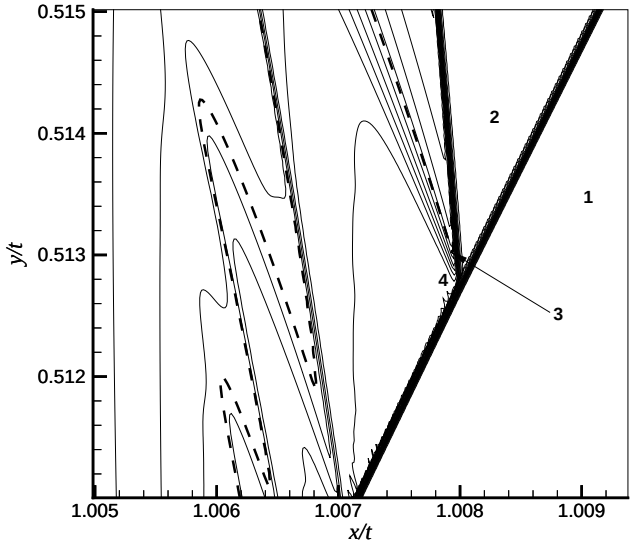
behind the shock is sonic. This point is very close to the corresponding triple point on the Mach shock, as shown in Figure 6. It is not possible, however, to determine from the numerical solution whether or not this sonic point coincides exactly with the triple point, as argued by Guderley [9] in the case of steady weak shock Mach reflections.

Three shocks and an expansion fan appear to connect four states at the leading triple point. We label these states 1–4 in Figure 6. Table 4.2 gives values of u and v for each of the states, computed from the numerical solution. For states 2–4, these values were computed at the locations indicated in the figure. The values of (u, v) for state 3 behind the reflected shock were computed close to the point where the flow behind the shock is sonic. This ensures that states 2 and 3 are connected by the reflected shock and not by any part of the expansion fan, which connects states 3 and 4. For state 1, the values for (u, v) were computed at a location sufficiently far ahead of the incident shock so that they were not influenced by the effects of numerical diffusion near the shock.

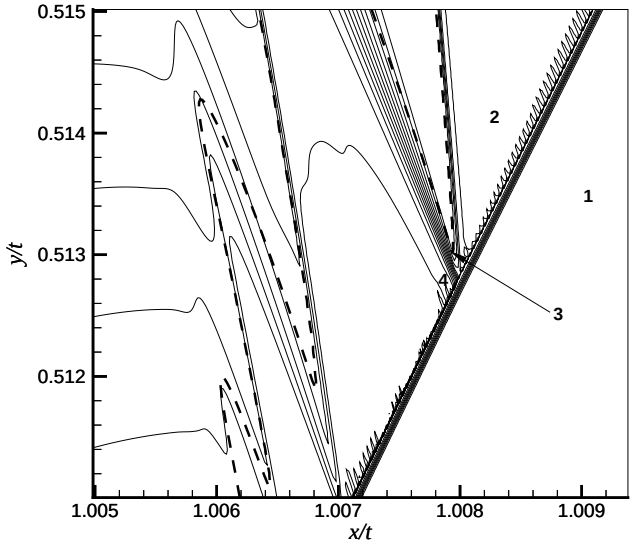
The velocity components $(\bar{u}, \bar{v})$ in a reference frame moving with the triple point are given by [12] as

$$\bar{u} = u - \left(\xi_* + \frac{1}{4}\eta_*^2\right), \qquad \bar{v} = v - \frac{1}{2}\eta_* u, \tag{4.1}$$

where (ξ_*, η_*) are the (ξ, η)-coordinates of the triple point. From the numerical solu-



(**a**)



(**b**)

FIG. 6. *An enlargement of the solution in Figure* 5 *near the leading triple point, showing* (a) *u-contours and* (b) *v-contours. The u-contour spacing is* 0.005*, and the v-contour spacing is* 0.001*. The dashed line in the plots is the sonic line. Table* 4.2 *gives the values of u and v from the numerical solution for the states labeled* 1–4 *in the plots.*

tion shown in Figure 6, we obtain $\xi_* = 1.008$, $\eta_* = 0.5128$. We show the corresponding values of $(\bar{u}, \bar{v})$ in Table 4.2. In Figure 7(a), we plot the shock and rarefaction curves for the steady transonic small disturbance equation [12] through each of the four states for $(\bar{u}, \bar{v})$. The plot in Figure 7(b) is an enlarged view of the shock and rarefaction curves for the states 2, 3, and 4. The curves coincide almost exactly with those of a triple point with an expansion fan. We show similar curves through the numerical values of the analogous states at the second triple point in Figure 7(c)–(d). These

TABLE 4.2

Numerically computed values for the four states at the leading and second triple points, from the solution for $a = 0.5$ (see Figure 6). The state ahead of the incident shock is denoted by 1, the state behind the incident shock by 2, the state behind the reflected shock by 3, and the state behind the Mach shock by 4. The states $1'$–$4'$ are the four analogous states at the second triple point. The variables $\bar{u}$ and $\bar{v}$ are defined in (4.1), with $\xi_* = 1.008$, $\eta_* = 0.5128$ for states 1–4, corresponding to the leading triple point, and $\xi_* = 1.007$, $\eta_* = 0.5108$ for states $1'$–$4'$, corresponding to the second triple point.

State	u	v	$\bar{u}$	$\bar{v}$
1	0	0	-1.074	0
2	0.997	-0.5000	-0.077	-0.756
3	1.073	-0.4963	-0.001	-0.771
4	1.047	-0.5062	-0.027	-0.775
$1'$	0	0	-1.072	0
$2'$	1.052	-0.5076	-0.020	-0.776
$3'$	1.072	-0.5047	0.000	-0.778
$4'$	1.060	-0.5088	-0.012	-0.779

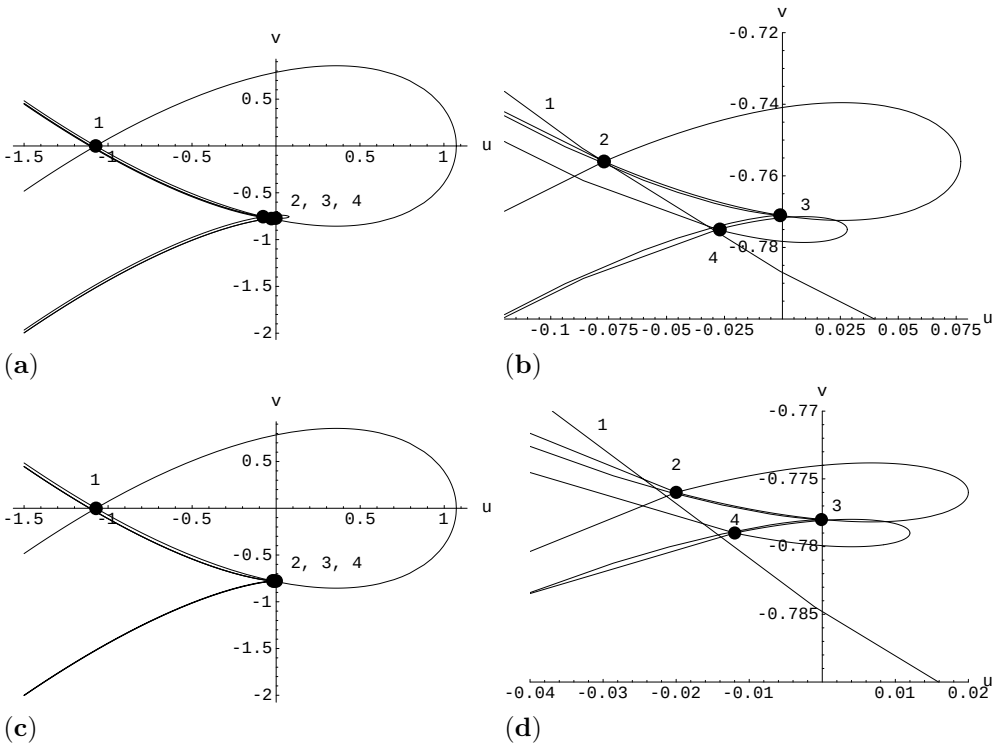


FIG. 7. *The plots in* (a)–(b) *show the theoretical shock and rarefaction curves through each of the four states for $(\bar{u}, \bar{v})$ at the leading triple point (see Figure 6). Their numerical values are given in Table 4.2. (The bars have been omitted from the axis labels.) The curves correspond almost exactly to those of a triple point with an expansion fan. The plots in* (c)–(d) *show similar shock and rarefaction curves for the second triple point. The states 2 and 4 lie slightly off the shock curve of 1; nevertheless, the overall agreement with the curves of a triple point with an expansion fan is excellent.*

plots show that the triple points with expansion fans that we observe numerically are consistent with theory.

To accelerate the convergence of the solution on a very refined grid, we partially

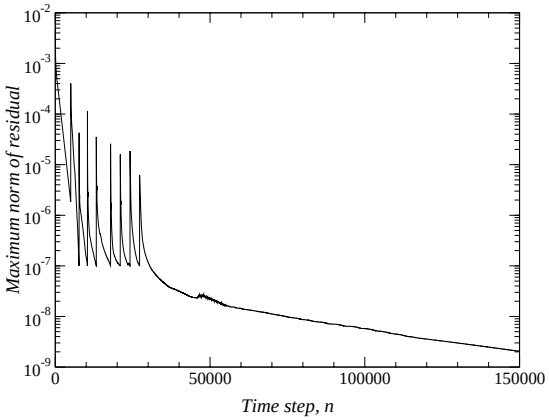


FIG. 8. *A plot of the maximum norm of the residual, showing partial convergence on a sequence of grids, followed by convergence on the most refined grid. The sharp local peaks correspond to interpolations onto more refined grids. The computation on the most refined grid begins at approximately* $n = 30000$. *The final stage of convergence to a value for the maximum norm of the residual of less than* 10^{-9} *is not shown in the plot.*

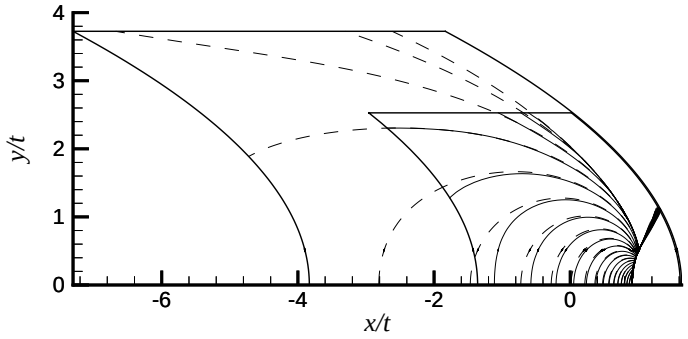


FIG. 9. *A check of the sensitivity of the solutions to the size of the numerical domain, showing* u*-contours for two solutions computed on different sized domains, for* $a = 0.5$. *The full numerical domains are shown, with* u*-contours for the large domain solution (dashed lines) and the small domain solution (solid lines) plotted at the same values of* u. *Contour lines for* u *and* v *near the triple point for both solutions shown here are compared in Figure* 10.

converged the solution on a coarse grid, interpolated the solution onto a refined grid, and repeated this process until the desired resolution was obtained. For example, Figure 4 shows a sequence of solutions obtained on four consecutive intermediate grids during the computation for $a = 0.5$. In Figure 8, we plot the maximum norm of the residual for a typical computation, in which nine grids were used. The sharp local peaks correspond to interpolations onto more refined grids. In the computation shown, the solution on each intermediate grid was converged until the maximum norm of the residual was less than 10^{-7}. The solution on the final grid in a computation was converged until no further change was observed in the details of the solution near the triple point, which typically occurred when the maximum norm of the residual was less than 10^{-9}.

We performed checks to determine the sensitivity of the solutions to the placement of the top and left numerical boundaries, which intersect the region where (2.8) is elliptic. In Figure 9, we plot u-contours for two solutions for $a = 0.5$ computed on

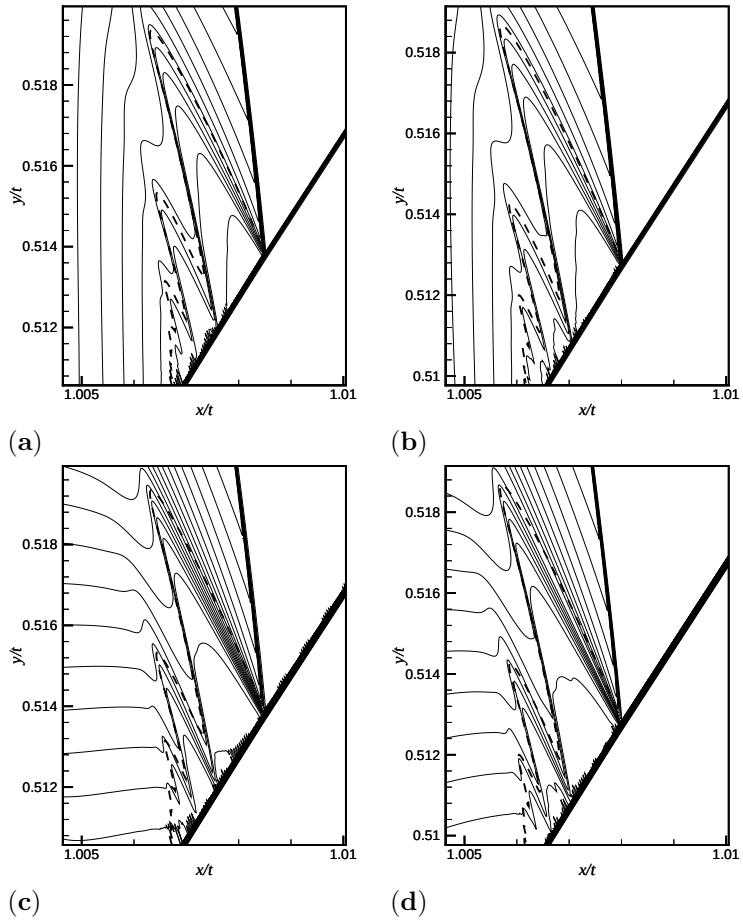


FIG. 10. *A comparison of u- and v-contours near the triple point for the two solutions shown in Figure* 9. *The plots in* (a) *and* (b) *show u-contours for the solutions computed on the larger and smaller domains, respectively, plotted at the same levels of u. The plots in* (c) *and* (d) *show v-contours for the solutions computed on the larger and smaller domains, respectively, plotted at the same levels of v. The dashed line in* (a)–(d) *is the sonic line. The u-contour spacing in* (a)–(b) *is* 0.005*, and the v-contour spacing in* (c)–(d) *is* 0.001.

different sized domains. In this study, the top and left numerical boundaries of the smaller domain were extended, as indicated in the figure, to approximately double the distance from these boundaries to the triple point. The contour lines are plotted at the same values of u for both solutions, with the dashed lines representing the u-contours of the solution on the larger domain. The contour lines approach each other and almost coincide near the triple point.

Figure 10 is an enlargement of the solutions near the triple point, showing u-contours and v-contours for the solutions on the larger and smaller domains. The u-contours in Figure 10(a)–(b) and the v-contours in Figure 10(c)–(d) are plotted at the same values of u and v, respectively, and the sizes of the regions shown in these plots are the same. The dashed line in Figure 10(a)–(d) is the sonic line. The structure of reflected shocks and expansion waves, supersonic patches, and repeating triple points did not change as a result of enlarging the computational domain, and the size of the supersonic region is nearly identical for the two solutions. The main

effect of extending the boundaries is a slight shift in the location of the leading triple point. The shift is approximately 0.05% in x/t and 0.2% in y/t.

5. Discussion. These numerical results raise the question of whether there is an infinite sequence of triple points in an inviscid weak shock Mach reflection. Gamba, Rosales, and Tabak [7] prove, under some mild assumptions, that the flow behind a triple point cannot be strictly subsonic for the unsteady transonic small disturbance equation. Therefore, if there were a finite sequence of supersonic triple points, there would presumably have to be a smooth transition from supersonic to subsonic flow at the rear of the final supersonic patch. Such a smooth transition appears unlikely to occur, however, because the resulting nonlinear mixed-type boundary value problem would be overdetermined [9, 15].

The most plausible alternative to a finite sequence of triple points terminated by a shock-free supersonic patch is an infinite sequence of more closely spaced triple points, weaker shock-expansion pairs, and smaller supersonic patches that accumulate at the rear sonic point of the supersonic region on the Mach shock. In this structure, the shock and expansion waves would reflect back and forth infinitely many times between the Mach shock and the sonic line, into the rear sonic point. The inviscid equations do not define a length scale so solutions may, in principle, develop arbitrarily small structures. We do not know, however, of a way to confirm or deny the existence of an infinite sequence of patches whose size shrinks to zero.

A remarkable feature of the numerical solutions is the extraordinarily small size of the supersonic region, especially for larger values of a. For example, when $a = 0.8$, the height of the supersonic region is approximately 0.05% of the height of the Mach shock. Once the inverse shock slope a is fixed, there are no further parameters in the problem, so the small size of the region cannot be explained by the dependence of the solution on a small parameter. The shock reflection pattern is produced by the requirement that the y-velocity component v, which is equal to $-a$ behind the incident shock, must return to zero at the wall $y = 0$. Thus, a global scale for v_η is

$$\alpha = \frac{a}{(y/t)_{\text{t.p.}}},$$

where $(y/t)_{\text{t.p.}}$ is the (y/t)-location of the triple point. The supersonic region is produced by the expansion fan that is formed when the leading reflected shock collides with the incident shock. If Δv is the change in v across this fan, then a local scale for v_η near the triple point is

$$\beta = \frac{\Delta v}{\Delta(y/t)},$$

where $\Delta(y/t)$ is the height of the supersonic region. From the numerical data, we find that α is much less than β for larger values of a, corresponding to a rapid change in the solution near the triple point and a tiny supersonic region. For example, when $a = 0.5$, we find from the numerical data that $\alpha/\beta \approx 1.0$, but when $a = 0.8$, we find that $\alpha/\beta \approx 0.05$. Since the largest value of a that we investigated is 0.8, we neither know if solutions for higher values of a contain a supersonic region with a sequence of triple points over the entire range $0.8 < a < \sqrt{2}$, nor know if the transition between regular and Mach reflection occurs exactly at $a = \sqrt{2}$.

A repeating structure of supersonic patches and triple points with expansion fans appears to provide a resolution of von Neumann's triple point paradox within

the framework of inviscid shock theory, and viscosity is not required to explain the structure of a weak shock Mach reflection. Nevertheless, in view of the extremely small size of the supersonic region, it is important to consider the likely effect of physical viscosity on the inviscid description. Since the triple point lies in the interior of the fluid, it is reasonable to expect that boundary layer effects do not influence the local structure of the solution. Thus, the main effect of viscosity is to thicken the shocks. If the size of the supersonic region is smaller than the viscous thickness of the reflected shock, then the sonic line is embedded inside the viscous profile of the reflected shock, and the local structure of the solution near the triple point is dominated by viscous effects. Since the numerical scheme includes numerical viscosity, which mimics the effect of physical viscosity, the plots in Figure 4 of the solution with increasing numerical resolution presumably indicate the effect of decreasing physical viscosity on the solution. At resolutions lower than the ones shown in Figure 4, the supersonic region disappears completely, and the sonic line runs down the inside of the reflected shock, through the triple point, and down the Mach shock.

To compare the width of the supersonic region with the viscous shock thickness, we suppose that the reflected shock Mach number is M_r and the mean free path in the gas is λ. The thickness δ of the reflected shock is then approximately given by [22]:

$$\delta = \frac{3\lambda}{M_r - 1}.$$

The incident and Mach shocks are thinner than the reflected shock because they are stronger. If the width of the supersonic region in x/t in the solution of the unsteady transonic small disturbance equation is $\Delta(x/t)$, then, from [12], the asymptotic width d of the supersonic region parallel to the wall in physical variables is given by

$$d = 2(M-1)\Delta(x/t)L.$$

Here, L is the distance traveled by the Mach shock along the wall, from the corner of the wedge to the reflection point, and M is the Mach number of the incident shock. Hence

$$\frac{d}{\delta} = c(M-1)^2\frac{L}{\lambda},$$

where the dimensionless constant c is defined by

$$c = \frac{2}{3}\Delta\left(\frac{x}{t}\right)[u]_r, \tag{5.1}$$

and $[u]_r$ is the ratio of the reflected and incident shock strengths,

$$[u]_r = \frac{M_r - 1}{M - 1}.$$

The supersonic region is much larger than the reflected shock structure if $d \gg \delta$, meaning that

$$L \gg \frac{\lambda}{c(M-1)^2}.$$

The value of c in (5.1) may be estimated from the numerical data in Table 4.1. The supersonic region is easier to observe for larger values of c, and the largest value of

c for the results obtained here is $c \approx 6 \times 10^{-5}$ for $a = 0.5$. For smaller values of a, the reflected shock becomes very weak and thick, while for larger values of a, the supersonic region becomes extremely small. The mean free path in argon at standard conditions is approximately $\lambda = 6 \times 10^{-5}$ mm. Therefore, for a shock reflection in argon with $a = 0.5$, we estimate that the supersonic region separates from the viscous profile of the reflected shock when $L \gg (M-1)^{-2}$ mm. Even for a relatively strong weak shock with $M = 1.1$, this estimate gives $L \gg 100$ mm. Thus, in order to observe the supersonic region in a shock tube experiment, the test section of the tube would have to be significantly longer than 100 mm.

It is striking that such a complex inviscid structure forms on a length scale that is comparable with, or less than, the viscous shock thickness in typical experiments.

6. Conclusion. We have presented numerical evidence of a structure of reflected shocks and expansion waves and a sequence of triple points and supersonic patches in a tiny region behind the leading triple point of an inviscid weak shock Mach reflection. The presence of the expansion fans at the triple points resolves the von Neumann paradox of weak shock reflection. Qualitative arguments, based on the well-posedness of mixed-type boundary value problems, suggest that there may be an infinite sequence of triple points and patches in an inviscid reflection, but a proof or disproof of this suggestion is lacking. The numerical solutions provide an estimate of the size of the supersonic region, which may enable its experimental detection.

REFERENCES

[1] J. Ackeret, F. Feldmann, and N. Rott, *Investigations of Compression Shocks and Boundary Layers in Fast Moving Gases*, report 10, ETH, Zurich, 1946.
[2] G. Birkhoff, *Hydrodynamics*, Revised ed., Princeton University Press, Princeton, NJ, 1960.
[3] W. Bleakney and A. H. Taub, *Interaction of shock waves*, Rev. Modern Phys., 21 (1949), pp. 584–605.
[4] M. Brio and J. K. Hunter, *Mach reflection for the two-dimensional Burgers equation*, Phys. D, 60 (1992), pp. 194–207.
[5] P. Colella and L. F. Henderson, *The von Neumann paradox for the diffraction of weak shock waves*, J. Fluid Mech., 213 (1990), pp. 71–94.
[6] M. van Dyke, *An Album of Fluid Motion*, Parabolic Press, Stanford, CA, 1982.
[7] I. M. Gamba, R. R. Rosales, and E. G. Tabak, *Constraints on possible singularities for the unsteady transonic small disturbance (UTSD) equations*, Comm. Pure Appl. Math., 52 (1999), pp. 763–779.
[8] K. G. Guderley, *Considerations of the Structure of Mixed Subsonic-Supersonic Flow Patterns*, Air Materiel Command technical report F-TR-2168-ND, ATI No. 22780, GS-AAF-Wright Field No. 39, U.S. Wright-Patterson Air Force Base, Dayton, OH, 1947.
[9] K. G. Guderley, *The Theory of Transonic Flow*, Pergamon Press, Oxford, 1962.
[10] L. F. Henderson, *Regions and boundaries for diffracting shock wave systems*, Z. Angew. Math. Mech., 67 (1987), pp. 73–86.
[11] J. K. Hunter, *Nonlinear geometrical optics*, in Multidimensional Hyperbolic Problems and Computations, IMA Vol. Math. Appl. 29, J. Glimm and A. Majda, eds., Springer-Verlag, New York, 1991, pp. 179–197.
[12] J. K. Hunter and M. Brio, *Weak shock reflection*, J. Fluid Mech., 410 (2000), pp. 235–261.
[13] H. Liepmann, *The interaction between a boundary layer and shock-waves in transonic flow*, J. Aeronautical Sciences, 13 (1946), pp. 623–637.
[14] C. S. Morawetz, *Potential theory for regular and Mach reflection of a shock at a wedge*, Comm. Pure Appl. Math., 47 (1994), pp. 593–624.
[15] C. S. Morawetz, *The mathematical approach to the sonic barrier*, Bull. Amer. Math. Soc. (N.S.), 6 (1982), pp. 127–145.
[16] J. von Neumann, *Collected Works*, Vol. 6, Pergamon Press, New York, 1963.
[17] R. Samtaney, *Computational methods for self-similar solutions of the compressible Euler equations*, J. Comput. Phys., 132 (1997), pp. 327–345.

[18] A. Sasoh, K. Takayama, and T. Saito, *A weak shock wave reflection over wedges*, Shock Waves, 2 (1992), pp. 277–281.
[19] J. Sternberg, *Triple-shock-wave intersections*, Phys. Fluids, 2 (1959), pp. 179–206.
[20] E. G. Tabak and R. R. Rosales, *Focusing of weak shock waves and the von Neumann paradox of oblique shock reflection*, Phys. Fluids, 6 (1994), pp. 1874–1892.
[21] A. M. Tesdall, *Self-Similar Solutions for Weak Shock Reflection*, Ph.D. Thesis, University of California at Davis, Davis, CA, 2001.
[22] P. A. Thompson, *Compressible Fluid Dynamics*, McGraw–Hill, New York, 1971.
[23] H. Q. Yang and A. J. Przekwas, *A comparative study of advanced shock-capturing schemes applied to Burgers' equation*, J. Comput. Phys., 102 (1992), pp. 139–159.
[24] A. R. Zakharian, M. Brio, J. K. Hunter, and G. Webb, The von Neumann paradox in weak shock reflection, J. Fluid Mech., 422 (2000), pp. 193–205.

SIAM J. APPL. MATH.
Vol. 63, No. 1, pp. 62–97

MULTIPLE BUMPS IN A NEURONAL MODEL OF WORKING MEMORY*

CARLO R. LAING†, WILLIAM C. TROY‡, BORIS GUTKIN§, AND G. BARD ERMENTROUT‡

Abstract. We study a partial integro-differential equation defined on a spatially extended domain that arises from the modeling of "working" or short-term memory in a neuronal network. The equation is capable of supporting spatially localized regions of high activity which can be switched "on" and "off" by transient external stimuli. We analyze the effects of coupling between units in the network, showing that if the connection strengths decay monotonically with distance, then no more than one region of high activity can persist, whereas if they decay in an oscillatory fashion, then multiple regions can persist.

Key words. short-term memory, integro-differential equation, coupling, homoclinic orbits

AMS subject classifications. 34B15, 34C23, 93C15, 34C11

PII. S0036139901389495

1. Introduction.

Working memory, which involves the holding and processing of information on the time scale of seconds, is a much studied area of neuroscience [3, 9, 24, 35, 37]. Experiments in primates [8, 15, 29] have shown that there exist neurons in the prefrontal cortex that have elevated firing rates during the period in which an animal is "remembering" the spatial location of an event before acting on the information being remembered. Realistic models for this type of activity have involved spatially extended systems of coupled neural elements and the study of spatially localized areas of high activity in these systems. Previous studies have involved "rate" models [1, 19, 22, 37] in which a neural element is described by a single scalar variable, e.g., a firing rate and more complicated "spiking" models [9, 24, 35], which take into account the intrinsic dynamics of single neurons.

In this paper we extend the 1977 work of Amari [1] who found *single* spatially localized regions of high activity ("bumps") in rate models of the form

$$\frac{\partial u(x,t)}{\partial t} = -u(x,t) + \int_{-\infty}^{\infty} w(x-y) f(u(y,t))\, dy + s(x,t) + h. \tag{1.1}$$

Equation (1.1) models a single layer of neurons. The function $u(x,t)$ denotes the "synaptic drive" or "synaptic input" to a neural element at position $x \in (-\infty, \infty)$ and time $t \geq 0$. The connection function $w(x)$ determines the coupling between elements, and the nonnegative function $f(u)$ gives the firing rate, or activity, of a neuron with input u. Neurons at a point x are said to be active if $f(u(x,t)) > 0$. The function $s(x,t)$ represents a variable external stimulus. Finally, the parameter h denotes a *constant* external stimulus applied uniformly to the entire neural field. Although

*Received by the editors May 16, 2001; accepted for publication (in revised form) February 15, 2002; published electronically August 5, 2002.
http://www.siam.org/journals/siap/63-1/38949.html
†Department of Physics, University of Ottawa, Ottawa ON, Canada K1N 6N5 (claing@science.uottawa.ca).
‡Department of Mathematics, University of Pittsburgh, Pittsburgh, PA 15260 (troy@vms.cis.pitt.edu, bard@pitt.edu).
§Unité de Neurosciences Intergratives et Computationelles, CNRS, Gif-sur-Yvete 91198, France (gutkin@cnbc.cmu.edu).

the model we study has been used to model working memory, similar equations arise in neural theory as applied to vision and robotic navigation [17], head direction systems [39], and cognitive development in infants [32]. We also mention recent analyses of wave propagation when inhomogeneities are present in the underlying neural substrate [4] and also in neural networks with axo-dendritic synaptic interactions [10].

Our goal is to extend Amari's results in two ways. First, in the next section we will extend the analysis of the original model in which $w(x)$ is assumed to have exactly one zero in $(0,\infty)$, and $f(u)$ is a step function. We will determine a simple set of assumptions on w and f for which (1.1) has stationary "single-bump" solutions. Our assumptions will allow us to obtain a more precise description of the shape of solutions. We will also investigate the existence of "double-bump" solutions.

In section 3 we relax the restrictions on w and f to include both oscillatory connection functions which change sign infinitely often and *continuous* firing rate functions. Our goal here is to show that "multi-bump" solutions of (1.1) exist over an appropriate range of parameters. The extension of $f(u)$ to a continuous function will allow us to derive an ordinary differential equation, specific solutions of which are steady-states of (1.1). This differential equation, which is derived in section 5, will be invaluable in proving the existence or otherwise of such "multi-bump" solutions. Sections 6 and 7 are devoted to studies of its N-bump solutions. In section 8 we extend the model to two space dimensions and present numerical evidence for multi–bumps solutions. Sections 9 and 10 contain proofs of two theorems stated in the text, and a summary of our results is given in section 11.

2. "Mexican hat" coupling. We begin with a description of the assumptions and conclusions obtained by Amari [1] where the coupling function $w(x)$ satisfies the following:

(H_1) $w(x)$ is symmetric, i.e., $w(-x) = w(x)$ for all $x \in \mathbf{R}$;
(H_2) $w(x) > 0$ on an interval $(-\bar{x}, \bar{x})$, and $w(-\bar{x}) = w(\bar{x}) = 0$;
(H_3) $w(x)$ is decreasing on $(0, \bar{x}]$;
(H_4) $w < 0$ on $(-\infty, -\bar{x}) \cup (\bar{x}, \infty)$.

An additional condition which Amari uses but does not explicitly state is

(H_5) w is continuous on $\mathbf{R}$, and $\int_{-\infty}^{\infty} w(y)\,dy$ is finite.

A coupling satisfying (H_2) and (H_4) produces "lateral inhibition" [14]. That is, condition (H_2) means that nearby neural elements excite one another, but (H_4) results in an "inhibitory effect" if the distance between neural elements is greater than a certain value, $\bar{x}$. Conditions (H_1), (H_3) and (H_5) are general requirements which allow for a tractable mathematical analysis of (1.1). In order to rigorously determine the shape of steady-state solutions of (1.1), we make one final assumption on the coupling function $w(x)$:

(H_6) $w(x)$ has a unique minimum on $\mathbf{R}^+$ at a point $x_0 > \bar{x}$, and $w(x)$ is strictly increasing on (x_0, ∞).

A connection function which satisfies conditions (H_1)–(H_6) is

$$w(x) = Ke^{-k|x|} - Me^{-m|x|}, \tag{2.1}$$

where $0 < M < K$ and $0 < m < k$. An example of this "Mexican hat" type function is given in Figure 1 for $K = 3.5$, $M = 3$, $k = 1.8$, and $m = 1.52$. For simplicity, Amari assumes (see Figure 1) that the firing rate $f(u)$ is the Heaviside step function

$$f(u) = \begin{cases} 0, & u \le 0, \\ 1, & u > 0. \end{cases} \tag{2.2}$$

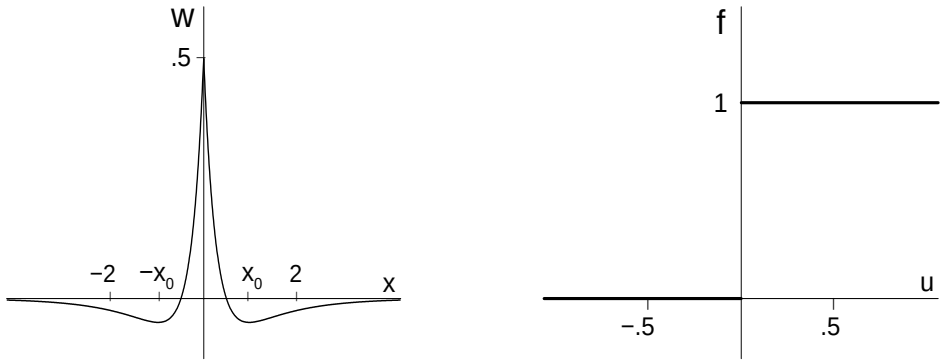


FIG. 1. *Mexican hat function* (2.1) *for parameters given in the text, and the Heaviside firing rate function* (2.2).

The effect of (2.2) is that a neuron fires at its maximum rate when the input exceeds the threshold value $u = 0$ and does not fire otherwise. Thus, (2.2) can be viewed as modeling neural elements whose firing rates "saturate" immediately, since increasing the input further does not cause the firing rate to increase, provided the input is above the threshold value.

Under assumptions (H_1)–(H_5), Amari analyzes the existence and stability of equilibrium solutions of (1.1) under the assumption that there is no "inhomogeneous" external stimulus $s(x,t)$. That is, he sets $\partial u(x,t)/\partial t = 0$ and $s(x,t) = 0$. This reduces (1.1) to the time independent equation

$$u(x) = \int_{-\infty}^{\infty} w(x-y) f(u(y))\, dy + h. \tag{2.3}$$

Solutions of (2.3) are called equilibrium or *stationary* solutions. An important observation is that the neural system is still subject to the constant external stimulus h applied uniformly to the entire neural field. Note that if $h \leq 0$, then the constant function $u = h$ is a solution of (2.3).

Single-bump solutions: For a given distribution $u(x)$, Amari defines its region of excitation to be the set

$$R(u) = \{x | u(x) > 0\}.$$

He then defines a *localized excitation* to be a pattern $u(x)$ whose region of excitation is a *finite* interval, i.e., $R(u) = (a_1, a_2)$. If $R(u)$ is connected, we refer to the pattern as a "single-bump", or "1-bump" solution. Furthermore, because (2.3) is homogeneous, it is easily verified that $u(x-a)$ is a solution whenever $u(x)$ is a solution. Thus, without loss of generality, we assume that the region of excitation for a single-bump solution has the form

$$R(u) = (0, a).$$

Remark. If (2.3) has a solution whose region of excitation consists of $N > 1$ disjoint, finite connected intervals, the solution is called an *N-bump* solution. A major goal of this paper is to show that multi-bump solutions exist for (2.3) when the restrictions on $w(x)$ and $f(u)$ are relaxed.

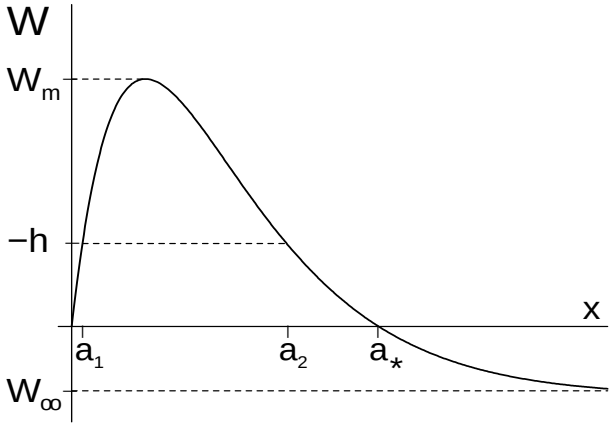


FIG. 2. $W(x)$, (2.4), *for parameters given in the text. We have chosen* h *to be negative, so that* $W_\infty < 0 < -h < W_m$.

In his analysis of single-bump solutions, Amari makes use of the function

$$W(x) = \int_0^x w(y)\,dy \tag{2.4}$$

and the related quantities

$$W_m = \max_{x>0} W(x) \quad \text{and} \quad W_\infty = \lim_{x\to\infty} W(x). \tag{2.5}$$

Conditions (H_1) and (H_5) imply that $W(x)$ is odd, and that W_∞ is finite, respectively. Amari observes that if (2.3) has a single-bump solution $u(x)$ whose region of excitation is given by $R(u) = (0, a)$, then $u(x)$ satisfies

$$u(x) = \int_0^a w(x-y)\,dy + h = W(x) - W(x-a) + h. \tag{2.6}$$

At the value $x = a$, (2.6) reduces to

$$W(a) = -h \tag{2.7}$$

since $W(x)$ is odd and $u(0) = u(a) = 0$. In turn, Amari claims that if $a > 0$ and $h < 0$ satisfy (2.7), then

$$u(x) = W(x) - W(x-a) + h \tag{2.8}$$

is a single-bump solution of (2.3) for which $R(u) = (0, a)$.

For a given $h \le 0$, (2.7) may have zero, one or two positive solutions. The exact number is determined by the relative values of W_∞, W_m, and h. In Figure 2 we construct the $W(x)$ corresponding to the Mexican hat function illustrated in Figure 1. That is, we use the formula for $w(x)$ given in (2.1) for the specific values $K = 3.5$, $k = 1.8$, $M = 3$, and $m = 1.52$. In Figure 2 we see that if $W_\infty < 0 < -h < W_m$, then there are two values, a_1 and a_2, which satisfy (2.7). Setting $a = a_1$ and $a = a_2$ in (2.8)

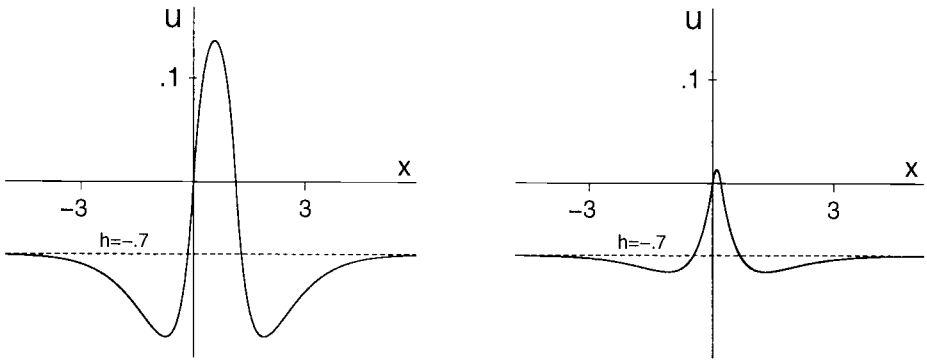


FIG. 3. *Stable (left) and unstable (right) single-bump solutions of* (2.3) *for the functions* w *and* f *shown in Figure* 1 *and* $h = -0.7$.

gives the corresponding single-bump solutions of (2.3). In Figure 3 we illustrate these two solutions for the value $h = -0.7$. Amari gives arguments that indicate that the large amplitude solution corresponding to $a = a_2$ (i.e., the first solution in Figure 3) is stable, while the second, smaller amplitude solution in Figure 3 corresponding to $a = a_1$ is unstable. Furthermore, as Figure 2 indicates, if $h = 0$, then (2.7) holds only at the positive value $a = a_2 = a_*$. Setting $a = a_*$ and $h = 0$ in (2.8), one can easily show that the resulting function is still a single-bump solution of (2.3).

We note that if (2.7) has a solution for some $a > 0$ and $h > 0$, then (2.8) implies that $u(x) > 0$ for all large x, contradicting the supposition that $R(u) = (0, a)$ is finite. Thus, single-bump solutions do not exist if $h > 0$.

Finally, we make a few observations concerning the *shape* of nonconstant single-bump solutions (see Figure 3). First, we conclude from hypotheses (H_1)–(H_4) and (2.8) that $u(x)$ is symmetric with respect to $x = a/2$ and that $u(x)$ is increasing on $(0, a/2)$ and decreasing on $(a/2, a)$. When we consider the additional hypotheses (H_5) and (H_6), it follows from standard analysis that the solution $u(x)$ has a unique minimum on $(0, \infty)$, and that $u(x) \to h$ from below as $x \to \infty$.

Double-bump solutions: We now consider the possible existence of double-bump solutions. A solution $u(x)$ of (2.3) is called a double-bump, or 2-bump, solution if there are values $0 < a < b < c$ such that

$$\begin{cases} u > 0 \quad \text{on} \quad (0, a) \cup (b, c), \\ u(0) = u(a) = u(b) = u(c) = 0, \\ u < 0 \quad \text{otherwise.} \end{cases} \tag{2.9}$$

Thus, a 2-bump solution is one whose region of excitation consists of two disjoint, connected intervals. The quantity $b - a$ is the distance between bumps. Our goal is to prove the existence or nonexistence of double-bump solutions of (2.3) which satisfy property (2.9). In general, a rigorous resolution of this problem is very difficult. Before stating our first result, we recall that x_0 denotes the unique positive value at which the coupling function $w(x)$ attains its global minimum and that $w(x)$ is strictly increasing on (x_0, ∞) (see Figure 1). In the following result we eliminate a class of 2-bump solutions.

THEOREM 2.1. *Under hypotheses* (H_1)–(H_6) *there is no value* $h \in \mathbf{R}$ *for which the problem* (2.2)–(2.3) *has a* 2*-bump solution such that the distance between bumps satisfies* $b - a \geq x_0$.

Remark. Theorem 2.1 does not completely eliminate the existence of all double-bump solutions. For example, our proof does not address the existence of general 2-bump solutions such that the distance $b - a$ satisfies $b - a < x_0$. However, it can be shown that under the assumptions $c - b = a$, i.e., equal width bumps, and $W_\infty < 0$, (2.2)–(2.3) can support (possibly unstable) 2-bump solutions [33] (and see [18]). We also have no results concerning existence or nonexistence of N-bump solutions where $N \geq 3$. The resolution of these problems remains open.

Because the proof of Theorem 2.1 is somewhat technical, we postpone the details until section 9. We proceed in the next section to describe the main focus of our investigation.

3. Statement of main results. The main goal of our investigation is to extend the analysis in section 2 and determine conditions on the connection and firing functions so that the integral equation (1.1) has stable *N-bump* solutions. For this we choose a specific $w(x)$ which changes sign infinitely often, and we let $f(u)$ be a continuous extension of the Heaviside function. For simplicity it is assumed that both $s(x,t) = 0$ and $h = 0$. Setting $h = 0$ will be compensated for by including a threshold in f. Thus, we study the problem

$$\frac{\partial u(x,t)}{\partial t} = -u(x,t) + \int_{-\infty}^{\infty} w(x-y) f(u(y,t))\, dy, \tag{3.1}$$

where

$$w(x) = e^{-b|x|}(b \sin|x| + \cos x) \tag{3.2}$$

and

$$f(u) = 2e^{-r/(u-th)^2} H(u - th). \tag{3.3}$$

Here $th > 0$, $b > 0$, and $r > 0$ are constants. The parameter b controls the rate at which the oscillations in w decay with distance. As shown in Figure 4, they decay more rapidly as b is increased. It is hoped that this oscillatory form of coupling better represents the connectivity known to exist in the prefrontal cortex, where labeling studies have shown that coupled groups of neurons form spatially approximately periodic stripes [16, 26, 27]. Interestingly, it has been proposed that disruption of this "lattice" of connectivity may be responsible for some of the symptoms of schizophrenia [27]. Note that we are not addressing the processes involved in the *formation* of these stripes, but are interested in the possible patterns of neural activity that can exist in the system once these patterns are in place. Also, although $w(x)$ does not have finite support we know that in the brain, connections cannot exist over arbitrarily large distances, so this is obviously an approximation to reality. It would be an interesting problem to analyze (3.1) with a function $w(x)$ that had more than one zero crossing for $x > 0$ yet had finite support. Finally, it is interesting to observe that the coupling given in (3.2) is differentiable at $x = 0$, and that $w'(0) = 0$. This is proved in [23] and easily follows from the formal definition of derivative. In contrast, the lateral inhibition coupling given in (2.1) is not differentiable at $x = 0$. However, we believe that the only significant feature for analysis of the models is continuity of w at $x = 0$.

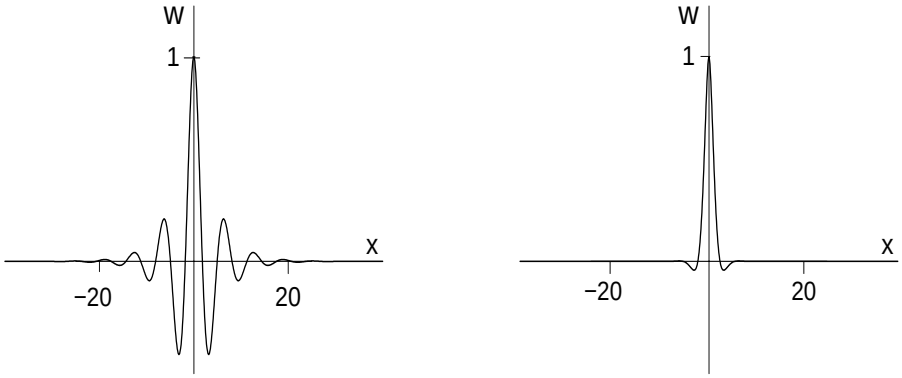


FIG. 4. $w(x)$, (3.2), *for* $b = 0.25$ *(left) and* $b = 1.0$ *(right).*

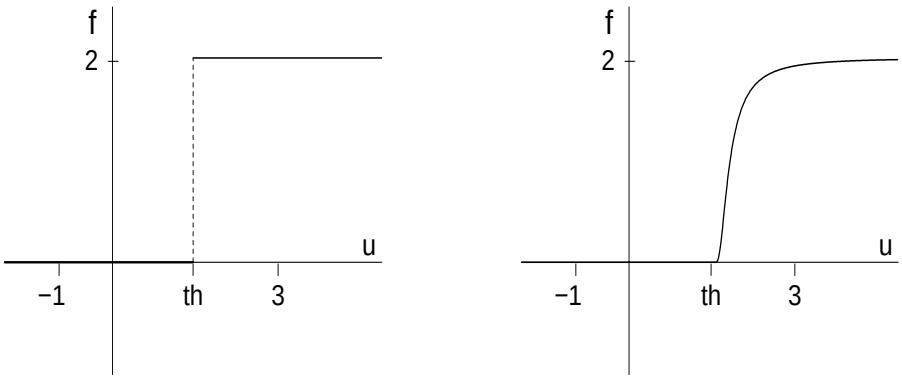


FIG. 5. $f(u)$, (3.3), *for* $r = 0$ *(left) and* $r = 0.1$ *(right), with* $th = 1.5$.

The parameter th denotes the *threshold* that is now included in $f(u)$. The coefficient of 2 in (3.3) was chosen merely for convenience. We note that $f(u) = 0$ if and only if $u \leq th$. Furthermore, $f(u)$ is a C^∞ function when $r > 0$, and r controls the rate of increase of $f(u)$ for u just past threshold. The differentiability of f will be useful when we derive a differential equation, specific solutions of which are equivalent to steady-state solutions of (3.1). In Figure 5 we set $th = 1.5$ and graph $f(u)$ for $r = 0$ (left) and $r = 0.1$ (right). When $r = 0$, $f(u)$ is just twice the Heaviside function. For $r > 0$, $f(u)$ is a *continuous* function which rapidly approaches 2 from below as u increases past th.

The choice of the functions (3.2) and (3.3) had some arbitrariness to it. The important features of (3.3) are that $f(u) = 0$ for $u \leq th$ and that $f(u)$ is sufficiently differentiable. The choice of (3.2) was made not only because it has the appropriate shape (decaying oscillations, with approximately the same distance between successive maxima), but also because the form of its Fourier transform makes the ordinary differential equation derived in section 5 particularly simple. Our hope is that the

qualitative details of the following results do not depend on the exact form of (3.2) and (3.3).

As before, we define a "stationary solution" to be a time independent solution of (3.1)–(3.3). Thus, a stationary solution satisfies the equation

$$u(x) = \int_{-\infty}^{\infty} w(x-y) f(u(y)) dy. \tag{3.4}$$

Before proceeding with our study of N-bump stationary solutions, we need to make precise the definition of the "region of excitation." For a solution of (3.4), we define its region of excitation to be the set

$$R(u) = \{x | u(x) > th\}. \tag{3.5}$$

A solution of (3.4) is an N-bump solution if its region of excitation consists of exactly N disjoint, finite connected intervals.

In the next section we begin our investigation of N-bump stationary solutions by considering the limiting value $r = 0$. As $r \to 0^+$ we note that the firing function tends to the discontinuous step function depicted in Figure 5 (left). In sections 5–7 we extend our studies to the case $r > 0$, for which the firing function $f(u)$ is *continuous.* As mentioned above, when $r > 0$ we find that there is an equivalent differential equation, some of whose solutions are solutions of (3.4). In section 5 we derive this fourth order equation and state our second theorem which determines a range of parameter values over which N-bump solutions can possibly exist. The differential equation will be especially useful to us in sections 6 and 7 where we give an extensive numerical investigation of the global behavior of entire families of N-bump solutions as parameters vary. Section 6 consists of a study of families of N-bump solutions for odd values of N, while section 7 covers even values of N.

4. The limiting problem: $r = 0$. It is natural to begin our investigation by considering the case $r = 0$ where $f(u)$ reduces to a multiple of the Heaviside function. In order to understand this case, we investigate the existence of N-bump solutions for a specific choice of the parameters b and th. For convenience we set $th = 1.5$ and $b = 0.25$ (see Figures 4 and 5 in the previous section). At these values our computations suggest that the problem (3.1)–(3.3) has at least four *stable* N-bump solutions. These are shown on the *left* in Figures 6–9, where the initial profile $u(x, 0)$ is represented by the dashed curve, and the solid curve represents $u(x, t)$ at $t = 60$. The formula for $u(x, 0)$ is given by

$$u(x,0) = \cos\left(\frac{Lx}{12.5\pi}\right) \exp\left(-\left(\frac{Lx}{12.5\pi}\right)^2\right), \quad -12.5\pi < x < 12.5\pi. \tag{4.1}$$

The parameter $L > 0$ allows us to vary the initial profile $u(x, 0)$. Equation (3.1) was numerically solved by spatially discretizing it on a uniform grid and then moving forward in time with an Euler step until convergence. The integral was approximated by a Reimann sum; note that the convolution can be performed more efficiently with a fast Fourier transform.

In the left panel in Figures 6 and 7 we set $L = 6$ and $L = 2.5$, and find that $u(x, t)$ approaches stable 1-bump and 2-bump solutions, respectively, as $t \to \infty$. Our computations imply that these also are solutions of

$$u(x) = \int_{-\infty}^{\infty} w(x-y) f(u(y)) dy. \tag{4.2}$$

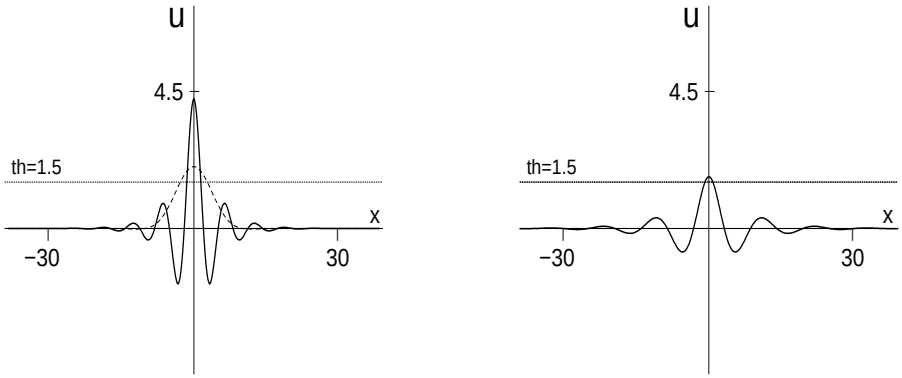


FIG. 6. *Stable (left) and unstable (right)* 1*-bump solutions:* $r = 0,\ th = 1.5,\ b = 0.25$.

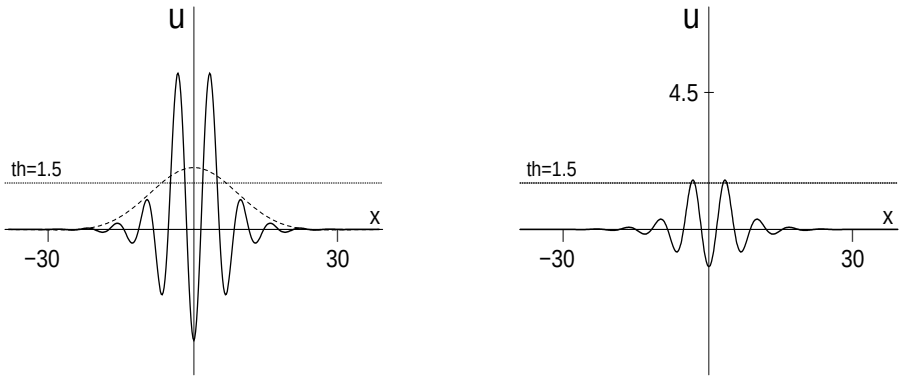


FIG. 7. *Stable (left) and unstable (right)* 2*-bump solutions:* $r = 0,\ th = 1.5,\ b = 0.25$.

Our computations also indicate that there exist *unstable* 1-bump and 2-bump stationary solutions. These are shown in the right panel in Figures 6 and 7. It is interesting to compare these unstable solutions with the unstable single-bump solution of the original Amari model described in section 2 (see Figure 3). Some of the stable solutions in Figures 6–9, Figures 14–17, Figure 19, Figure 23, Figures 25–28, and Figure 30 were found by numerically integrating (3.1) to a steady state, and the continuation program Auto97 [12, 13] was used to find the unstable solutions and reconfirm some of the stable solutions already found. We provide more detail in section 6.

Even though the system (3.1)–(3.3) is defined on an infinite domain, when numerically integrating (3.1) it must be finite. We have chosen a domain size of 25π, centered at $x = 0$. While it is unlikely that the boundaries have a significant effect on the spatially localized solutions shown in Figures 6 and 7, they will have a greater effect on broader solutions such as those in Figures 8 and 9. When comparing homoclinic orbits for the differential equation derived in section 5 (which represent solutions on an infinite domain) with solutions obtained from the numerical integration of (3.1),

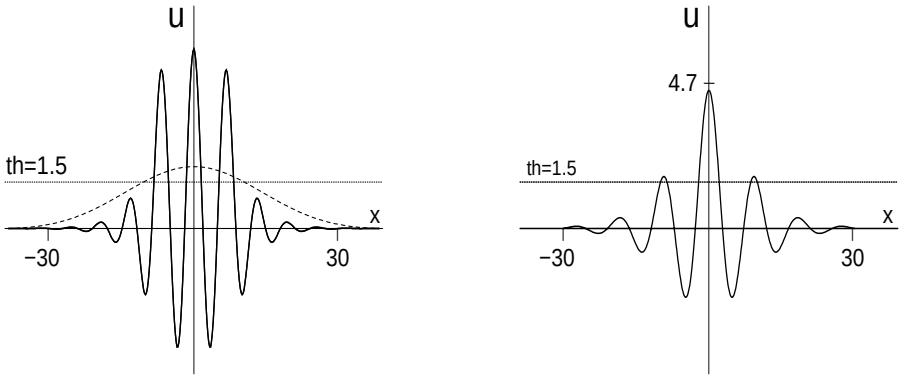


FIG. 8. *Stable (left) and unstable (right)* 3*-bump solutions:* $r = 0,\ th = 1.5,\ b = 0.25$.

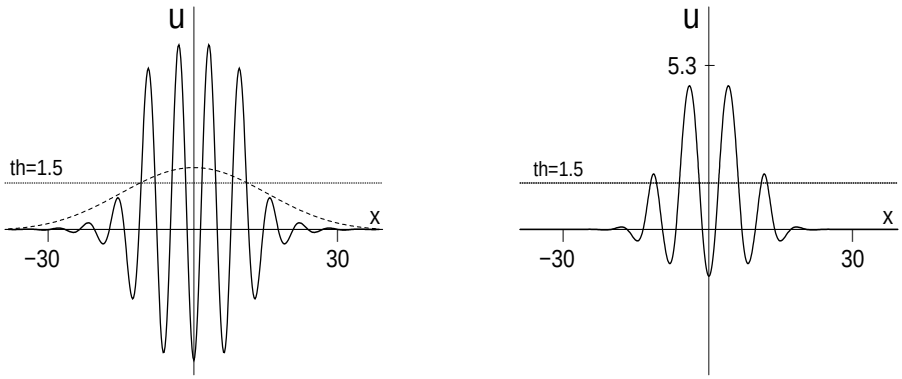


FIG. 9. *Stable (left) and unstable (right)* 4*-bump solutions:* $r = 0,\ th = 1.5,\ b = 0.25$.

the difference in domains should be kept in mind.

In the left panel of Figures 8 and 9 we let $L = 1.6$ and $L = 1.5$, respectively, and found that $u(x,t)$ tended to stable 3-bump and 4-bump stationary solutions as $t \to \infty$. Again, our computations indicate that there exist corresponding *unstable* 3-bump and 4-bump stationary solutions. These are shown in the right panels of Figures 8 and 9. Although we do not show the results, our computations indicate that if $L = 1$, then $u(x,t)$ tends to a stable 5-bump stationary solution as $t \to \infty$. For the values $r = 0$, $b = 0.25$, and $th = 1.5$, and a sufficiently large domain, we conjecture that both stable and unstable N-bump stationary solutions exist for each $N \geq 1$. We leave the resolution of this conjecture as an open problem.

We now develop a necessary mathematical criterion for the existence of 1-bump solutions of (4.2) when $r = 0$. In this case the firing function $f(u)$ defined in (3.3) reduces to twice the Heaviside function, as shown in the left panel of Figure 5. The solutions computed in Figures 6–9 are symmetric with respect to $x = 0$. Thus, we first look for single-bump *symmetric* solutions. We assume that there is a value $a > 0$

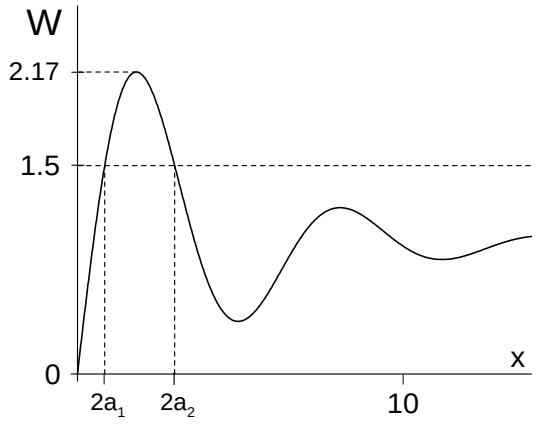


FIG. 10. $W(x)$, (4.4): $th = 1.5,\ b = 0.25$.

such that $u(x) > th$ on $(-a, a)$ and $u(x) < th$ if $|x| > a$. Under these assumptions, (4.2) reduces to

$$u(x) = \int_{-a}^{a} 2w(x-y)\,dy. \tag{4.3}$$

In analogy with section 2, we define

$$W(x) \equiv \int_{0}^{x} 2w(y)\,dy, \tag{4.4}$$

and note that $W(0) = 0$. From (4.3) and (4.4) it follows that

$$u(x) = W(x+a) - W(x-a). \tag{4.5}$$

Thus, we conclude that the condition $u(a) = th$ can be written as

$$W(2a) = th. \tag{4.6}$$

Figure 10 shows that, when $b = 0.25$ and $th = 1.5$, there are exactly two positive values, a_1 and a_2, for which (4.6) is satisfied. In Figure 11 we keep $th = 1.5$ and decrease b from $b = 0.25$. The left panel shows that there is a critical $b \approx 0.057$ at which a third value $a = a_3$ appears which satisfies $W(2a_3) = th$. For $0 < b < 0.057$ there are *at least* four solutions of (4.6). For example, we set $b = 0.03$ and illustrate this property in the right panel of Figure 11. As b decreases further, the number $\nu = \nu(b)$ of solutions of (4.6) (i.e., the number of symmetric 1-bump solutions of (4.2)) continues to increase, with $\nu(b) \to +\infty$ as $b \to 0^+$. In Figure 12 we see that the number $\nu(b)$ of solutions of (4.6) also increases if we keep b fixed at $b = 0.25$ and then lower the value of th from $th = 1.5$. Here we find that there is a critical value $th^* \equiv W(\infty) = 4b/(b^2+1)$ such that $\nu(b) \to +\infty$ as $th \to th^*$. We conjecture that each solution of (4.6) corresponds to a single-bump solution of the integral equation

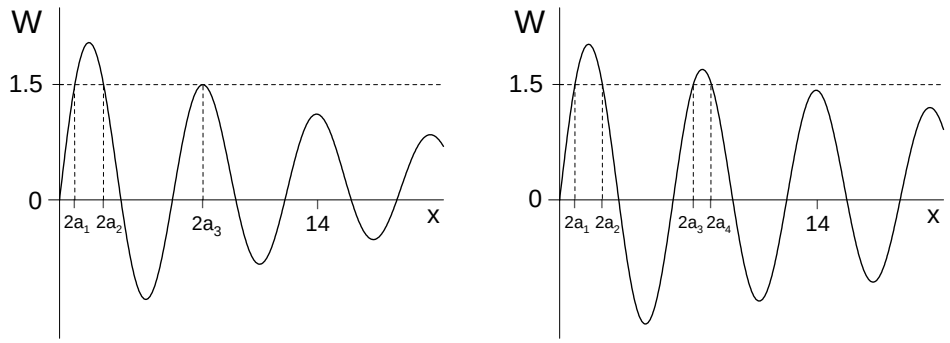


FIG. 11. $W(x)$, (4.4): $th = 1.5$; $b = 0.057$ *(left) and* $b = 0.03$ *(right).*

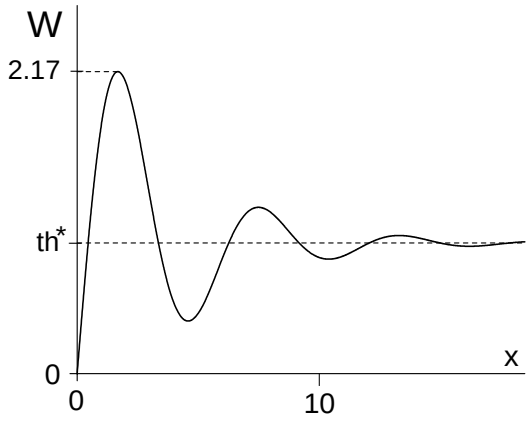


FIG. 12. $W(x)$, (4.4): $b = 0.25$, $th^* = 0.94$.

(4.2). In order to prove this conjecture, one would need to check that for all values of a satisfying (4.6), the function defined in (4.5) satisfied

$$u(x) > th \text{ for } -a < x < a \text{ and } u(x) < th \text{ for } x < -a \text{ or } a < x, \tag{4.7}$$

i.e., that the form of $u(x)$ given in (4.5) is actually a 1-bump solution. It would also be interesting to develop a criterion for the existence of N-bump solutions when $N > 1$. We leave these questions as open problems for future research.

5. The continuous case: $r > 0$. We now turn to the case $r > 0$, for which $f(u)$ is a *continuous* function. Thus, we study the existence of N-bump solutions of the equation

$$u(x) = \int_{-\infty}^{\infty} w(x-y) f(u(y))\, dy, \tag{5.1}$$

where $w(x)$ is given in (3.2) and $f(u)$ is given by (3.3), with $r > 0$. When $r > 0$, both the mathematical and computational analysis of (5.1) become more tractable. This is due to the fact that N-bump solutions of an associated differential equation problem also are solutions of (5.1). To derive the differential equation we make use of the Fourier transform, defined by

$$\mathrm{F}(g) = \int_{-\infty}^{\infty} e^{-i\alpha\eta} g(\eta)\, d\eta, \tag{5.2}$$

where $g \in L^1(\mathbf{R})$ and $\alpha \in \mathbf{R}$. Note that $\mathrm{F}(g)$ is a function of α.

We assume that u is a solution of (5.1), that u, u', u'', u''', and u'''' are continuous on $\mathbf{R}$, and that

$$(u, u', u'', u''') \to (0, 0, 0, 0) \tag{5.3}$$

exponentially fast as $x \to \pm\infty$. Under these assumptions, an application of the Fourier transform to (5.1) is justified and gives

$$\mathrm{F}(u) = \mathrm{F}(w)\mathrm{F}(f(u)). \tag{5.4}$$

An evaluation of $\mathrm{F}(w)$ converts (5.4) to

$$\mathrm{F}(u) = \frac{4b(b^2+1)}{\alpha^4 + 2\alpha^2(b^2-1) + (b^2+1)^2}\mathrm{F}(f(u)). \tag{5.5}$$

Next, multiply both sides of (5.5) by the denominator of $\mathrm{F}(w)$ and use the identities

$$\mathrm{F}(u'''') = \alpha^4\mathrm{F}(u) \quad \text{and} \quad \mathrm{F}(-u'') = \alpha^2\mathrm{F}(u) \tag{5.6}$$

to obtain

$$\mathrm{F}[u'''' - 2(b^2-1)u'' + (b^2+1)^2 u - 4b(b^2+1)f(u)] = 0. \tag{5.7}$$

We claim that (5.7) is satisfied if u is a solution of the problem

$$\begin{cases} u'''' - 2(b^2-1)u'' + (b^2+1)^2 u = 4b(b^2+1)f(u), \\ \lim_{x\to\pm\infty}(u, u', u'', u''') = (0, 0, 0, 0). \end{cases} \tag{5.8}$$

Because $r > 0$, it follows from the definition of $f(u)$ and standard analysis that if u is a solution of (5.8), then u, u', u'', u''', and u'''' are continuous on $\mathbf{R}$, hence (5.7) holds. It then follows that properties (5.4)–(5.7) also hold. From this we conclude that any solution of (5.8) also is a solution of the integral equation (5.1). This reduces the problem of finding N-bump solutions of (5.1) to the study of N-bump solutions of (5.8).

The first goal of our investigation of (5.8) is to extend the results of the previous section where we considered the special case $r = 0$. Thus, we keep $th = 1.5$ and choose an $r > 0$. Our numerical experiments for the case $r = 0$ indicate the existence of even solutions. Thus, when $r > 0$ we will restrict our attention to even solutions of (5.8). These satisfy

$$u'(0) = u'''(0) = 0. \tag{5.9}$$

In the next two sections we use the program AUTO97 [12, 13] to obtain an understanding of the global behavior of families of N-bump solutions of (5.8) as the parameter b varies.

Our second goal is to give global estimates on the range of r, th, and b for which N-bump solutions of (5.8) can exist. We have the following result.

THEOREM 5.1. *Let $r > 0$ and $th > 0$. If there is a value $b > 0$ for which* (5.8) *has a nonconstant solution, then*

$$0 < b \leq \frac{4 + \sqrt{|16 - th^2|}}{th}. \tag{5.10}$$

Remarks. (i) It would be interesting to extend the results of Theorem 5.1 to the special case $r = 0$. When $r = 0$ the function $f(u)$ is discontinuous and the differential equation in (5.8) no longer has a continuous right-hand side. However, since $f(u)$ will now be piecewise constant and the left-hand side of the differential equation is linear, it may be possible to solve (5.8) over restricted domains, piecing together these solutions into a continuous solution for all $x \in (-\infty, \infty)$. We leave this as an open problem.

(ii) The proof of Theorem 5.1 will be postponed until section 10.

(iii) As will be seen in section 6, the upper bound for b in Theorem 5.1 is not particularly tight, but the main purpose of this theorem is to show that there do not exist nonconstant solutions for all positive b.

The differential equation in (5.8) is fourth order, and for $th > 0$ it has a fixed point at the origin. The eigenvalues of the linearization of (5.8) about the origin are $b \pm i$ and $-b \pm i$. Thus, in (u, u', u'', u''') phase space, solutions of (5.8) are homoclinic orbits leading to the bifocus-type fixed point $(u, u', u'', u''') = (0, 0, 0, 0)$ [25]. We note that the differential equation is not generic since the sum of the eigenvalues is zero for all parameter values. This is a simple consequence of the fact that the differential equation in (5.8) is conservative and, in fact, Hamiltonian. This is easily verified, since solutions $u(x)$ satisfy the first integral

$$u'u''' - \frac{(u'')^2}{2} - (b^2 - 1)(u')^2 + (b^2 + 1)^2 Q(u) = 0, \tag{5.11}$$

where $Q(u)$ is defined by

$$Q(u) \equiv \int_0^u \left(s - \left(\frac{8b}{b^2 + 1} \right) e^{-r/(s-th)^2} H(s - th) \right) ds. \tag{5.12}$$

We also note that the differential equation is *reversible* since it contains only even order derivatives.

In recent years, higher order reversible, Hamiltonian equations have played an increasingly important role in modeling pattern formation in physical systems. We mention, for example, the encyclopedic paper by Cross and Hohenberg [11] which describes a wide array of higher order scalar equations. In two recent survey papers, Champneys [5, 6] gives a dynamical systems approach to the analysis of multi-bump, homoclinic orbits in higher order reversible models arising in physics, fluid mechanics, and optics. We also mention the recent book by Peletier and Troy [30] in which methods of analysis of pattern formation in higher order equations are developed from the alternative topological shooting point of view. In the models considered in these works, families of N-bump homoclinic orbits often arise through a Hamiltonian–Hopf

bifurcation from a constant solution. Furthermore, in many of these models the terms involving u are polynomials of degree greater than one. Thus, these terms exhibit *superlinear* growth as $|u| \to \infty$. However, in the model proposed in this paper, the terms involving u exhibit only *linear* growth for large $|u|$. In addition, the rapidly increasing sigmoidal function $f(u)$ given in (3.3) is poorly approximated by polynomials. Finally, as we shall see in the next two sections, our numerical investigation of (5.8) indicates that families of N-bump solutions do not come into existence through a Hamiltonian–Hopf bifurcation from a constant solution. Because of these fundamental differences from other higher order equations, a rigorous proof of existence of N-bump solutions of problem (5.8) should prove to be a challenging problem.

6. Families of N-bump solutions: N odd. In this section we use AUTO97 [12, 13] to determine the global behavior of families of even 1-bump, 3-bump, and 5-bump solutions of the problem

$$\begin{cases} u'''' - 2(b^2-1)u'' + (b^2+1)^2 u = 4b(b^2+1)f(u) \\ \lim_{x\to\pm\infty}(u, u', u'', u''') = (0,0,0,0), \end{cases} \tag{6.1}$$

where

$$f(u) = 2e^{-r/(u-th)^2} H(u-th), \tag{6.2}$$

and $th > 0$, $b > 0$, and $r > 0$ are constants.

In Figure 13 we set $th = 1.5$ and $r = 0.095$, and let b vary, and compute the bifurcation curve for families of even 1-bump and 3-bump solutions of (6.1)–(6.2). The horizontal axis is b and the vertical axis gives the global maximum of u for the corresponding solutions. Figures 14–17 show solutions at specific points $P_0, \ldots, P_7$ on the curve.

Using MATLAB [28], we numerically integrate (3.1)–(3.3) to a steady state, choosing an initial condition which evolves, as $t \to \infty$, into a 1-bump solution at $b = 0.25$. This solution, which we conjecture to be stable, is labeled P_4 on the bifurcation diagram, and is illustrated in the right panel of Figure 16. We then use AUTO97 to continue this solution as b varies. Figure 13 shows 1-bump solutions along the lower branch ${\Gamma_1}^-$ *between* P_1 and P_3. We conjecture that these solutions are unstable. Solutions at P_2 and P_3 are shown in Figure 15. As b decreases along ${\Gamma_1}^-$, solutions cease to be 1-bump solutions at P_1 (the right panel in Figure 14). As b decreases towards zero, solutions acquire arbitrarily many bumps. For example, the point P_0 corresponds to the 3-bump solution shown in the left panel of Figure 14. Note that when $b = 0$, the only bounded even solution of the ordinary differential equation (ODE) in (6.1) is $u(x) = \cos x$, and it is to this that solutions tend as $b \to 0$.

Remark. The first solution in Figure 15 is computed at $b = 0.25$. As $r \to 0^+$, our computations indicate that this solution tends to the 1-bump solution shown in the right panel of Figure 6 in section 4.

Next, we consider the middle branch ${\Gamma_1}^+$ in Figure 13. Along ${\Gamma_1}^+$ we find a second family of 1-bump solutions, some of which we conjecture are stable, *between* P_5 and P_3. As b decreases along ${\Gamma_1}^+$, solutions cease to be 1-bump solutions at P_5 (the left panel in Figure 16). The solution in the right panel of Figure 16 was computed at $b = 0.25$. As $r \to 0^+$, our computations indicate that this solution is stable and tends to the 1-bump solution shown in the left panel of Figure 6 in section 4.

We let ${\Gamma_3}^-$ denote the upper branch of the diagram in Figure 13. Along this branch our computations indicate that solutions are *unstable* 3-bump solutions. Specific solutions at P_6 and P_7 are given in Figure 17. The solution at P_6 is computed

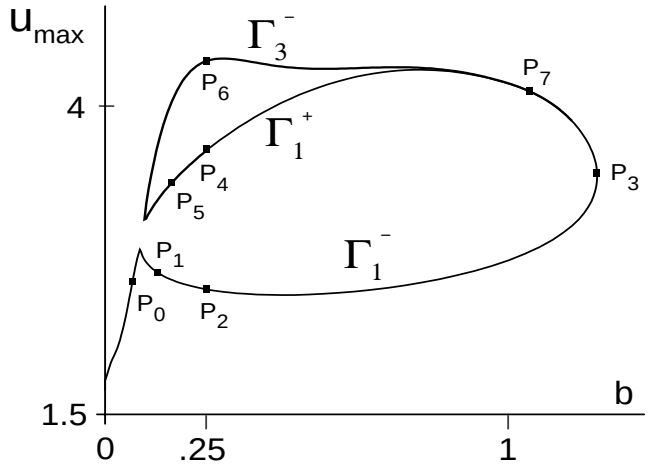


FIG. 13. *Bifurcation curve for* (6.1)–(6.2) *showing* 1*-bump and* 3*-bump solutions. Parameters are* $th = 1.5$ *and* $r = 0.095$. u_{max} *is the maximum of* u *over all* x. *Particular solutions at the points* $P_0, \ldots P_7$ *are shown in Figures* 14–17, *and the labeling of the curves is discussed in the text.*

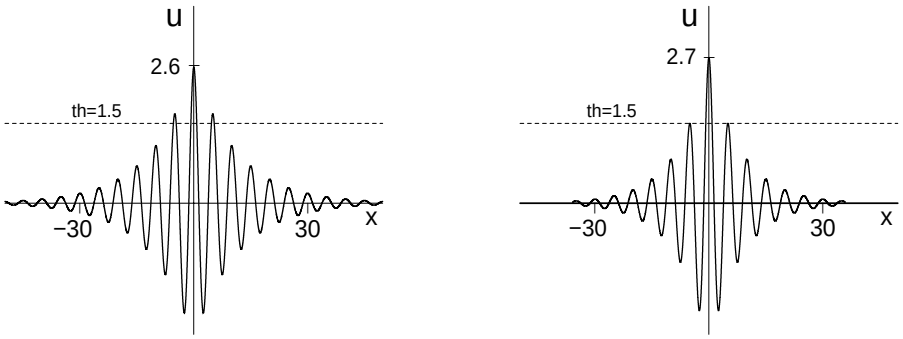


FIG. 14. *Solutions on* $\Gamma_1{}^-$ *at* P_0 *(left) and* P_1 *(right) in Figure* 13.

at $b = 0.25$, and as $r \to 0^+$ our computations indicate that it tends to the solution shown in the right panel of Figure 8.

We have also investigated the existence of 3-bump and 5-bump solutions. Our computations show that these solutions lie on yet another branch leading to the original bifurcation curve in Figure 13. This branch of solutions is labeled $\Gamma_3{}^+$ and $\Gamma_5{}^-$ in Figure 18. In Figure 19 we give specific solutions on $\Gamma_3{}^+$ and $\Gamma_5{}^-$ at $b = 0.25$. Our computations indicate that the solution in the left panel of Figure 19 is stable. Furthermore, as $r \to 0^+$ this solution tends to the solution in the left panel of Figure 8.

We can use data from Figures 13 and 18 to compare the largest values of b for which nonconstant solutions exist with the upper bound given in Theorem 5.1.

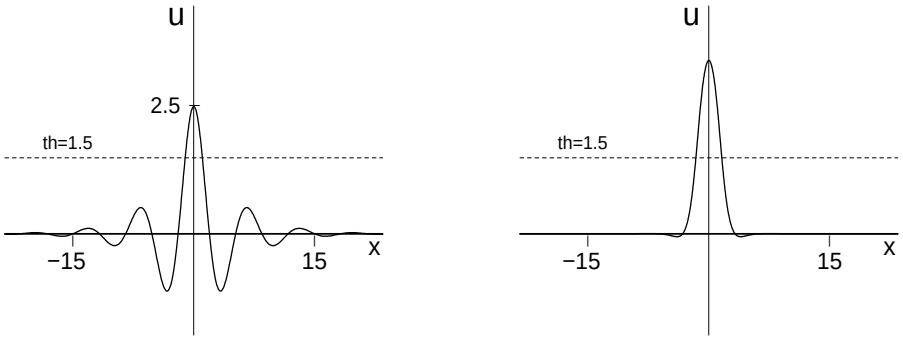


FIG. 15. *Solutions on* $\Gamma_1{}^-$ *at* P_2 *(left) and* P_3 *(right) in Figure* 13.

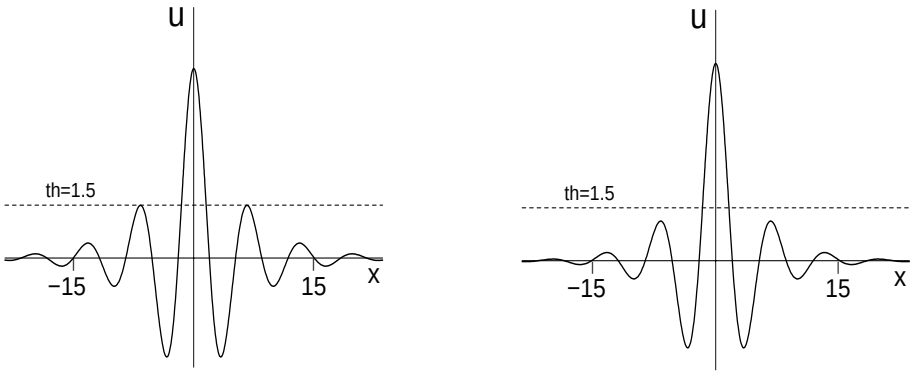


FIG. 16. *Solutions on* $\Gamma_1{}^+$ *at* P_5 *(left) and* P_4 *(right) in Figure* 13.

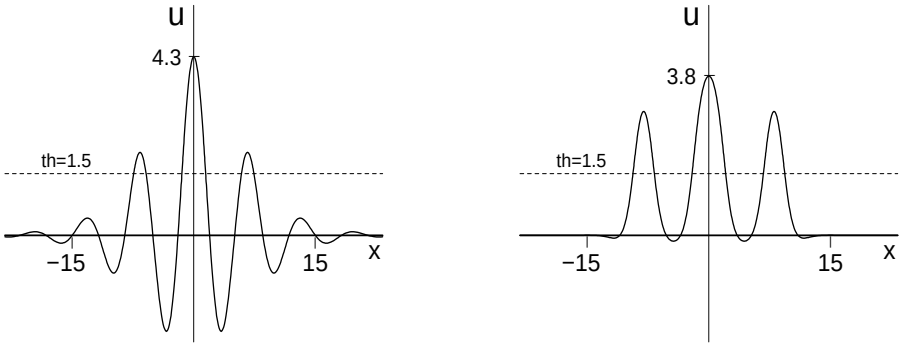


FIG. 17. *Solutions on* $\Gamma_3{}^-$ *at* P_6 *(left) and* P_7 *(right) in Figure* 13.

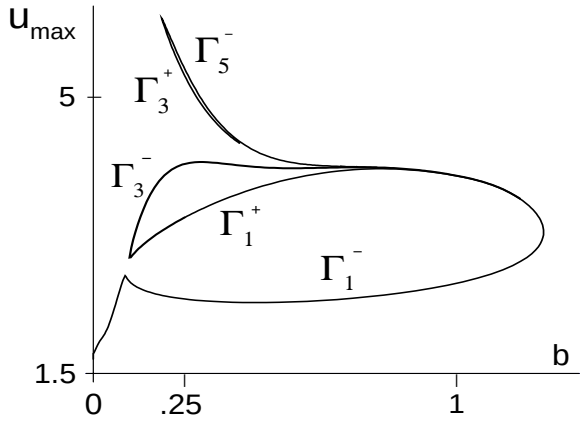


FIG. 18. *Bifurcation curve for* (6.1)–(6.2) *showing* 1, 3, *and* 5*-bump solutions. This Figure is an extension of Figure* 13.

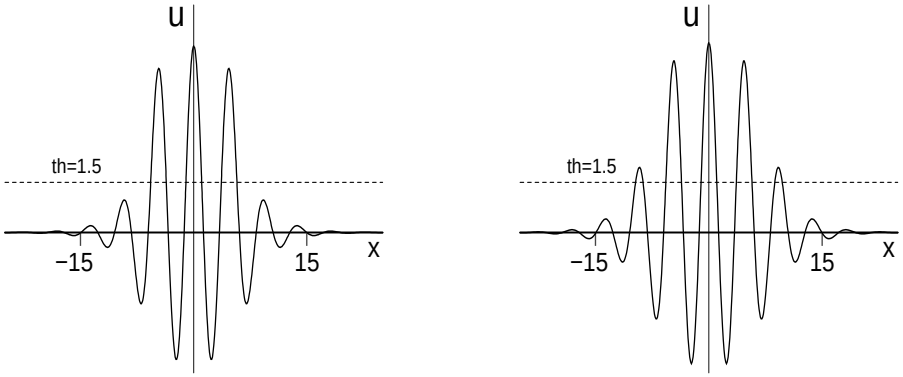


FIG. 19. *Solutions on the curves* $\Gamma_3{}^+$ *(left) and* $\Gamma_5{}^-$ *(right) at* $b = 0.25$ *in Figure* 18.

In Figure 20 we show saddle-node bifurcations of 1-, 3-, and 5-bump solutions in the b, th plane for $r = 0.095$. The curve γ_1 is the continuation of the point P_3 in Figure 13, and the curves γ_3 and γ_5 are the corresponding continuations for 3- and 5-bump homoclinic orbits, respectively. The dashed line (A) is the function given by the equality in (5.10), i.e., the value of b above which Theorem 5.1 states that no nonconstant solutions of (6.1)–(6.2) can exist. We see that the solutions studied in this section are compatible with Theorem 5.1, but that the bound given there is not particularly tight.

We have done one further experiment which shows how quickly the global behavior of solutions can change. In Figures 13 and 18 we set $th = 1.5$ and $r = 0.095$ and found that two "cusps" form on the left side of the bifurcation diagram. In Figure 21 we have increased r from $r = 0.095$ to $r = 0.1$ and repeated our computations. In

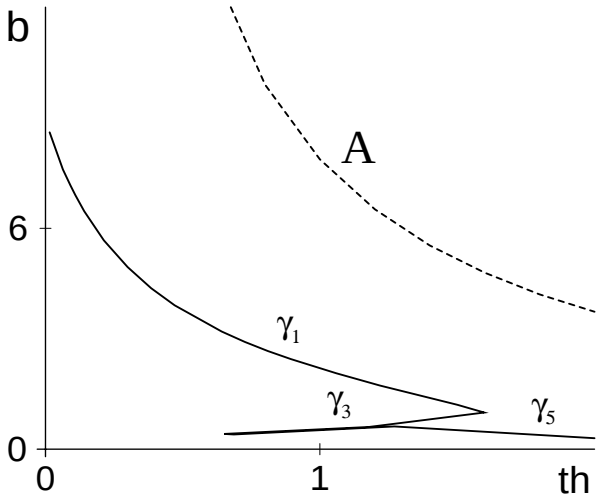


FIG. 20. *The continuation of the saddle-node bifurcations marking the largest values of b for which various orbits exist, compared with the upper bound given in Theorem* 5.1. γ_1 *is the continuation of the point* P_3 *in Figure* 13, *while* γ_3 *and* γ_5 *are continuations of the corresponding points for* 3- *and* 5-*bump homoclinic orbits. The curve "A" is the function* $b = (4+\sqrt{|16-th^2|})/th$, *given in* (5.10).

this case we find that the cusps have now joined and the 1-bumps solutions lie on an isolated closed curve. The lower branch Γ_1^- consists of small amplitude 1-bump solutions, which are conjectured to be unstable. The upper branch Γ_1^+ consists of large amplitude 1-bump solutions, some of which are conjectured to be stable. In order to see the separation of curves more clearly, in Figure 22 we have redrawn the bifurcation diagram of Figure 21 but now we have replaced u_{max} on the vertical axis with the L^2 norm of the solution (the default L^2 norm of AUTO97 is used). Figure 22 suggests that a "snaking" phenomenon occurs in the branches of the bifurcation curve and that solutions acquire more bumps as the L^2 norm increases (e.g., see Figure 23). Similar snaking phenomena occur in other physical systems modeled by higher order scalar equations [21, 30, 38], as well as in systems where homoclinic orbits are present [20].

7. Families of *N*-bump solutions: *N* even. In this section we determine the global behavior of families of 2-bump, 4-bump, and 6-bump solutions of the problem

$$\begin{cases} u'''' - 2(b^2-1)u'' + (b^2+1)^2 u = 4b(b^2+1)f(u), \\ \lim_{x\to\pm\infty}(u,u',u'',u''') = (0,0,0,0), \end{cases} \tag{7.1}$$

where

$$f(u) = 2e^{-r/(u-th)^2}H(u-th). \tag{7.2}$$

Here $H(\cdot)$ is the Heaviside function, $th > 0$ is the threshold, and $b > 0$, $r > 0$ are constants.

In Figure 24 we again keep $th = 1.5$ and $r = 0.095$, and let b vary, and compute the bifurcation curve for families of even 2-bump and 4-bump solutions of (7.1)–(7.2).

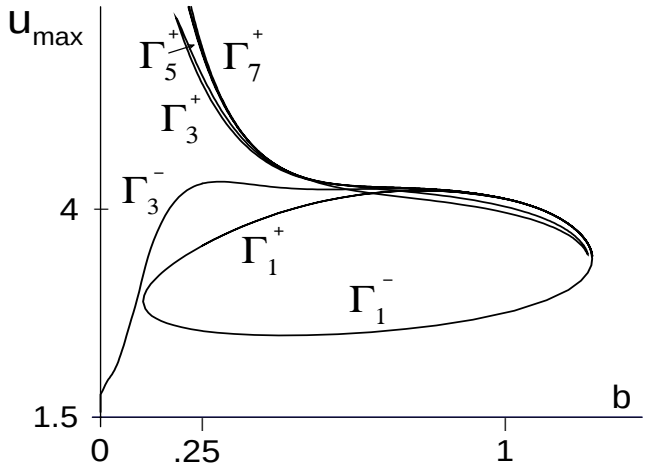


FIG. 21. *Bifurcation curve for* (6.1)–(6.2) *showing* 1-, 3-, 5-, *and* 7-*bump solutions. Parameters are* $th = 1.5$ *and* $r = 0.1$. *Compare this with Figure* 18.

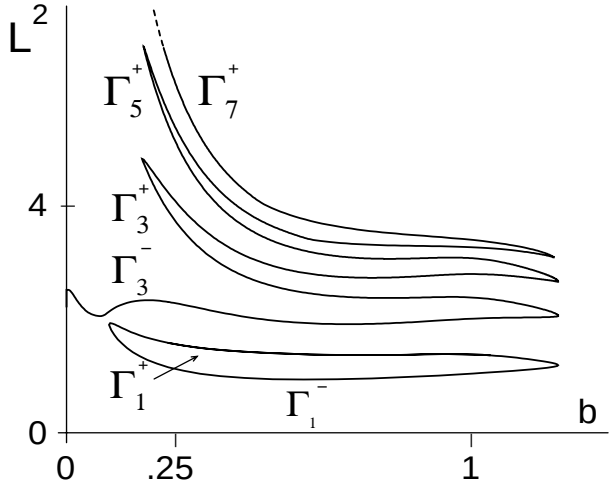


FIG. 22. *The same curves as in Figure* 21*, but the vertical axis is now the* L^2 *norm of the solutions.*

Figures 25–28 show solutions at specific points $P_0, \ldots, P_7$ on this curve. To compute the curve in Figure 24 we first set $b = 0.25$ and integrate (3.1)–(3.3) with an initial condition chosen so that the solution converges, as $t \to \infty$, to the 2-bump (apparently stable) solution indicated by point P_4, and illustrated in the right panel of Figure 27. We then use AUTO97 to continue this solution as b varies. In Figure 24 we find 2-bump solutions, which are conjectured to be unstable, along the lower branch $\Gamma_2{}^-$

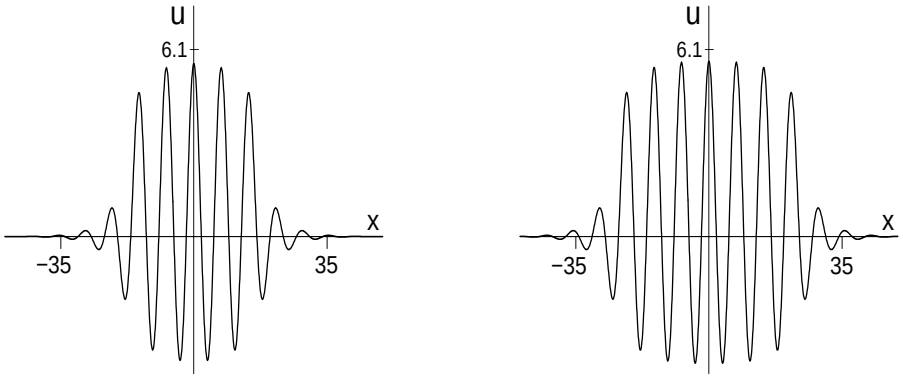


FIG. 23. *Solutions on the curves* Γ_5^+ *(left) and* Γ_7^+ *(right) in Figure* 22. *Parameters are* $r = 0.1$, $th = 1.5$, and $b = 0.25$.

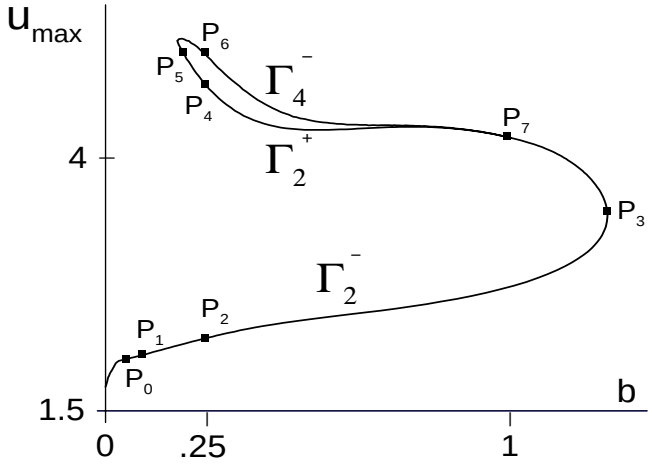


FIG. 24. *Bifurcation curve of* 2*-bump and* 4*-bump solutions for* (7.1)–(7.2). *Solutions at the points* $P_0, \ldots, P_7$ *are shown in Figures* 25–28. *Parameters are* $r = 0.095$, $th = 1.5$. *Compare with Figure* 13.

between P_1 ($b = 0.045$) and P_3 ($b = 1.23$). Solutions at P_2 ($b = 0.25$) and P_3 are shown in Figure 26. As b decreases along $\Gamma_2{}^-$, solutions cease to be 2-bump solutions at P_1 (right panel in Figure 25). To the left of P_1 our computations imply that solutions acquire arbitrarily many bumps as $b \to 0^+$, as was the case for bumps with N odd. For example, at $b = 0.03$ the point P_0 corresponds to the 4-bump solution in the left panel of Figure 25.

Remark. The solution in the left panel of Figure 26 is computed at $b = 0.25$. As $r \to 0^+$, our computations indicate that this solution is unstable and tends to the 2-bump solution shown in the right panel of Figure 7.

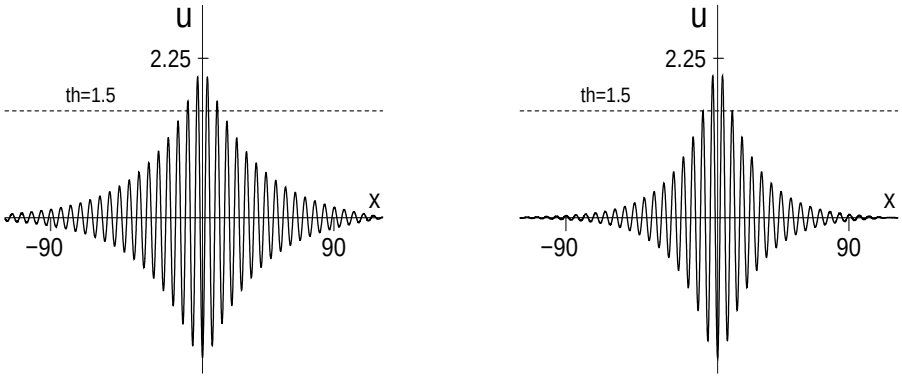


FIG. 25. *Solutions on* $\Gamma_2{}^-$ *at* P_0 *(left) and* P_1 *(right) in Figure* 24.

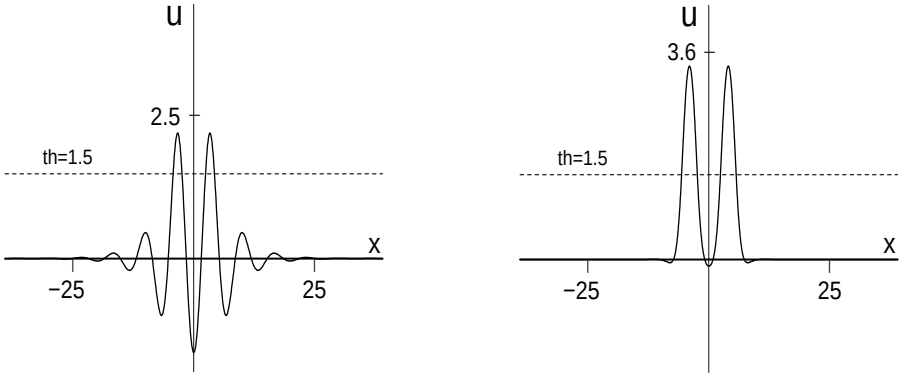


FIG. 26. *Solutions on* $\Gamma_2{}^-$ *at* P_2 *(left) and* P_3 *(right) in Figure* 24.

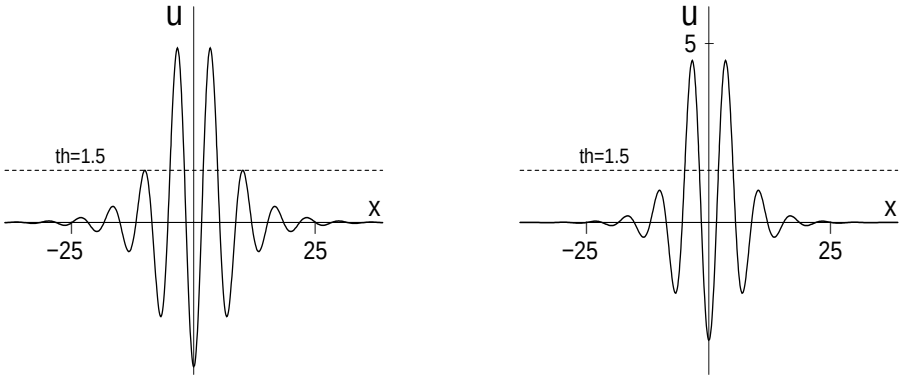


FIG. 27. *Solutions on* $\Gamma_2{}^+$ *at* P_5 *(left) and* P_4 *(right) in Figure* 24.

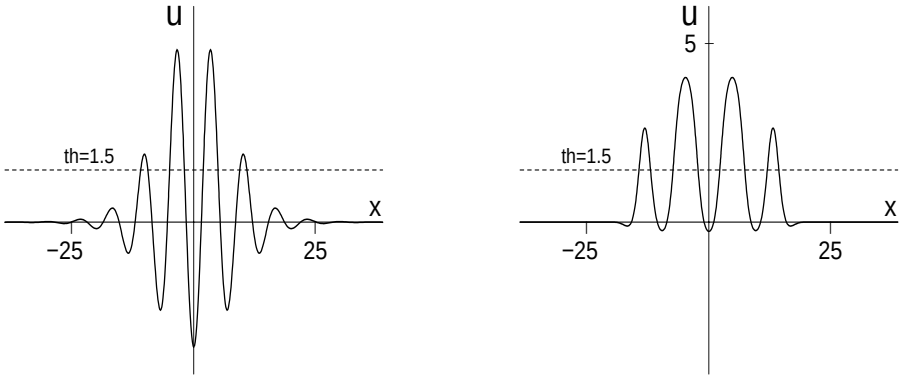


FIG. 28. *Solutions on* $\Gamma_4{}^-$ *at* P_6 *(left) and* P_7 *(right) in Figure* 24.

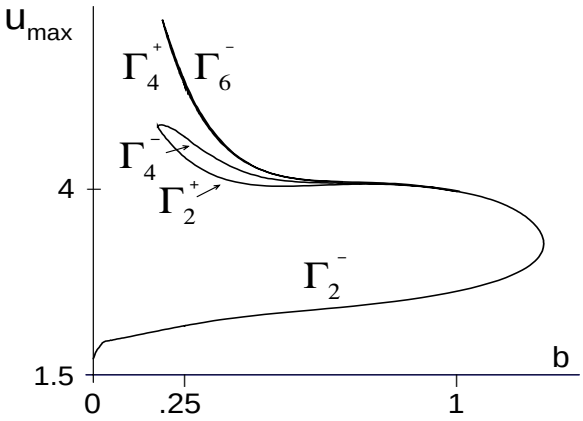


FIG. 29. *Bifurcation curve for* 2, 4, *and* 6*-bump solutions of* (7.1)–(7.2). *This figure is an extension of Figure* 24.

Next, along the middle branch $\Gamma_2{}^+$ in Figure 24 we find a family of 2-bump solutions, some of which are conjectured to be stable, between P_5 ($b = 0.187$) and P_3 ($b = 1.23$). As b decreases along $\Gamma_2{}^+$, solutions cease to be 2-bump solutions at P_5 (shown in the left panel of Figure 27). The solution in the right panel of Figure 27 corresponds to P_4 ($b = 0.25$) in Figure 24. As $r \to 0^+$ this solution tends to the 2-bump solution shown in the left panel of Figure 7.

We let $\Gamma_4{}^-$ denote the upper branch in Figure 24. Along this branch our computations indicate that solutions are *unstable* 4-bump solutions. The solutions at P_6 ($b = 0.25$) and P_7 ($b = 0.99$) are shown in Figure 28. We have also found another family of 4-bump solutions, as well as 6-bump solutions. These solutions lie on a second branch leading to the original curve in Figure 24. The lower and upper curves on this branch are given by $\Gamma_4{}^+$ and $\Gamma_6{}^-$ in Figure 29. In Figure 30 we give specific solutions on $\Gamma_4{}^+$ and $\Gamma_6{}^-$ at $b = 0.25$. Our computations indicate that the solution in the left panel of Figure 30 is stable and tends to the solution in the left panel of Figure 9 as $r \to 0^+$.

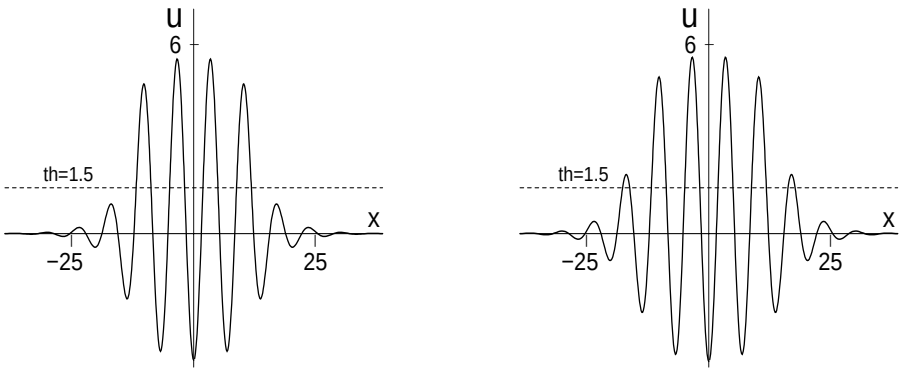


FIG. 30. *Solutions on* $\Gamma_4{}^+$ *(left) and* $\Gamma_6{}^-$ *(right) at* $b = 0.25$ *in Figure* 29.

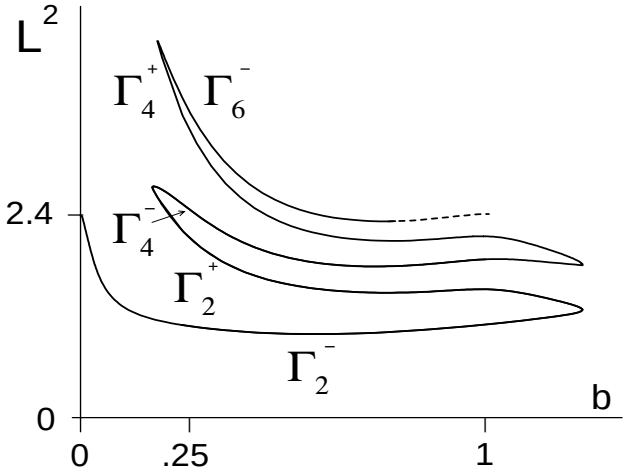


FIG. 31. *The same curves as in Figure* 29, *but the vertical axis is now the* L^2 *norm of the solutions.*

As in the previous section, we redraw in Figure 31 the bifurcation curve shown in Figure 29 but using the L^2 norm for the vertical axis. This allows us to see the separation of branches and, once again, a snaking diagram results.

While we have only looked at multi-bump solutions for which successive maxima of u monotonically increase and then decrease as a function of x, there may also exist "$(n + m)$-bumps" for integer $n, m \geq 1$. These would have the approximate form of an n-bump "glued" to an m-bump, with sufficient low-amplitude oscillations between them. The linearization of (5.8) about the origin has the form necessary for these "composite" orbits to exist, and to confirm this conjecture one would need to check that the N-bump orbits studied above were formed by transverse intersections of the stable and unstable manifolds of the origin (a generic property). See [7] and references

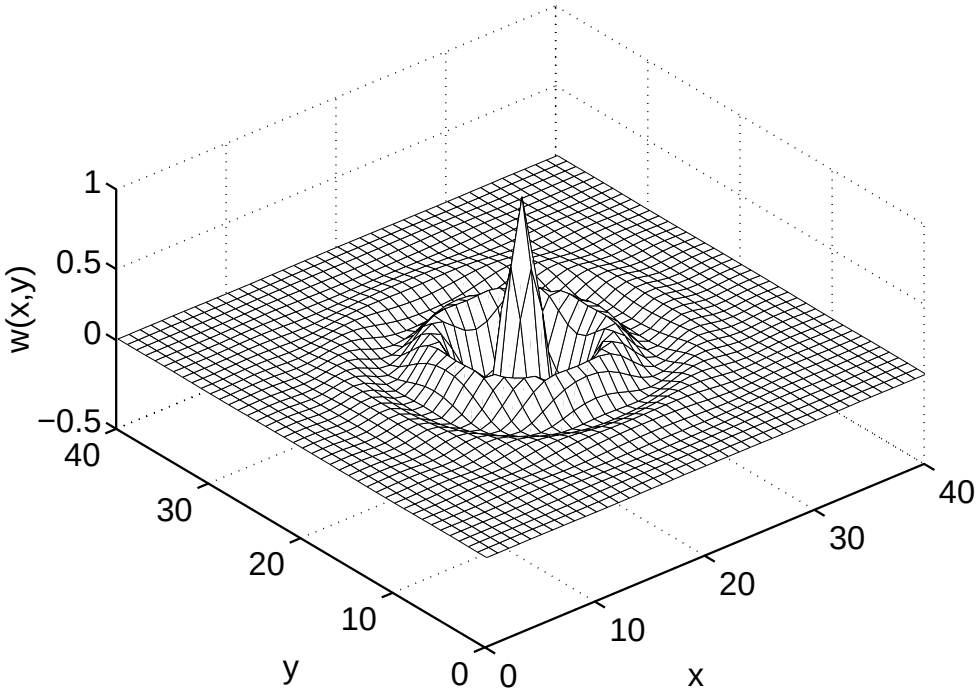


FIG. 32. *Coupling function $w(x, y)$, (8.2), for $b = 0.3$, centered at the center of the domain.*

therein for more details.

8. Extension to two space dimensions. In this section we extend our model to include two spatial dimensions. The system we study, an analogy of (3.1)–(3.3), is the following:

$$\frac{\partial u(x,y,t)}{\partial t} = -u(x,y,t) + \iint_{\Omega} w(x-q, y-s) f(u(q,s,t))\, dq\, ds, \tag{8.1}$$

where

$$w(x,y) = e^{-b\sqrt{x^2+y^2}} \left(b \sin\left(\sqrt{x^2+y^2}\right) + \cos\left(\sqrt{x^2+y^2}\right)\right), \tag{8.2}$$

and

$$f(u) = 2e^{-r/(u-th)^2} H(u-th). \tag{8.3}$$

The coupling function (8.2) is the same as (3.2), with distance in one dimension now replaced by distance in two dimensions. An example is shown in Figure 32. The rate function, (8.3), is identical to (3.3).

A typical stable solution of (8.1)–(8.3) is shown in Figure 33 for the parameters $r = 0.1$, $th = 1.5$, and $b = 0.45$. The initial condition was $u(x, y, 0) = 5$ for $16 < x < 25.6$ and $8 < y < 24$, and $u(x, y, 0) = 0$ otherwise. The domain, Ω, is a square of side-length 40, discretized by a regular 50×50 grid, with open boundaries; i.e., there are no constraints on u or any of its derivatives at the boundaries, and the integral

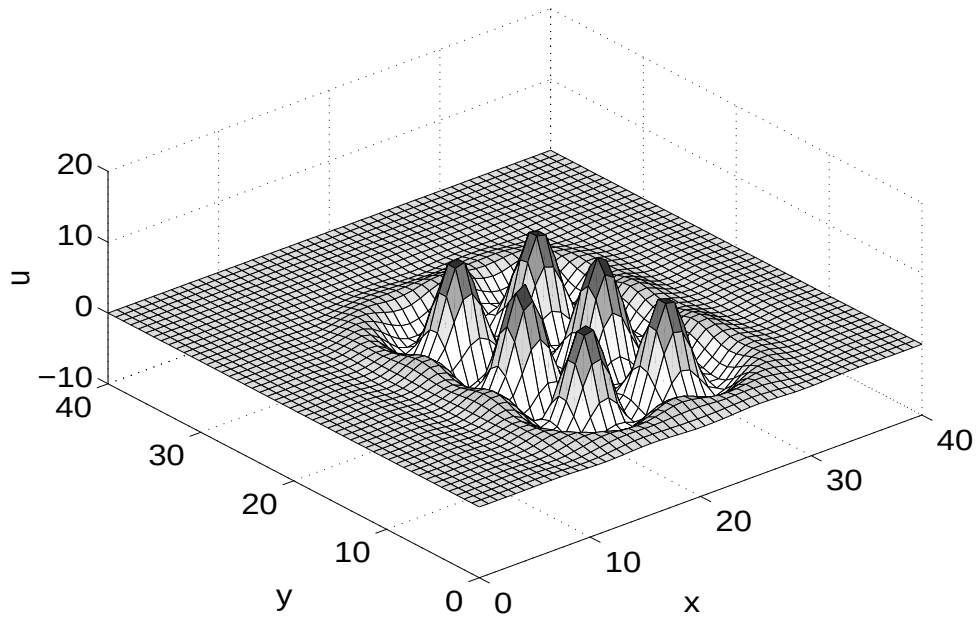


FIG. 33. *A "6-bump" stable solution of* (8.1)–(8.3). *Parameters are* $b = 0.45$, $r = 0.1$, $th = 1.5$.

in (8.1) is taken over only Ω. Note that while the coupling function (8.2) is radially symmetric, the domain is not, and so we do not expect the resulting solutions to have radial symmetry. The equation (8.1) was integrated using an Euler step until the solution converged to a steady state, and at each time step the double integral was approximated by a Riemann integral using the values of u on the grid mentioned above. Note that the convolution can be performed more efficiently by using the two-dimensional fast Fourier transform.

Figure 33 shows the resultant 6-bump solution, and the distance between local maxima is approximately the same as the distance between successive maxima of the coupling function (2π). The regularity is a reflection of the initial condition; more irregular initial conditions lead to an irregular cluster of bumps with similar spacing between local maxima (not shown). That is, keeping $r = 0.1$, $th = 1.5$, and $b = 0.45$, it is possible to find other stable clusters with small numbers of bumps, with the exact number and position being determined by the initial condition. This is analogous with the one-dimensional model (3.1)–(3.3) where stable multi-bump solutions coexist for $b = 0.25$ (see Figure 19 (left), Figure 23, Figure 27 (right), and Figure 30 (left)). In the two-dimensional model, as b is decreased from $b = 0.45$ it seems more difficult to find localized clusters of multi-bump solutions. Instead, for smaller b, either an initial set of u values will die down to $u = 0$ if b is too small or else the entire domain will be filled with bumps. An example with $b = 0.3$ and the other parameters the same (i.e., $r = 0.1$ and $th = 1.5$) is shown in Figure 34. This "progressive recruitment" phenomenon is the same as that seen by Gutkin, Ermentrout, and O'Sullivan in a one-dimensional model [16]. Similar patterns were also found by Usher, Stemmler, and Olami [34] in a neural model with short-range excitation and long-range inhibition.

For larger b, stable attractors also form, but they do not seem to retain the structure of a cluster of bumps observed in Figure 33. However, there still appears

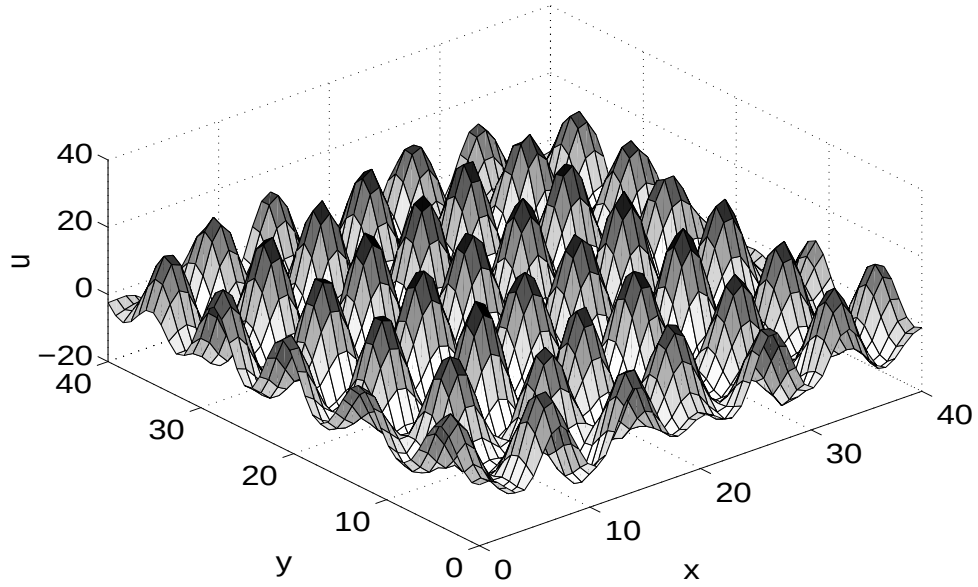


FIG. 34. *A stable solution of* (8.1)–(8.3). *Parameters are* $b = 0.3$, $r = 0.1$, $th = 1.5$. *The initial* u *was spatially localized.*

to be a characteristic length similar to the interbump spacing seen for lower b. In Figure 35, keeping $r = 0.1$ and $th = 1.5$, we increase b to $b = 0.7$ and illustrate an example of this type of stable attractor. For still larger b values, the whole domain becomes active and there are no structures with characteristic length 2π. This is probably due to the lack of a significant inhibitory component to w when b is large—see Figure 4, right panel, for an illustration of this effect in the one-dimensional setting.

In this section, we have presented only numerical results. We leave the possible derivation of a differential equation problem whose solutions describe steady states of (8.1)–(8.3), and any further analysis, as open problems. Although few mathematical results exist for two-dimensional neural models, some interesting results have been obtained relating to the study of circular stationary solutions [2, 31, 36].

9. Proof of Theorem 2.1. In this section we prove Theorem 2.1 concerning the nonexistence of a class of 2-bump solutions of problem (2.2)–(2.3). Recall from section 2 that $u(x)$ is a 2-bump solution of (2.2)–(2.3) if there are values $0 < a < b < c$ such that

$$\begin{cases} u > 0 \quad \text{on} \quad (0,a) \cup (b,c), \\ u(0) = u(a) = u(b) = u(c) = 0, \\ u < 0 \quad \text{otherwise.} \end{cases} \tag{9.1}$$

We define the "distance between bumps" to be $b - a$. Also, we recall from section 2 that under hypotheses (H_1)–(H_6), the function $w(x)$ is symmetric with respect to $x = 0$, that $w(x)$ attains a unique local minimum on $\mathbf{R}^+$ at a value $x_0 > 0$, and that $w(x)$ is *increasing* on (x_0, ∞) (see Figure 1). We will use these properties in our proof of the following result (a restatement of Theorem 2.1).

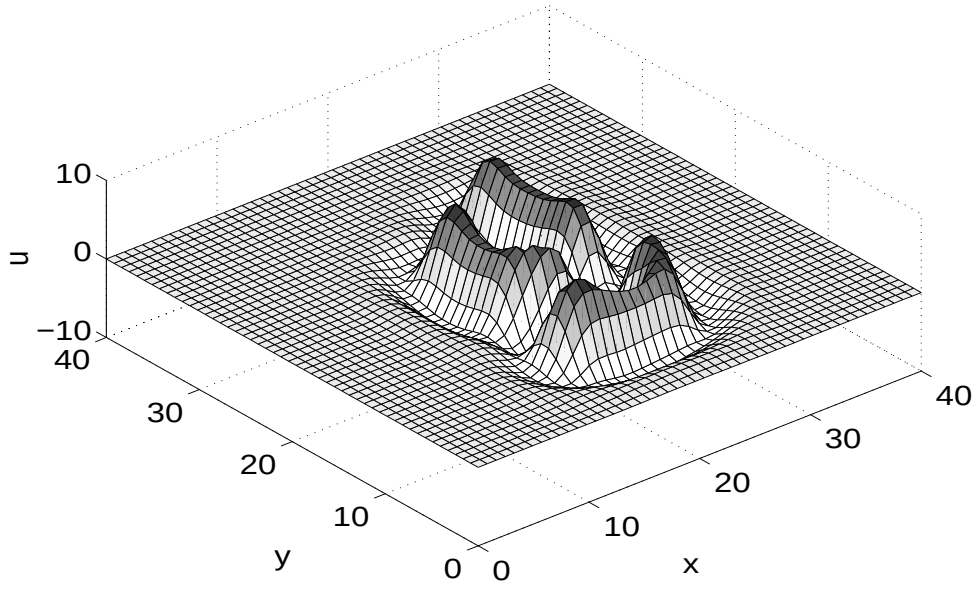


FIG. 35. *A stable solution of* (8.1)–(8.3). *Parameters are* $b = 0.7$, $r = 0.1$, $th = 1.5$. *The initial condition was random but spatially localized.*

THEOREM 9.1. *Under hypotheses* (H_1)–(H_6) *there is no value* $h \in \mathbf{R}$ *for which the problem* (2.2)–(2.3) *has a* 2*-bump solution satisfying* (2.9) *such that the distance between bumps satisfies* $b - a \geq x_0$.

Proof. We assume that there is an $h \in \mathbf{R}$ for which (2.2)–(2.3) has a solution satisfying (2.9), with $b - a \geq x_0$. Using (H_1)–(H_6), we will obtain a contradiction of this assumption. From (2.2), (2.3) and (2.9), it follows that $u(x)$ can be written in the form

$$u(x) = \int_0^a w(x-y)\,dy + \int_b^c w(x-y)\,dy + h \qquad \forall x \in \mathbf{R}. \tag{9.2}$$

Next, recall from (2.4) that $W(x)$ is defined by

$$W(x) = \int_0^x w(y)\,dy \qquad \forall x \in \mathbf{R}. \tag{9.3}$$

Hypotheses (H_1)–(H_6) imply that $W(x)$ is odd. That is,

$$W(x) = -W(-x) \qquad \forall x \in \mathbf{R}. \tag{9.4}$$

Using (9.3), we write (9.2) as

$$u(x) = W(x) - W(x-a) + W(x-b) - W(x-c) + h. \tag{9.5}$$

Because $u(b) = u(c) = W(0) = 0$, it follows from (9.5) that

$$u(c) = W(c) - W(c-a) + W(c-b) + h = 0, \tag{9.6}$$

and

$$u(b) = W(b) - W(b-a) - W(b-c) + h = 0. \tag{9.7}$$

We note that $W(c-b) = -W(b-c)$ since $W(x)$ is odd. Thus, a subtraction of (9.7) from (9.6) leads to

$$W(c) - W(b) = W(c-a) - W(b-a). \tag{9.8}$$

Recalling the definition of $W(x)$ from (9.3), we write (9.8) as

$$\int_b^c w(y)dy = \int_{b-a}^{c-a} w(y)dy. \tag{9.9}$$

Also, our hypothesis that $b - a \geq x_0$ implies that

$$x_0 \leq b - a < c - a. \tag{9.10}$$

We need to consider two cases to complete the proof. The first case is

$$x_0 \leq b - a < c - a \leq b < c. \tag{9.11}$$

From (H_6) and (9.10) we conclude that $w(x)$ is increasing on $(b-a, c)$. Thus, $w(x) > w(b)$ on (b, c), and $w(x) < w(c-a)$ on $(b-a, c-a)$. This implies that

$$\int_b^c w(y)dy > w(b)(c-b), \tag{9.12}$$

and

$$\int_{b-a}^{c-a} w(y)dy < w(c-a)(c-b). \tag{9.13}$$

Combining (9.9), (9.11), (9.12), and (9.13), we conclude that

$$w(b) < w(c-a). \tag{9.14}$$

However, since (H_6) implies that $w(x)$ is nondecreasing on $[c-a, b]$, it follows that $w(b) \geq w(c-a)$, contradicting (9.14). The second case we need to consider is

$$x_0 \leq b - a < b < c - a < c. \tag{9.15}$$

Then (9.9) can be written as

$$\int_b^{c-a} w(y)dy + \int_{c-a}^{c} w(y)dy = \int_{b-a}^{b} w(y)dy + \int_b^{c-a} w(y)dy.$$

This reduces to

$$\int_{c-a}^{c} w(y)dy = \int_{b-a}^{b} w(y)dy. \tag{9.16}$$

Again, we use the fact that $w(x)$ is increasing on (x_0, c), together with (9.15), and conclude that

$$\int_{c-a}^{c} w(y)dy > w(c-a)a \tag{9.17}$$

and

$$\int_{b-a}^{b} w(y)dy < w(b)a. \tag{9.18}$$

From (9.16)–(9.18) it follows that $w(b) > w(c-a)$. However, this is a contradiction since $w(x)$ increases on $(b, c-a)$. The proof of Theorem 2.1 is now complete.

10. Proof of Theorem 5.1. In this section we prove Theorem 5.1 and determine a global parameter regime over which nonconstant solutions of the problem

$$\begin{cases} u'''' - 2(b^2-1)u'' + (b^2+1)^2 u = 4b(b^2+1)f(u), \\ \lim_{x\to\pm\infty}(u,u',u'',u''') = (0,0,0,0) \end{cases} \tag{10.1}$$

might possibly exist. We recall that $f(u)$ is defined by

$$f(u) = 2e^{-r/(u-th)^2}H(u-th), \tag{10.2}$$

where $H(u-th)$ is the Heaviside function (see Figure 5). For convenience we restate our result (Theorem 5.1) below.

THEOREM 10.1. *Let $r > 0$ and $th > 0$. If there is a value $b > 0$ for which (10.1)–(10.2) has a nonconstant solution, then*

$$0 < b \le \frac{4+\sqrt{|16-th^2|}}{th}.$$

Proof. Suppose that $u(x)$ is a nonconstant solution of (10.1)–(10.2) for some

$$r > 0,\ th > 0, \quad \text{and} \quad b > \frac{4+\sqrt{|16-th^2|}}{th}. \tag{10.3}$$

We will obtain a contradiction of this assumption. First, we observe that

$$\frac{4+\sqrt{|16-th^2|}}{th} \ge 1 \qquad \forall\ th > 0. \tag{10.4}$$

It then follows from (10.3) and (10.4) that $b > 1$. Next, from (10.1)–(10.2) it is easily verified that $u(x)$ must satisfy the first integral

$$u'u''' - \frac{(u'')^2}{2} - (b^2-1)(u')^2 + (b^2+1)^2 Q(u) = 0, \tag{10.5}$$

where $Q(u)$ is defined by

$$Q(u) \equiv \int_0^u \left(s - \left(\frac{8b}{b^2+1}\right) e^{-r/(s-th)^2} H(s-th) \right) ds. \tag{10.6}$$

Over the range given in (10.3), we claim that the integrand in (10.6) satisfies

$$u - \left(\frac{8b}{b^2+1}\right) e^{-r/(u-th)^2} H(u-th) > 0 \qquad \forall u > 0. \tag{10.7}$$

First, suppose that $0 < u \le th$. Then $f(u) = 0$ by (10.2), and therefore the left side of (10.7) must be positive. If $u > th$, then

$$u - \left(\frac{8b}{b^2+1}\right) e^{-r/(u-th)^2} H(u-th) > th - \frac{8b}{b^2+1} > 0,$$

since we assume that $th > 0$, $r > 0$, and $b > (4+\sqrt{|16-th^2|})/th$. Thus (10.7) is proved. From (10.6) and (10.7) we conclude that $Q(0) = 0$,

$$Q(u) > 0 \quad \text{if} \quad |u| > 0, \qquad \lim_{|u|\to\infty} Q(u) = \infty, \tag{10.8}$$

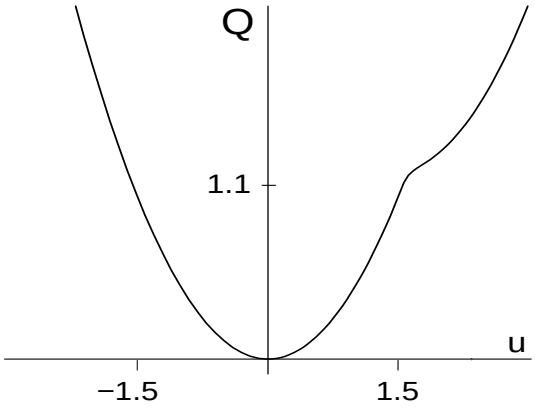


FIG. 36. $Q(u)$, (10.6), *for parameter values* $r = 0.005,\ th = 1.5,\ b = 5.2$.

and

$$\frac{dQ}{du} < 0 \qquad \forall u < 0, \qquad \frac{dQ}{du} > 0 \qquad \forall u > 0. \tag{10.9}$$

For example, the parameters $r = 0.005$, $th = 1.5$, and $b = 5.2$ satisfy (10.3), and in Figure 36 we graph the corresponding $Q(u)$.

Next, because (10.1)–(10.2) is autonomous, we may assume that the solution $u(x)$ attains its global maximum at $x = 0$. We claim that $u(0) > th$. If, on the contrary, $u(0) \le th$, then $u(x) \le th$ for all $x \in \mathbf{R}$, and it follows from (10.2) that $f(u) = 0$ for all $x \in \mathbf{R}$. This reduces the integral equation (5.1) to $u(x) = 0$, and we arrive at a contradiction since we assume that $u(x)$ is a *nonconstant* solution of (10.1)–(10.2), and solutions of (10.1)–(10.2) also are solutions of (5.1). Thus, at $x = 0$ it must be the case that

$$u(0) > th,\ u'(0) = 0, \quad \text{and} \quad u''(0) \le 0. \tag{10.10}$$

Substituting (10.10) into (10.5), and using (10.8), we conclude that

$$u''(0) = -(b^2 + 1)\sqrt{2Q(u(0))} < 0. \tag{10.11}$$

Without loss of generality we may assume that $u'''(0) \le 0$. Otherwise, if $u'''(0) > 0$, then it would suffice to consider the function $v(x) = u(-x)$ which also is a solution of (10.1)–(10.2) and satisfies the initial conditions

$$v(0) > th,\ v'(0) = 0,\ v''(0) < 0, \quad \text{and} \quad v'''(0) < 0.$$

Thus, it may be assumed that the solution $u(x)$ satisfies

$$u(0) > th,\ u'(0) = 0,\ u''(0) < 0, \quad \text{and} \quad u'''(0) \le 0. \tag{10.12}$$

Our goal in the remainder of the proof is to show that there is an $\bar{x} > 0$ such that $u(\bar{x}) > u(0)$. This will contradict the fact that $u(x)$ attains its global maximum

at $x = 0$. Thus, we need to follow the solution as x increases from $x = 0$. Throughout we will make extensive use of the first integral (10.5) and the associated functional $Q(u(x))$. In Figures 37 and 38 we follow $u(x)$ and $Q(u(x))$, respectively, and keep track of the points where the solution $u(x)$ attains its maxima and minima.

From (10.1)–(10.4), (10.7), and (10.12) it follows that $u''''(0) < 0$. This and (10.12) imply that $u'''(x) < 0$ on an interval $(0, \epsilon)$. We set

$$\sigma = \sup\{\hat{x} > 0 | u'''(x) < 0 \quad \forall x \in (0, \hat{x})\}. \tag{10.13}$$

If $\sigma = \infty$, then $u''(x) < u''(0) < 0$ for all $x > 0$, hence $u''(\infty) < 0$, contradicting the condition $u''(\infty) = 0$ given in (10.1). Thus, it must be the case that $\sigma < \infty$, $u'''(\sigma) = 0$, and

$$u(x) < u(0),\ u'(x) < 0, \quad \text{and} \quad u''(x) < u''(0) < 0 \qquad \forall x \in (0, \sigma]. \tag{10.14}$$

Next, it follows from (10.8) and (10.9) that there is a unique, negative value $u_1 < 0$ (see Figure 38) such that

$$Q(u) < Q(u(0)) \qquad \forall u \in (u_1, u(0)), \quad \text{and} \quad Q(u_1) = Q(u(0)). \tag{10.15}$$

We need to show that $u(\sigma) < u_1$. If $u(\sigma) \geq u_1$, then from (10.11), (10.14), and (10.15) it follows that $(u'')^2$ *increases* on $(0, \sigma)$ so that

$$\frac{(u(x)'')^2}{2} > (b^2 + 1)^2 Q(u(x)) \qquad \forall x \in (0, \sigma]. \tag{10.16}$$

Setting $x = \sigma$ in (10.5), and using (10.3), (10.4), (10.14), and (10.16), we obtain

$$-(u'(\sigma))^2(b^2 - 1) > 0,$$

a contradiction since $u'(\sigma) < 0$ and $b > 1$. Therefore it must be the case that $u(\sigma) < u_1$. Thus, there is an $x_1 \in (0, \sigma)$ such that (see Figure 37)

$$u'(x) < 0,\ u''(x) < 0,\ u'''(x) < 0 \qquad \forall x \in (0, x_1], \quad \text{and} \quad u(x_1) = u_1. \tag{10.17}$$

Since $u(\infty) = 0$, it follows from (10.17) that there is an $x_2 > x_1$ such that

$$u'(x) < 0 \qquad \forall x \in [x_1, x_2), \quad \text{and} \quad u'(x_2) = 0. \tag{10.18}$$

We conclude from (10.5) and (10.18) that

$$u(x_2) < u_1 < 0,\ u'(x_2) = 0, \text{ and } u''(x_2) = (b^2 + 1)\sqrt{2Q(u(x_2))} > 0. \tag{10.19}$$

We need to determine the sign of $u'''(x_2)$. Because $u''(x_1) < 0$ and $u''(x_2) > 0$, there is an $\tilde{x} \in (x_1, x_2)$ where $u''(\tilde{x}) = 0$ and $u'''(\tilde{x}) \geq 0$. This, (10.3), (10.4), and (10.18) give

$$u'''(\tilde{x}) - 2(b^2 - 1)u'(\tilde{x}) > 0. \tag{10.20}$$

Next, because $u(x) < u_1 < 0$ on $[\tilde{x}, x_2]$, it follows from (10.1)–(10.2) that

$$(u''' - 2(b^2 - 1)u')' = -(b^2 + 1)^2 u > 0 \qquad \forall x \in [\tilde{x}, x_2]. \tag{10.21}$$

From (10.19), (10.20), and (10.21) we conclude that

$$u'''(x_2) > 0. \tag{10.22}$$

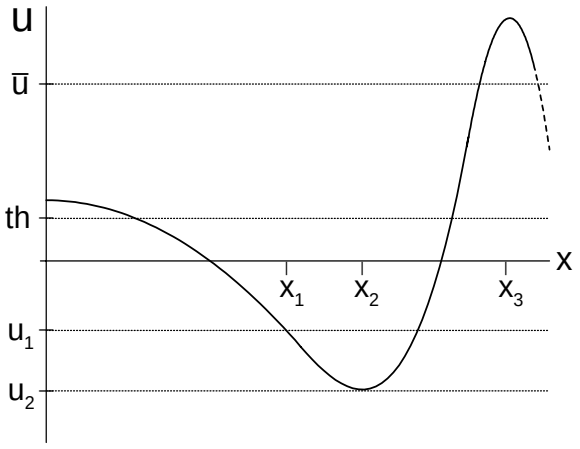


FIG. 37. *A sketch of* $u(x)$ *for* (10.1)–(10.2): $u(x_1) = u_1,\ u(x_2) = u_2$, *and* $u(x_3) > \bar{u}$.

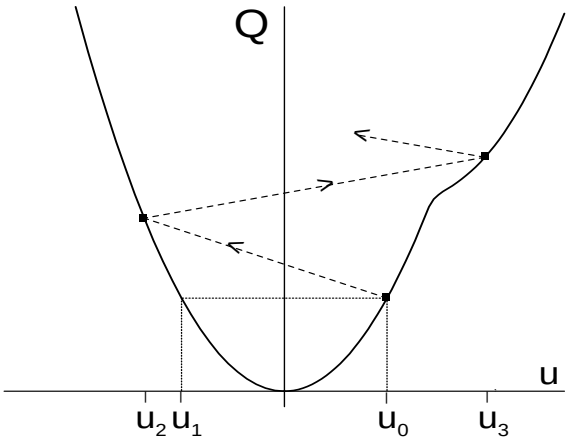


FIG. 38. $Q(u)$, (10.6): $u_0 = u(0),\ u_1 = u(x_1) = u_0,\ u_2 = u(x_2),\ u_3 = u(x_3)$.

In Figure 38 we set $u_1 = u(x_1)$ and $u_2 = u(x_2)$. As $u(x)$ decreases from u_1 to u_2, properties (10.8) and (10.9) imply that $Q(u)$ *increases*, and therefore

$$Q(u(x_2)) > Q(u_1) = Q(u(0)). \tag{10.23}$$

In the final step of the proof we follow $u(x)$ as x increases from $x = x_2$, and we show that there is an $x_3 > x_2$ such that $u(x_3) = u_3 > u(0)$ (see Figures 37 and 38). We first observe from (10.8)–(10.9) that there is a unique $\bar{u} > 0$ such that $Q(\bar{u}) = Q(u(x_2))$.

It follows from (10.23), and the fact that $Q(u)$ is increasing for $u > 0$, that

$$\bar{u} > u(0). \tag{10.24}$$

Next, define

$$x_3 = \sup\{\hat{x} > x_2 | u'''(x) > 0 \ \forall x \in (x_2, \hat{x})\}. \tag{10.25}$$

Because of (10.24), if we show that $u(x_3) > \bar{u}$, we will obtain a contradiction of the fact that $u(x)$ has its global maximum at $x = 0$. From (10.19), (10.22), and (10.25) it follows that

$$u'(x) > 0, \quad u''(x) > u''(x_2) = (b^2+1)\sqrt{2Q(u(x_2))} > 0 \qquad \forall x \in (x_2, x_3]. \tag{10.26}$$

If $x_3 = \infty$, then (10.26) implies that $u''(\infty) > 0$, contradicting the condition $u''(\infty) = 0$ given in (10.1). Thus, $x_3 < \infty$ and it follows from (10.25) that

$$u'''(x_3) = 0. \tag{10.27}$$

Finally, suppose that

$$u(x_2) < u(x) \leq \bar{u} \qquad \forall x \in (x_2, x_3).$$

Then (10.8) and (10.9) imply that

$$0 \leq Q(u(x)) \leq Q(u(x_2)) \qquad \forall x \in (x_2, x_3). \tag{10.28}$$

Combining (10.26), (10.27), and (10.28), and setting $x = x_3$ in (10.5), we obtain

$$-(b^2-1)(u'(x_3))^2 = \frac{(u''(x_3))^2}{2} - (b^2+1)^2 Q(u(x_3)) > 0,$$

a contradiction since $u'(x_3) > 0$ and $b > 1$. Thus, it must be the case that $u(x_3) > \bar{u} > u(0)$ as claimed. However, as described earlier, this contradicts the fact that $u(x)$ has its global maximum at $x = 0$. This completes the proof.

11. Summary. In this paper we have studied steady states of a partial integro-differential equation that has been used to model working memory in a neuronal network. We have extended previous results for "Mexican hat" coupling to the case where the connectivity function changes sign infinitely often, in the hope of more realistically modeling the connectivity known to exist in the prefrontal cortex. Our main results include (a) a proof of the nonexistence of a type of "multiple bump" solution when the connectivity is of Mexican hat type, (b) an upper bound on the decay rate of an oscillatory connectivity function, above which only trivial solutions exist, and (c) a numerical investigation of the possible solutions and the bifurcations they undergo for a particular oscillatory connectivity function.

For the one-dimensional model, many of the numerical results were obtained as a result of noting that stationary solutions of the partial integro-differential equation (5.1) are equivalent to homoclinic orbits in the related fourth order ordinary differential equation problem (5.8). This property allowed us to use the software package AUTO97 [12, 13], with its facilities for continuing homoclinic orbits, to follow both stable and unstable solutions as parameters were varied. We are presently pursuing a rigorous proof of existence of the families of N-bump solutions found here. Already,

it has been proved in [23] that any bounded solution of the ordinary differential equation in (5.8) also is a solution of the integral equation (5.1). Thus, in addition to homoclinic orbits, we are also investigating the existence of other families of solutions, including periodic, aperiodic, and chaotic solutions. While many of our results were derived by exploiting the specific form of an oscillatory connectivity function, we believe that the qualitative aspects of our results will hold for any qualitatively similar function.

For the two-dimensional extension of our model we used a MATLAB [28] code to generate stable multi-bump solutions. For appropriate parameter values we found that N-bump solutions exist and that they retain many of the characteristic qualities of solutions of the one-dimensional model. However, we also found stable solutions which were not predicted by our one-dimensional studies. In future research we will continue our investigation of the different types of stable patterns of solutions of the two-dimensional problem.

Acknowledgment. The authors thank Edward Krisner and the referees for making several helpful suggestions.

REFERENCES

[1] S. Amari, *Dynamics of pattern formation in lateral-inhibition type neural fields,* Biol. Cybern., 27 (1977), pp. 77–87.
[2] S. Amari, *Mathematical Theory of Neural Networks,* Sangyo-Tosho, Tokyo, 1978.
[3] D. J. Amit and N. Brunel, *A model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex*, Cereb. Cortex., 7 (1997), pp. 237–252.
[4] P. C. Bressloff, *Travelling fronts and wave propagation failure in an inhomogeneous neural network,* Phys. D, 155 (2001), pp. 83–100.
[5] A. R. Champneys, *Homoclinic orbits in reversible systems and their applications in mechanics, fluids and optics,* Phys. D, 112 (1998), pp. 158–186.
[6] A. R. Champneys, *Homoclinic orbits in reversible systems* II*: Multibumps and saddle centers,* CWI Quarterly, 12 (1999), pp. 185–212.
[7] A. R. Champneys and P. J. McKenna, *On solitary waves of a piecewise linear suspended beam model,* Nonlinearity, 10 (1997), pp. 1763–1782.
[8] C. L. Colby, J. R. Duhamel, and M. E. Goldberg, *Oculocentric spatial representation in parietal cortex,* Cereb. Cortex., 5 (1995), pp. 470–481.
[9] A. Compte, N. Brunel, P. Goldman-Rakic, and X.-J. Wang, *Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model,* Cereb. Cortex., 10 (2000), pp. 910–923.
[10] S. Coombes, G. J. Lord, and M. R. Owen, *Waves and Bumps in Neuronal Networks with Axo-Dendritic Synaptic Interactions,* preprint, Loughborough University, Leicestershire, UK, 2002.
[11] M. C. Cross and P. C. Hohenberg, *Pattern formation outside of equilibrium,* Rev. Modern Phys., 65 (1993), pp. 851–1112.
[12] E. J. Doedel. *Auto: A program for the automatic bifurcation analysis of autonomous systems,* in Proceedings of the Tenth Manitoba Conference on Numerical Mathematics and Computing, University of Manitoba, Winnipeg, Canada, 1981, pp. 265–284.
[13] E. Doedel, A. R. Champneys, T. F. Fairgrieve, Y. A. Kuznetsov, B. Sandstede, and X. Wang, *AUTO*97*: Continuation and Bifurcation Software for Ordinary Differential Equations (with HomCont)*, report, CMVL, Concordia University, Montreal, 1997.
[14] G. B. Ermentrout, *Neural networks as spatio-temporal pattern forming systems,* Rep. Progr. Phys., 61 (1998), pp. 353–430.
[15] S. Funahashi, C. J. Bruce, and P. S. Goldman-Rakic, *Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex,* J. Neurophysiol., 61 (1989), pp. 331–349.
[16] B. Gutkin, G. B. Ermentrout, and J. O'Sullivan, *Layer* 3 *patchy recurrent connections may determine the spatial organization of sustained activity in the primate frontal cortex,* Neurocomputing, 32–33 (2000), pp. 391–400.
[17] M. A. Giese, *Dynamic Neural Field Theory for Motion Perception,* Kluwer Academic, Boston, 1998.

[18] Y. Guo and C. Chow, *Localized Persistent States in Neural Networks,* preprint, University of Pittsburgh, Pittsburgh, PA, 2002.
[19] D. Hansel and H. Sompolinsky, *Modeling feature selectivity in local cortical circuits,* in Methods in Neuronal Modeling, 2nd ed., C. Koch and I. Segev, eds., MIT Press, Cambridge, MA, 1998.
[20] P. Hirschberg and E. Knobloch, *Šil'nikov–Hopf bifurcation,* Phys. D, 62 (1993), pp. 202–216.
[21] G. W. Hunt, M. A. Peletier, A. R. Champneys, P. D. Woods, M. Ahmer Wadee, C. J. Budd, and G. L. Lord, *Cellular buckling in long structures,* Nonlinear Dynam., 21 (2000), pp. 3–29.
[22] K. Kishimoto and S. Amari, *Existence and stability of local excitations in homogeneous neural fields,* J. Math. Biol., 7 (1979), pp. 303–318.
[23] E. Krisner, W. Troy, and C. R. Laing, *N–Bump Solutions of a Model of Short Term Memory,* manuscript, 2002.
[24] C. R. Laing and C. C. Chow, *Stationary bumps in networks of spiking neurons,* Neural Comp., 13 (2001), pp. 1473–1494.
[25] C. Laing and P. Glendinning, *Bifocal homoclinic bifurcations,* Phys. D, 102 (1997), pp. 1–14.
[26] J. B. Levitt, D. A. Lewis, T. Yoshioka, and J. S. Lund, *Topography of pyramidal neuron intrinsic connections in macaque monkey prefrontal cortex (areas* 9 *and* 46*),* J. Comp. Neurol., 338 (1993), pp. 360–376.
[27] D. A. Lewis and S. A. Anderson, *The functional architecture of the prefrontal cortex and schizophrenia,* Psychol. Med., 25 (1995), pp. 887–894.
[28] Matlab, The MathWorks, Natick, MA.
[29] E. K. Miller, C. A. Erickson, and R. Desimone, *Neural mechanisms of visual working memory in prefrontal cortex of the Macaque,* J. Neurosci., 16 (1996), pp. 5154–5167.
[30] L. A. Peletier and W. C. Troy, *Patterns: Higher order models in physics and chemistry,* Birkhäuser, Boston, 2001.
[31] J. G. Taylor, *Neural "bubble" dynamics in two dimensions: Foundations,* Biol. Cybern., 80 (1999), pp. 393–409.
[32] E. Thelen, G. Schöner, C. Scheier, and L. B. Smith, *The dynamics of embodiment: a field theory of infant perseverative reaching,* Behavioral and Brain Sciences, 24 (2001), pp. 1–34.
[33] W. C. Troy and C. R. Laing, *Two–bump solutions of Amari's model of working memory*, Phys. D, submitted.
[34] M. Usher, M. Stemmler, and Z. Olami, *Dynamic pattern formation leads to 1/f noise in neural populations,* Phys. Rev. Lett., 74 (1995), pp. 326–329.
[35] X. J. Wang, *Synaptic basis of cortical persistent activity: The importance of NMDA receptors to working memory,* J. Neurosci., 19 (1999), pp. 9587–9603.
[36] H. Werner and T. Richter, *Circular stationary solutions in two–dimensional neural fields,* Biol. Cybern., 85 (2001), pp. 211–217.
[37] H. R. Wilson and J. D. Cowan, *A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue,* Kybernetik, 13 (1973), pp. 55–80.
[38] P. D. Woods and A. R. Champneys, *Heteroclinic tangles and homoclinic snaking in the unfolding of a degenerate reversible Hamiltonian–Hopf bifurcation,* Phys. D, 129 (1999), pp. 147–170.
[39] K. Zhang, *Representation of spatial orientation by the intrinsic dynamics of the head–direction cell ensemble: A theory,* J. Neurosci., 16 (1996), pp. 2112–2126.

SIAM J. APPL. MATH.
Vol. 63, No. 1, pp. 98–115

DESORPTION OVERSHOOT IN POLYMER-PENETRANT SYSTEMS: ASYMPTOTIC AND COMPUTATIONAL RESULTS*

DAVID A. EDWARDS† AND RICHARD A. CAIRNCROSS‡

Abstract. Many practically relevant polymers undergoing desorption change from the rubbery (saturated) to the glassy (nearly dry) state. The dynamics of such systems cannot be described by the simple Fickian diffusion equation due to viscoelastic effects. The mathematical model solved numerically is a set of two coupled PDEs for concentration and stress. Asymptotic solutions are presented for a moving boundary-value problem for the two states in the short-time limit. The solutions exhibit *desorption overshoot*, where the penetrant concentration in the interior is less than that on the surface. In addition, it is shown that if the underlying time scale of the equations is ignored when postulating boundary conditions, nonphysical solutions can result.

Key words. asymptotic expansions, desorption, moving boundary-value problems, perturbation methods, polymer-penetrant systems, finite-element method

AMS subject classifications. 35B20, 35C15, 35C20, 35K60, 35R35, 65N30, 74D10, 76M10, 76R99, 80A22

PII. S0036139901390428

1. Introduction.

Over the past few decades, much experimental and theoretical work has been devoted to the study of polymer-penetrant systems. In particular, the desorption of penetrants from saturated polymer matrices has been examined due to its wide industrial applicability. One unusual feature of such systems is the change in the polymer from a *rubbery* state when it is nearly saturated to a *glassy* state when it is nearly dry. As part of the drying process, a glassy *skin* often develops at the exposed surface of a polymer whose properties are significantly different from the rest of the polymer-penetrant solution [1], [2], [3], [4], [5]. This phenomenon, called *skinning* [6], [7], [8], has many industrial applications [8], [9], [10], [11], [12], [13], [14], [15], [16].

There are many different theories for why the skinning process occurs, including phase separation [17], crystallization [18], and diffusion-induced convection [19]. Nevertheless, for the systems we wish to study, most scientists agree that one important factor is a viscoelastic stress in the polymer entanglement network, which can be as important to the transport process as the well-understood Fickian dynamics [20], [21], [22]. The size of this stress is related to the *relaxation time* of the viscoelastic polymer matrix. In the glassy skin, the relaxation time is finite, so the stress is an important effect, but in the rubbery region the relaxation time is nearly zero [15], [20], [23]. Nevertheless, we will show that in order for the mathematical model to yield physically meaningful results, at some level the short relaxation time in the rubber must also be taken into account.

Numerical and analytical solutions are derived here for model equations for the

*Received by the editors June 6, 2001; accepted for publication (in revised form) January 21, 2002; published electronically August 5, 2002.

http://www.siam.org/journals/siap/63-1/39042.html

†Department of Mathematical Sciences, University of Delaware, Newark, DE 19716-2553 (edwards@math.udel.edu). The work of this author was supported by the National Science Foundation grant DMS-9972013.

‡Department of Chemical Engineering, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104 (cairncro@coe.drexel.edu). The work of this author was supported by a 3M Non-Tenured Faculty Grant.

system described above. Our equations are the same, to leading order, as those for general polymer-penetrant systems derived in detail by Edwards and Cohen [24], [25], Edwards [26], Cairncross and Durning [8], Durning [27], and Durning and Tabor [28]. These models, which are presented in section 2, consist of a set of coupled PDEs for the concentration and stress. The parameters in the numerical simulation vary smoothly with concentration, so the glass-rubber interface $x = s(t)$ between the two states is simply an isocline of concentration. In contrast, the parameters in the analytical model are assumed to be piecewise constant in the rubber and glass. Thus, a moving boundary-value problem similar to a Stefan problem results. In each of the regions a different partial differential operator holds, and continuity conditions at the glass-rubber interface dictate its motion.

In section 3 we construct a perturbation solution to the equations. The solutions are expressed as integrals of Green's functions convolved with fictitious boundary conditions which provide the new unknowns for which we must solve. In section 4 we construct short-time asymptotic solutions of the concentration and stress fields. The form of the short-time solutions necessitates a corner layer, where the full system of equations holds. The solutions exhibit *desorption overshoot*, where the minimum in the concentration occurs in the interior of the domain.

In addition, if we use a standard high-mass-transfer-coefficient approximation common in diffusion and heat conduction problems, it is possible for the concentration to become negative. This result is confirmed numerically in section 5. In section 6 it is explained that the unphysical negative concentration appears because the limit of high mass transfer coefficient imposes a jump in the exterior concentration faster than the underlying time scales of the operator. Physically, the polymer is *self-regulating* for desorption as well as sorption [24]. A new boundary condition is postulated which incorporates the time scale in the stress evolution equation, and it is shown that such a boundary condition does not lead to negative concentrations.

2. Preliminaries.

2.1. Governing equations.

We examine the following dimensionless system of equations for anomalous desorption in a polymer of finite dimensionless length L:

$$\frac{\partial C}{\partial t} = \frac{\partial}{\partial x}\left(D(C)\frac{\partial C}{\partial x} + \frac{\partial \sigma}{\partial x}\right), \qquad 0 \le x \le L, \tag{2.1a}$$

$$\frac{\partial \sigma}{\partial t} + \frac{\beta(C)}{\beta_{\mathrm{g}}}\sigma = \gamma\epsilon C + \frac{\partial C}{\partial t}, \tag{2.1b}$$

where C is the dimensionless concentration of penetrant in the polymer, γ is a dimensionless constant, and L is the length of the slab scaled with the length scale of stress evolution [29].

The system is described in general in [29] and specialized in [26], but some discussion is required. The flux in (2.1a) can be derived by postulating that the chemical potential is a function of both C and σ [24], which in one dimension corresponds to the stress in the polymer network [24], [30], [31], [32]. In (2.1b), the coefficient of $\partial\sigma/\partial x$ has been chosen constant, in contrast to the models of Durning and colleagues [8], [28], [33].

$D(C)$ is a normalized diffusion coefficient measuring the ratio of the Fickian to non-Fickian effects in the flux. Also $\beta(C)$ is the inverse of the relaxation time, which measures the speed at which changes in one part of the polymer are communicated

to other parts of the polymer. Both increase dramatically as the polymer goes from the glassy to rubbery state [15], [20], [23], [34], [35], [36]. In contrast, the differences in these parameters *within* states are qualitatively negligible. Therefore, for our numerical work we assume the following forms for these functions:

$$\begin{aligned} D(C) &= D_{\mathrm{g}} - \frac{D_{\mathrm{g}} - D_{\mathrm{r}}}{2}\left[1 + \tanh(\alpha(C - C_*))\right], \\ \beta(C) &= \beta_{\mathrm{g}} - \frac{\beta_{\mathrm{g}} - \beta_{\mathrm{r}}}{2}\left[1 + \tanh(\alpha(C - C_*))\right], \end{aligned} \qquad \alpha \gg 1, \tag{2.2}$$

where C_* is the value of the concentration at which the glass-rubber transition occurs. Other physically appropriate forms for β and D are presented in [4], [8], [28], [33], [34], [35], [36], [37].

We examine a polymer that is initially saturated (and hence rubbery) and unstressed, which leads to the initial conditions

$$C^{\mathrm{r}}(x,0) = 1, \qquad \sigma^{\mathrm{r}}(x,0) = 0. \tag{2.3}$$

The end $x = L$ is insulated, while at the exposed surface $x = 0$, the flux is proportional to the difference between the surface concentration and the environment concentration C_{ext}:

$$\left(D(C)\frac{\partial C}{\partial x} + \frac{\partial \sigma}{\partial x}\right)(L,t) = 0, \tag{2.4a}$$

$$\left(D(C)\frac{\partial C}{\partial x} + \frac{\partial \sigma}{\partial x}\right)(0,t) = k[C(0,t) - C_{\mathrm{ext}}], \tag{2.4b}$$

where k is a constant measuring the mass transfer coefficient of the exposed interface.

2.2. Two-state formulation. We solve (2.1)–(2.4) numerically in section 5, but in order to obtain direct dependence of our solution on the physical parameters in the system, we will solve the problem analytically, which necessitates some simplifications.

As $\alpha \to \infty$, the parameters in (2.2) become piecewise constant:

$$\begin{aligned} D(C) &= \begin{cases} D_0\epsilon, & 0 \le C \le C_*, \\ D_{\mathrm{r}}, & C_* \le C \le 1, \end{cases} \\ \beta(C) &= \begin{cases} \beta_{\mathrm{g}}, & 0 \le C \le C_*, \\ \beta_{\mathrm{r}}, & C_* < C \le 1. \end{cases} \end{aligned} \tag{2.5}$$

The rubber is closest to the Fickian regime because the relaxation time is almost instantaneous; thus $\beta_{\mathrm{g}}/\beta_{\mathrm{r}} = \epsilon \ll 1$. It has been shown experimentally [16] that the diffusion coefficient in the glassy region is quite small, so we let $D_{\mathrm{g}} = D_0\epsilon$, where D_0 is an $O(1)$ constant.

With the functional forms in (2.5), it is natural to model the physical system as a two-state problem with a moving boundary $x = s(t)$ representing the glass-rubber interface. Thus, making our substitutions into (2.1), we obtain the following in the glassy region:

$$\frac{\partial C^{\mathrm{g}}}{\partial t} = D_0\epsilon\frac{\partial^2 C^{\mathrm{g}}}{\partial x^2} + \frac{\partial^2 \sigma^{\mathrm{g}}}{\partial x^2}, \tag{2.6a}$$

$$\frac{\partial \sigma^{\mathrm{g}}}{\partial t} + \sigma^{\mathrm{g}} = \gamma\epsilon C^{\mathrm{g}} + \frac{\partial C^{\mathrm{g}}}{\partial t}, \tag{2.6b}$$

while in the rubbery region we have

$$\frac{\partial C^{\mathrm{r}}}{\partial t} = D_{\mathrm{r}} \frac{\partial^2 C^{\mathrm{r}}}{\partial x^2} + \frac{\partial^2 \sigma^{\mathrm{r}}}{\partial x^2}, \tag{2.7a}$$

$$\frac{\partial \sigma^{\mathrm{r}}}{\partial t} + \frac{\sigma^{\mathrm{r}}}{\epsilon} = \gamma \epsilon C^{\mathrm{r}} + \frac{\partial C^{\mathrm{r}}}{\partial t}. \tag{2.7b}$$

With such a formulation, we must have conditions that hold at $x = s(t)$. We impose continuity of concentration at the glass-rubber transition value C_*:

$$C^{\mathrm{r}}(s(t), t) = C^{\mathrm{g}}(s(t), t) = C_*. \tag{2.8}$$

In addition, we require continuity of stress and flux:

$$\sigma^{\mathrm{r}}(s(t), t) = \sigma^{\mathrm{g}}(s(t), t), \tag{2.9a}$$

$$\left(D_{\mathrm{r}} \frac{\partial C^{\mathrm{r}}}{\partial x} + \frac{\partial \sigma^{\mathrm{r}}}{\partial x} \right)(s(t), t) = \left(D_0 \epsilon \frac{\partial C^{\mathrm{g}}}{\partial x} + \frac{\partial \sigma^{\mathrm{g}}}{\partial x} \right)(s(t), t). \tag{2.9b}$$

3. Perturbation solution. We assume perturbation expansions for our dependent variables in ϵ, the small ratio of the relaxation times:

$$C \sim C_0 + O(\epsilon), \qquad \sigma \sim \sigma_0 + O(\epsilon), \tag{3.1}$$

where the same expansions hold for the rubber and glass.

3.1. The glassy region. Substituting (3.1) into (2.6) yields

$$\frac{\partial C_0^{\mathrm{g}}}{\partial t} = \frac{\partial^2 \sigma_0^{\mathrm{g}}}{\partial x^2}, \tag{3.2a}$$

$$\frac{\partial \sigma_0^{\mathrm{g}}}{\partial t} + \sigma_0^{\mathrm{g}} = \frac{\partial C_0^{\mathrm{g}}}{\partial t}. \tag{3.2b}$$

It is simpler to solve for the stress in the glassy region first; hence we combine (3.2) to obtain

$$\frac{\partial \sigma_0^{\mathrm{g}}}{\partial t} + \sigma_0^{\mathrm{g}} = \frac{\partial^2 \sigma_0^{\mathrm{g}}}{\partial x^2}, \qquad 0 < x < s(t). \tag{3.3}$$

In many industrial applications, fast drying is desirable in order to reduce production time and cost. Thus, we consider the case where $k \to \infty$, which corresponds to high mass transfer coefficient or large driving force. (In certain scaling limits, this can also correspond to thick films.) Making this substitution in (2.4b), we obtain

$$C_0^{\mathrm{g}}(0, t) = C_{\mathrm{ext}} < C_*, \tag{3.4}$$

which locates the glass-rubber interface at the origin for $t = 0$. This sort of Dirichlet condition is routinely used in diffusion or heat conduction problems, instead of the more physically realistic flux or activity balance conditions. Nevertheless, we shall see that in this context, imposing such a simple boundary condition can produce unphysical results.

Equation (3.4) implies that the concentration jumps discontinuously at the origin from 1 to C_{ext}, so we have the following:

$$\frac{dC}{dt}(0,t) = (C_{\text{ext}} - 1)\delta(t), \tag{3.5}$$

and upon substituting this result into (3.2b) evaluated at $x = 0$, we obtain

$$\sigma_0^{\text{g}}(0,t) = (C_{\text{ext}} - 1)e^{-t}. \tag{3.6}$$

Note the exponential decay of surface stress from its initial value, reflecting the memory effects in the glassy polymer.

In order to solve the problem, we use an integral method first introduced by Boley [38] and used extensively in this context by Edwards and Cohen [24] and Edwards [26], [29], [39], [40]. Essentially, we wish to write the solution of (3.3) and (3.6) as a Green's function convolved with a *fictitious* initial condition $\sigma_0^{\text{g}}(x,0) = f^{\text{i}}(x)$. This condition is fictitious because the polymer is not glassy at $t = 0$. Thus we extend our domain beyond the region $0 < x < s(t)$. By writing our solution in this form, we reduce the problem from a PDE to an integrodifferential equation.

Since all expressions for $x > s(t)$ are fictitious anyway, we embed the problem in the semi-infinite domain $x > 0$. The solution then is found to be

$$\sigma_0^{\text{g}}(x,t) = (C_{\text{ext}} - 1)e^{-t}\operatorname{erfc}\left(\frac{x}{2\sqrt{t}}\right)$$

$$+\frac{e^{-t}}{2\sqrt{\pi t}}\int_0^\infty f^{\text{i}}(z)\left\{\exp\left[-\frac{(x-z)^2}{4t}\right] - \exp\left[-\frac{(x+z)^2}{4t}\right]\right\}dz. \tag{3.7}$$

3.2. The rubbery region. In the rubbery region we substitute (3.1) into (2.7b) to obtain

$$\sigma_0^{\text{r}}(x,t) = 0. \tag{3.8}$$

Since the γ term does not contribute to the dynamics in either the glassy or the rubbery regions, our model (2.1) contains exactly those dynamical processes as in the models of Cairncross and Durning [8], Durning [27], and Durning and Tabor [28].

Substituting (3.1) and (3.8) into (2.7a) yields

$$\frac{\partial C_0^{\text{r}}}{\partial t} = D_{\text{r}}\frac{\partial^2 C_0^{\text{r}}}{\partial x^2}, \qquad s(t) < x < L, \quad 0 < t < t_L, \tag{3.9a}$$

where $s(t_L) = L$. To use Boley's method to rewrite our solution, we note that upon substituting (3.1) and (3.8) into (2.4a), we obtain

$$\frac{\partial C_0^{\text{r}}}{\partial x}(L,t) = 0, \tag{3.9b}$$

and hence $x = L$ is a line of symmetry. Thus by the method of images

$$C_0^{\text{r}}(x,t) = 1 - [T^{\text{r}}(x,t) + T^{\text{r}}(2L - x,t)] \tag{3.10a}$$

is a solution to (3.9) and (2.3) if $T^{\text{r}}(x,t)$ is a solution of the heat equation. Since the rubber occupies the region $s(t) < x < L$, the fictitious condition is $T^{\text{r}}(0,t) = f^{\text{b}}(t)$, so T^{r} is given by

$$T^{\text{r}}(x,t) = \frac{x}{2\sqrt{D_{\text{r}}\pi}}\int_0^t \frac{f^{\text{b}}(z)}{(t-z)^{3/2}}\exp\left[-\frac{x^2}{4D_{\text{r}}(t-z)}\right]dz. \tag{3.10b}$$

Substituting (3.1) and (3.8) into (2.8) and (2.9), we obtain

$$C_0^{\mathrm{r}}(s(t),t) = C_0^{\mathrm{g}}(s(t),t) = C_*, \tag{3.11a}$$

$$\sigma_0^{\mathrm{g}}(s(t),t) = 0, \tag{3.11b}$$

$$D_{\mathrm{r}}\frac{\partial C_0^{\mathrm{r}}}{\partial x}(s(t),t) = \frac{\partial \sigma_0^{\mathrm{g}}}{\partial x}(s(t),t). \tag{3.12}$$

Upon substitution of (3.7) and (3.10) into (3.11) and (3.12), we will obtain three integrodifferential equations for the unknowns f^{i}, f^{b}, and s.

4. Short-time solutions.

4.1. The outer solution. We examine the small-time asymptotics of our analytic solution as in Edwards [26] by letting

$$f^{\mathrm{i}}(x) \sim f_0^{\mathrm{i}}, \qquad f^{\mathrm{b}}(t) \sim f_0^{\mathrm{b}}, \qquad s(t) = 2s_0 t^n, \qquad t \to 0, \quad x \to 0. \tag{4.1}$$

We substitute (4.1) into (3.10) and (3.7) to obtain expressions for our unknowns for $L = O(1)$. Substituting these expressions into (3.11) and (3.12), we obtain

$$C_* \sim 1 - f_0^{\mathrm{b}} \operatorname{erfc}\left(\frac{s_0 t^{n-1/2}}{\sqrt{D_{\mathrm{r}}}}\right), \tag{4.2a}$$

$$0 \sim (C_{\mathrm{ext}} - 1)\operatorname{erfc}(s_0 t^{n-1/2}) + f_0^{\mathrm{i}} \operatorname{erf}(s_0 t^{n-1/2}), \tag{4.2b}$$

$$\frac{D_{\mathrm{r}} f_0^{\mathrm{b}}}{\sqrt{\pi D_{\mathrm{r}} t}} \exp\left(-\frac{s^2}{4D_{\mathrm{r}} t}\right) \sim [f_0^{\mathrm{i}} - (C_{\mathrm{ext}} - 1)]\frac{e^{-t}}{\sqrt{\pi t}} \exp\left(-\frac{s^2}{4t}\right). \tag{4.3}$$

Equation (4.2a) can be satisfied if and only if $n \geq 1/2$. Equation (4.2b) can be satisfied if and only if $n \leq 1/2$. Therefore $n = 1/2$ and initially the front moves in a purely Fickian way because the nonlinear memory effects have not yet had time to develop. Using this result, we obtain

$$g_1(s_0) \equiv \sqrt{D_{\mathrm{r}}}\frac{1 - C_*}{\operatorname{erfc}(s_0/\sqrt{D_{\mathrm{r}}})} \exp\left(-\frac{s_0^2}{D_{\mathrm{r}}}\right) = \frac{1 - C_{\mathrm{ext}}}{\operatorname{erf} s_0} \exp\left(-s_0^2\right) \equiv g_2(s_0). \tag{4.4}$$

Figure 4.1 shows plots of $g_2 - g_1$ for various values of C_*. The s_0-intercept marks the value of the front speed. Note that as C_* increases, the front speed increases since not as much penetrant has to desorb to move the front along.

Figure 4.2 shows the variance in the front speed as D_{r} and C_{ext} vary. Note that as C_{ext} decreases, the front speed increases because of a larger driving force. In addition, as D_{r} increases, the front speed decreases because it is easier to diffuse penetrant to the front.

Using (4.4), we may derive the value of f_0^{b} and hence obtain

$$C_0^{\mathrm{r}}(x,t) = 1 - \frac{1 - C_*}{\operatorname{erfc}(s_0/\sqrt{D_{\mathrm{r}}})}\left[\operatorname{erfc}\left(\frac{x}{2\sqrt{D_{\mathrm{r}} t}}\right) + \operatorname{erfc}\left(\frac{2L - x}{2\sqrt{D_{\mathrm{r}} t}}\right)\right]. \tag{4.5}$$

As $L \to \infty$, the second term drops out and we are left with exactly the expression in Edwards [26] for the case of a semi-infinite domain. In addition, as $t \to 0$ the second term modeling "reflections" from $x = L$ is negligible.

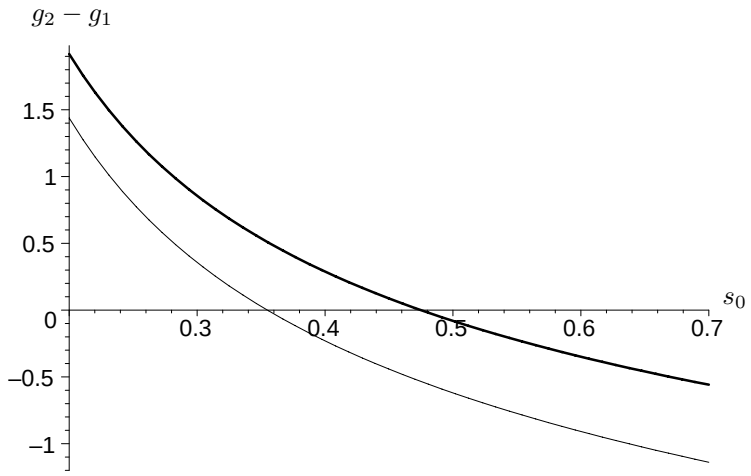


FIG. 4.1. $g_2 - g_1$ *versus* s_0 *for* $D_\mathrm{r} = 7$, $C_\mathrm{ext} = 1/3$. *Thin line:* $C_* = 1/2$. *Thick line:* $C_* = 2/3$.

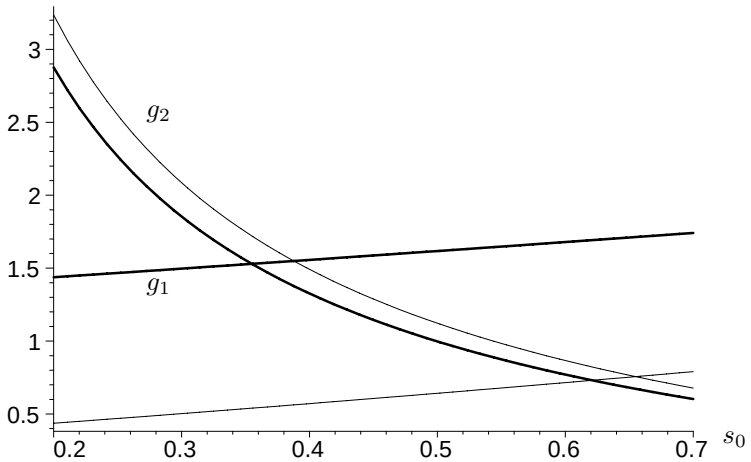


FIG. 4.2. g_1 *and* g_2 *versus* s_0 *for* $C_* = 1/2$. *Thick lines:* $D_\mathrm{r} = 0.4$, $C_\mathrm{ext} = 1/4$. *Thin lines:* $D_\mathrm{r} = 7$, $C_\mathrm{ext} = 1/3$.

We may also use (4.4) to derive the value of f_0^i and hence obtain

$$\sigma_0^\mathrm{g}(x,t) \sim (C_\mathrm{ext} - 1)e^{-t}\left[1 - \frac{1}{\operatorname{erf} s_0}\operatorname{erf}\left(\frac{x}{2\sqrt{t}}\right)\right]. \tag{4.6}$$

Substituting (4.6) into (3.2a) and solving using (3.11a), we have the following:

$$\begin{aligned} C^\mathrm{g}(x,t) = C_* + \frac{C_\mathrm{ext} - 1}{2\operatorname{erf} s_0}\Bigg\{ & e^{-x}\left[\operatorname{erfc}\left(-\sqrt{t} + \frac{x}{2\sqrt{t}}\right) - \operatorname{erfc}\left(s_0 - \frac{x}{2s_0}\right)\right] \\ & + e^{x}\left[\operatorname{erfc}\left(\sqrt{t} + \frac{x}{2\sqrt{t}}\right) - \operatorname{erfc}\left(s_0 + \frac{x}{2s_0}\right)\right]\Bigg\}, \end{aligned} \tag{4.7}$$

where the x/s_0 terms come from the asymptotic expansion of $s^{-1}(x)$, the inverse function for the front position:

$$s^{-1}(x) \sim \left(\frac{x}{2s_0}\right)^2, \qquad x \to 0. \tag{4.8}$$

Unfortunately, we note that if we substitute $x = 0$ into (4.7), we obtain

$$\lim_{x\to 0} C_0^{\mathrm{g}}(x,t) = C_* + C_{\mathrm{ext}} - 1 \neq C_{\mathrm{ext}} = C^{\mathrm{g}}(0,t). \tag{4.9}$$

The discontinuity near $x = 0$ must be resolved by a boundary layer, but even the solution to the full problem near $x = 0$ will be less than C_{ext}. We call this excessive drying near the exposed surface *desorption overshoot*, as the minimum of the concentration now occurs inside the film. The terminology is motivated by the related phenomenon of *sorption overshoot*, where the concentration rises above its equilibrium value during a sorption experiment [41].

Moreover, it is certainly possible for $C_* + C_{\mathrm{ext}} - 1 < 0$, which would yield the physically unrealistic result of a negative concentration. This unphysical aspect is not an artifact of the asymptotics; rather it is the direct result of the Dirichlet condition (3.4), as discussed in section 6.

4.2. The corner layer. The discontinuity about $x = 0$ is caused by the form of the operator in (3.2a). As long as the evolution equation for C has only a $\partial C/\partial t$ term in it, then σ and C will differ *everywhere* only by a function of x. Since both C and σ_0^{r} are constants along the front, that difference must also be a constant at the front. This causes a discontinuity because

$$\lim_{t\to 0} C(s(t),t) \neq \lim_{t\to 0} C(0,t).$$

σ_0^{r} does not vary along the front due to the ϵ^{-1} term in (2.7b). This term can be counteracted if we introduce a corner layer near the origin *via* the following substitutions:

$$C^{\mathrm{r}}(x,t) = C^+(\xi,\tau), \qquad \sigma^{\mathrm{r}}(x,t) = \sigma^+(\xi,\tau), \qquad \xi = \frac{x}{\epsilon^{1/2}}, \qquad \tau = \frac{t}{\epsilon}. \tag{4.10}$$

Substituting (4.10) into (2.7), we obtain

$$\frac{\partial C^+}{\partial \tau} = D_{\mathrm{r}} \frac{\partial^2 C^+}{\partial \xi^2} + \frac{\partial^2 \sigma^+}{\partial \xi^2}, \tag{4.11a}$$

$$\frac{\partial \sigma^+}{\partial \tau} + \sigma^+ = \frac{\partial C^+}{\partial \tau}, \tag{4.11b}$$

which is just the full system (2.7) without the γ term. Hence even in the corner layer, our model matches that of Cairncross and Durning [8], Durning [27], and Durning and Tabor [28]. Note that τ is the time scale for relaxation in the rubber.

To solve this system, we must proceed numerically. Nevertheless, we note that due to the exponential decay inherent in (4.11b), curves of constant C are not curves of constant σ. This fact will remove the discontinuity, which was caused by the fact that the front was an isocline for both outer solutions.

5. Numerical computations. We compare our asymptotic results to those from a finite-element code previously used to solve a similar model [8]. The code solves (2.1)–(2.4) using finite elements with quadratic basis functions. In order to resolve the boundary layer, the domain was discretized into sixty fixed but unequally spaced elements, with more elements placed near $x = 0$.

Application of the finite element method to the model results in a system of nonlinear coupled ODEs for the nodal values of concentration and stress. The system of ODEs was integrated in time using a stiff DAE solver, DASSL [42]. DASSL uses an Adams–Bashforth–Moulton predictor-corrector algorithm with a variable-order backward differentiation formula. The corrector is implicit and the nonlinear system is solved by Newton's method with an analytical Jacobian matrix. The time step is automatically updated to control the estimated error within a specified tolerance. The error tolerance and number and distribution of elements were adjusted until the results were insensitive to the size of these parameters.

5.1. Comparison with asymptotics. The parameters chosen for use in both the analytical model and the numerical simulations were as follows:

$$C_* = 1/2, \qquad D_{\mathrm{r}} = 4, \qquad C_{\mathrm{ext}} = 1/4, \qquad L = 3, \qquad D_{\mathrm{g}} = 4 \times 10^{-4}, \tag{5.1a}$$

$$\beta_{\mathrm{g}} = 1, \qquad \beta_{\mathrm{r}} = 10^4, \qquad \alpha = 80, \qquad k = 1.33 \times 10^4. \tag{5.1b}$$

These parameters essentially correspond to an ϵ value of 10^{-4}. Also, with these parameters, $s_0 \approx 0.4550$.

Figure 5.1 shows a comparison of the asymptotic and numerical predictions of the front position for small time. The speed of the front decreases with time as predicted by both methods. The agreement between the asymptotic and numerical results is excellent.

In Figure 5.2 we show a graph of the concentration for the parameters in (5.1) and the times listed. The interval in x is restricted near $x = 0$; the grid spacing decreases as we reach that endpoint. There is excellent agreement between the numerical and outer solutions in the region away from the boundary layer, and the discontinuity in

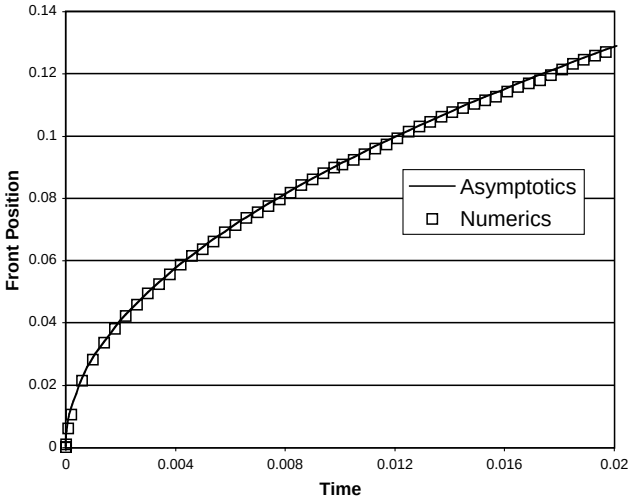


FIG. 5.1. *Asymptotic and numerical calculations of $s(t)$ for the parameters in* (5.1).

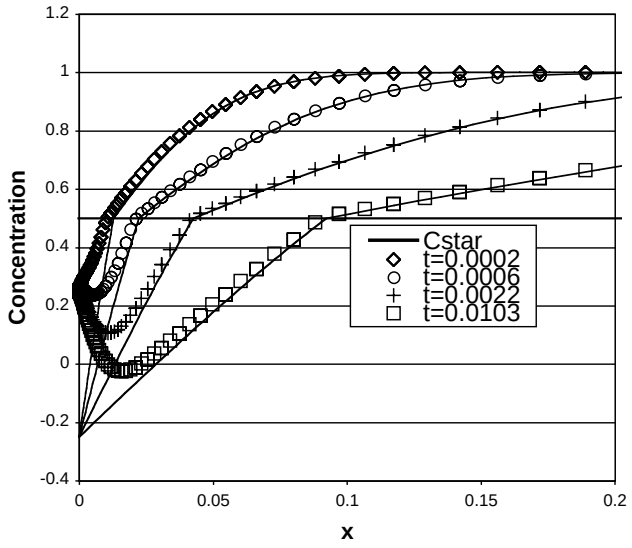


FIG. 5.2. *Aymptotic solution $C_0(x,t)$ (lines) and numerical solution (symbols) versus x. The numerical and asymptotic solutions are indistinguishable beyond $x = 0.2$.*

$\partial C/\partial x$ at the glass-rubber transition is accurately predicted by both techniques. Note that at $t = 0.0103$ even the numerical solution goes negative, so we have confirmed that negative concentration values are not an artifact of the asymptotic solution. We shall examine the root causes of this phenomenon in the next section.

Figure 5.3 shows a graph of the stress versus x for the times listed. There is excellent agreement between the asymptotic and numerical solutions for the glassy stress. In addition, the zero-stress approximation (3.8) and the numerical calculation

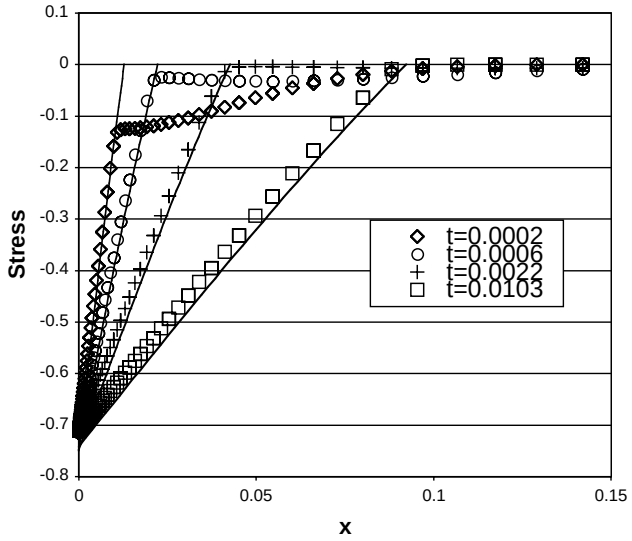


FIG. 5.3. *Aymptotic solution $\sigma_0^{\mathrm{g}}(x,t)$ (lines) and numerical solution (symbols) versus x. The rubbery stress is zero. The numerical and asymptotic solutions are indistinguishable beyond $x = 0.15$.*

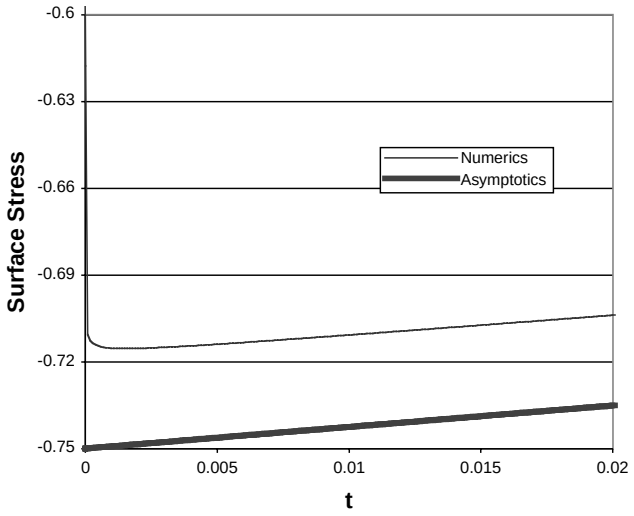


FIG. 5.4. *Comparison of asymptotic (from equation* (3.6)*) and numerical expressions for surface stress.*

in the rubbery region match for larger times. From the graph at $t = 2 \times 10^{-4}$ we see that due to the rapid drop in C at the surface, the stress in the rubbery region is initially $O(1)$, as shown in section 4.2. For longer times, the stress decays exponentially as predicted by (4.11b), until at $t = 2.2 \times 10^{-3}$, the numerical calculation is virtually indistinguishable from (3.8).

Figure 5.4 shows a graph of the surface stress at $x = 0$. Though the outer and numerical solutions decay on the same e^{-t} scale, there is a persistent gap because the outer solution in (3.6) assumes an instantaneous change in C at $t = 0$, while the numerical simulations follow (2.4b). Thus, initially the surface concentration and surface stress evolve on a time scale roughly proportional to k^{-1}. This time scale will become important later on when we examine the reason for the negative concentration values.

5.2. Long-time results. Though the validity of the asymptotics ends for moderate times, we can certainly continue the numerical calculations into that region. Figure 5.5 shows the computed concentration profiles for various times. Note that between $t = 2$ and $t = 4$, the film becomes entirely glassy. (For more discussion of the time at which the front reaches the back of the film, see the appendix.) Since the glass has a longer relaxation time, the change in the concentration between $t = 2$ and $t = 4$ is relatively small.

The unphysical negative concentration is not a brief anomaly; it continues for moderate time, and the size of the dip actually increases. It should be noted that the desorption overshoot disappears if k is smaller, which corresponds to a slower change in the surface concentration. For more discussion of this topic, see section 6.

Figure 5.6 shows the computed stress profiles for the same series of times. The nearly linear stress in the glassy region implies a constant non-Fickian flux. Thus, the evolution of the concentration in this region is dominated by the Fickian flux. Note that the surface stress continues its exponential decay to a final limiting value of zero.

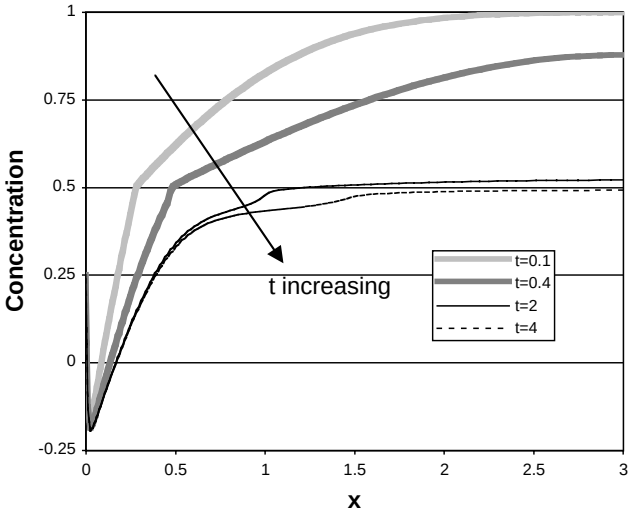


FIG. 5.5. *Computed concentration profiles versus x for the times listed in the legend.*

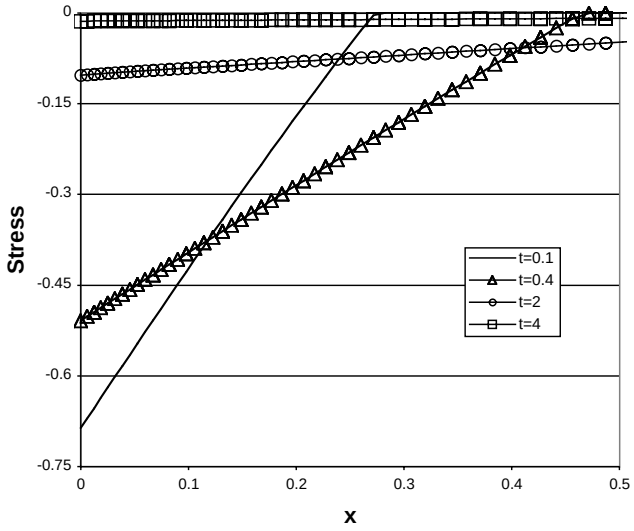


FIG. 5.6. *Computed stress profiles versus x for the times listed in the legend.*

6. Explaining negative concentrations. The unphysical negative concentration is not an artifact of the asymptotics, as the numerical solutions in Figure 5.2 show. To explain the phenomenon, we solve (2.6b) for short times, using (3.6) and (3.2). After some work, we obtain the following expression:

$$\lim_{x\to 0} C_0^{\mathrm{g}}(x,t) = \lim_{t\to 0} \sigma^{\mathrm{g}}(0,t) + C_*. \tag{6.1}$$

Hence the discontinuity *in the outer solution* exists for all time unless

$$\lim_{t\to 0} \sigma^{\mathrm{g}}(0,t) = C_{\mathrm{ext}} - C_*. \tag{6.2}$$

Moreover, the concentration will go negative whenever

$$\lim_{t\to 0} \sigma^{\mathrm{g}}(0,t) < -C_*. \tag{6.3}$$

How then to avoid satisfying this condition?

Following common practice for diffusion and heat transfer problems, we took $k \to \infty$ in (2.4b) to obtain the Dirichlet condition (3.4). The resulting discontinuous jump at $t = 0$ forces $\sigma(0, 0^+) = C_{\mathrm{ext}} - 1$, as can be seen from (3.6), and this value can violate (6.3). Why does the standard Dirichlet trick not work for this model?

In a standard diffusion problem, lines of constant t are characteristics. Thus these equations transmit disturbances with infinite signal speed to the entire domain. As can be intuited from the leading-order outer equations (3.3) and (3.9a), the same is true for this model. This explains why the solution does not break down in any mathematical sense; it just goes negative, which offends our physical sensibilities.

The key difference rather is the delay term inherent in (2.1b). When a jump occurs very quickly, the stress cannot relax fast enough (even with an $O(\epsilon)$ relaxation time in the rubber) to equilibrate it. Once a large stress gradient has been introduced at the exposed surface, there is no mechanism in (3.2a) to stop the concentration from going negative. This is related to the observation in [41] that other models for anomalous diffusion will have negative concentration values if a "retardation time" is not included.

There are several mechanisms one can introduce to moderate the concentration dip. For instance, consider the case where the $\partial\sigma/\partial x$ term in (2.1a) is multiplied by a "stress diffusion coefficient" $E(C)$, where $E(0) = 0$. This term would remain at leading order in the equation analogous to (3.2a), causing $\partial C_0^{\mathrm{g}}/\partial t(C = 0) = 0$ and preventing the concentration from going negative. Moreover, preliminary numerical calculations indicate that if $E(C) \ll 1$ in the glassy region, this change can eliminate negative concentrations while maintaining desorption overshoot.

Another remedy is to slow the change so that it occurs on the fast relaxation time scale of the rubbery polymer. Thus we replace (3.4) by

$$C(0,\tau) = C_{\mathrm{ext}} + (1 - C_{\mathrm{ext}})e^{-\lambda\tau}, \qquad \lambda \neq 1, \tag{6.4}$$

where τ is the time scale defined in (4.10). The exponential form is chosen to match the analysis in Edwards [24] and the forms in Hui et al. [34] and Long and Richman [43]; $\lambda \neq 1$ is taken for simplicity. As λ increases, the driving force increases and the transition between rubber and glass steepens.

As given by (6.4), the interface is now rubbery for some interval. We may substitute (6.4) into the leading orders of (2.6b) and (2.7b) and solve to obtain the stress boundary condition. Since τ is an initial-layer variable, we may take the limit of this condition as $\tau \to \infty$ to find the limiting value of the outer boundary condition. This is found to be

$$\lim_{t\to 0} \sigma^{\mathrm{g}}(0,t) = -\frac{C_* - C_{\mathrm{ext}}}{1-\lambda} + \frac{\lambda(1 - C_{\mathrm{ext}})}{1-\lambda}\left(\frac{C_* - C_{\mathrm{ext}}}{1 - C_{\mathrm{ext}}}\right)^{1/\lambda}. \tag{6.5}$$

Thus our matching condition, and hence $C^{\mathrm{g}}(x, 0)$, depends on λ. As $\lambda \to \infty$, (6.4) approaches a step function and our result from section 3 holds:

$$\lim_{t\to 0} \sigma^{\mathrm{g}}(0,t) = -(1 - C_{\mathrm{ext}}), \qquad \lambda \to \infty. \tag{6.6a}$$

If instead $\lambda \to 0$, we obtain the following:

$$\lim_{t\to 0} \sigma^{\mathrm{g}}(0,t) = -(C_* - C_{\mathrm{ext}}), \quad \lambda \to 0, \tag{6.6b}$$

which from (6.2) is exactly the condition required to eliminate the boundary layer at the exposed surface. In this case the exterior concentration varies on a time scale slower than that of the rubber relaxation time, so the *entire* polymer can equilibrate to the exterior.

Last, we note that from (6.3) that in order to maintain a positive concentration, the expression in (6.5) must be greater than $-C_*$. Hence the generation of negative concentrations in our model can be remedied by imposing more physically realistic boundary conditions. If changes in the exterior happen on a faster time scale than the rubber relaxation time scale, the surface cannot immediately equilibrate. Thus, the polymer exhibits a sort of "self-regulation" which puts restrictions on the speed at which the surface concentration can change. This sort of self-regulation has been seen in similar models of sorption processes [24].

7. Conclusions. During the desorption of saturated polymers near the glass-rubber transition temperature, a glassy skin will form near the exposed surface. One mechanism for the formation of such a skin is viscoelastic relaxation in the polymer network. The mathematical model presented here has captured this behavior in previous numerical [5], [8] and analytical [26], [29] studies. However, never before has the model been studied in both ways simultaneously. This merging of techniques involved restricting the analytical study to a more realistic finite domain and adapting the numerical parameter scheme to approximate a piecewise-constant approach. By approaching the solution in two ways, we validated both approaches. In particular, we established that negative concentrations were the result of neither a computational bug nor an erroneous asymptotic approximation, but were rather the predictable and robust result of a mathematically simple, but physically unrealistic, boundary condition.

In the asymptotic work, the parameters are taken as piecewise constant and the system is treated in a manner similar to a Stefan problem. Since the system is not amenable to similarity solutions, an integral method based on the one in Boley [38] is used. The finite domain is extended to a semi-infinite one in both cases, and the method of images is used to handle the insulated boundary condition at $x = L$.

The asymptotic and numerical results match well, showing a quick transition to the glassy region near the exposed surface. The glass-rubber interface initially moves like $t^{1/2}$, reflecting the fact that the viscoelastic memory effects have not yet had time to develop. The numerical solutions demonstrated desorption overshoot, where a minimum in the concentration occurs in the interior of the domain. This is mirrored in the asymptotic outer solution, which is less than the imposed surface concentration as $x \to 0$.

The overshoot can be traced to our replacement of a flux balance condition with a Dirichlet condition. Such approximations are routinely used in diffusion and heat conduction problems instead of the more complicated (but physically realistic) activity or flux conditions. However, in our case taking the limit of large k leads to negative concentrations. Essentially, we are attempting to force the surface concentration to vary faster than the polymer can adapt. The intrinsic time scale in the model then reduces the set of boundary conditions that can lead to physically meaningful results.

In section 6 we proposed two remedies for negative concentrations. A stress diffusion coefficient can be introduced which shuts down further penetrant diffusion

when the polymer is dry. Alternatively, if we vary the surface concentration on the fast τ time scale, we eliminate the negative concentrations. Essentially, the τ-variance is the fastest that the actual physical system can accommodate. This type of self-regulation has been demonstrated in sorption models [24].

Though the numerical and asymptotic profiles match well for small t, the appendix shows that due to the diffusive nature of the operators considered, the fictitious boundary conditions must be approximated very closely to guarantee accurate results for moderate t. Nevertheless, the agreement for small time provides a sturdy background on which to base further work. Not only do the asymptotic solutions verify the numerics, but they also demonstrate parameter ranges which produce unphysical results.

Appendix. Some remarks on the intersection point. We conclude by examining the solution near the time $t = t_L$ where $s(t_L) = L$. Due to the symmetry about the line $x = L$, $s^{-1}(x)$ should be even about $x = L$, so the first terms in our expansion for $s(t)$ should be

$$s(t) \sim L - s_1\sqrt{r}, \qquad r = t_L - t > 0, \qquad s_1 > 0. \tag{A.1}$$

Using (3.10a) and (A.1) in (3.11), we may construct an expansion in r, eventually reaching the following terms at $O(r)$:

$$(s_1^2 - 2D_{\mathrm{r}})\frac{\partial^2 T^{\mathrm{r}}}{\partial x^2}(L, t_L) = 0, \tag{A.2a}$$

$$\left(\frac{s_1^2}{2} - 1\right)\frac{\partial^2 \sigma_0^{\mathrm{g}}}{\partial x}(L, t_L) = 0. \tag{A.2b}$$

It can be shown that if the second derivative of T^{r} vanishes at $x = L$, *all* even derivatives of T^{r} must vanish there. But the numerical solutions shown later in Figure A.1 do not support this transcendental vanishing. Thus, we set $s_1 = \sqrt{2D_{\mathrm{r}}}$.

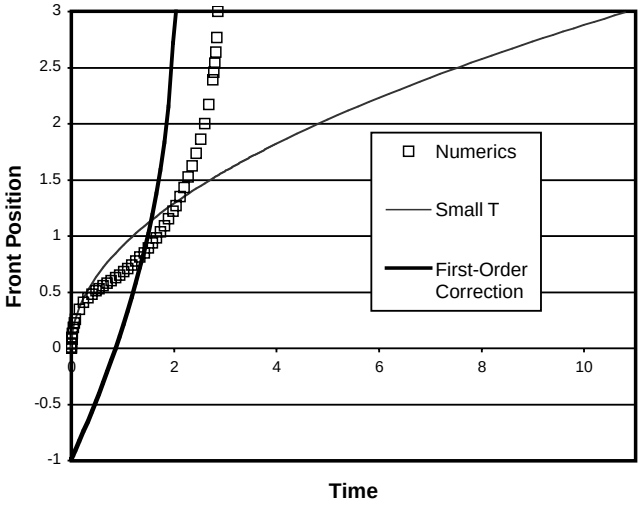


FIG. A.1. *Comparison of short- and intersection-time expressions for $s(t)$ with true front position. Here $s_1 = \sqrt{8}$, $s_L = 1.0634$, and $t_L = 1.990$.*

(Formally, this must be done in a limiting way by setting the stress in the glass equal to a small quantity representing σ_1^{r}, then taking that term to zero.)

To determine t_L, we introduce another degree of freedom into our fictitious condition as follows:

$$f^{\mathrm{i}}(x) \sim f_0^{\mathrm{i}} + f_1^{\mathrm{i}} x. \tag{A.3}$$

Substituting (A.3) into (3.10a) using (4.8), we may combine the resulting equations to obtain the following equation involving t_L:

$$\operatorname{erf} s_0 = \operatorname{erf} s_L - \frac{2s_L}{\sqrt{\pi}} e^{-s_L^2}, \qquad s_L = \frac{L}{2\sqrt{t_L}}. \tag{A.4}$$

Note the underlying parabolic nature of the operator, as evidenced by the relationship between the definition of s_L and the diffusion equation similarity variable. It can be shown that (A.4) has exactly one root s_L. In addition, $s_L > s_0$, so the front must speed up as time passes. Figure A.1 shows a graph of the short- and intersection-time expressions for $s(t)$ as compared with the numerical calculations.

By replacing the one-term expansion for $f^{\mathrm{i}}(x)$ with a two-term expression, we obtain closer agreement near $x = L$, as desired. In particular, we note that by combining the asymptotic results, we obtain the change in concavity of the graph and an acceptable estimate of the inflection time.

Unfortunately, though the two-term expansion provides an improved estimate of t_L, it is still not very accurate. Due to the diffusive nature of the underlying problem, the estimate of the initial condition must be highly accurate to obtain reasonable predictions for moderate t. In addition, the constructed solution does not work well for small times ($r = O(1)$). Thus, as a next step one should construct a three-term expansion for $f^{\mathrm{i}}(x)$ that satisfies the leading-order conditions at both $x = 0$ and $x = L$. This sort of iterative process, where one continually improves the form of f^{i}, should converge to the correct solution on finite domains. Infinite domains are fundamentally different since $t \to \infty$. This case can be treated asymptotically using appropriately chosen expansion functions [26], [29], [39], [40].

Acknowledgments. We thank the reviewers for many helpful comments that improved the manuscript.

REFERENCES

[1] N. Thomas and A. H. Windle, *Transport of methanol in poly-(methyl-methocry-late)*, Polymer, 19 (1978), pp. 255–265.

[2] M. Vinjamur and R. A. Cairncross, *A high airflow drying experimental set-up to study drying behavior of polymer solvent systems*, Drying Tech. J., 19 (2001), pp. 1591–1612.

[3] M. Vinjamur and R. A. Cairncross, *Experimental investigations of trapping skinning*, J. Appl. Polym. Sci., 83 (2002), pp. 2269–2273.

[4] M. Vinjamur and R. A. Cairncross, *A non-Fickian non-isothermal model to predict anomalous trapping skinning behaviour during drying of polymer coatings*, AIChE J., to appear.

[5] M. Vinjamur and R. A. Cairncross, *Guidelines for dryer design based on results from non-Fickian model*, J. Appl. Polym. Sci., to appear.

[6] R. A. Cairncross, L. F. Francis, and L. E. Scriven, *Competing drying and reaction mechanisms in the formation of sol-to-gel films, fibers, and spheres*, Drying Tech. J., 10 (1992), pp. 893–923.

[7] R. A. Cairncross, L. F. Francis, and L. E. Scriven, *Predicting drying in coatings that react and gel: Drying regime maps*, AIChE J., 42 (1996), pp. 55–67.

[8] R. A. Cairncross and C. J. Durning, *A model for drying of viscoelastic coatings*, AIChE J., 42 (1996), pp. 2415–2425.

[9] E. Ben-Yoseph and R. W. Hertel, *Computer modeling of sugar crystallization during drying of thin sugar films*, J. Crys. Growth, 198/199 (1999), pp. 1294–1298.
[10] E. Ben-Yoseph, R. W. Hertel, and D. Howling, *Three dimensional model of phase transition of thin sucrose films during drying*, J. Food Eng., 44 (2000), pp. 13–22.
[11] S. Simal, A. Femenia, P. Llull, and C. Rossello, *Dehydration of aloe vera: Simulation of drying curves and evaluation of functional properties*, J. Food Eng., 43 (2000), pp. 109–114.
[12] J. Crank, *The influence of concentration-dependent diffusion on rate of evaporation*, Proc. Phys. Soc., 63 (1950), pp. 484–491.
[13] J. Crank, *Diffusion in media with variable properties, part* III: *Diffusion coefficients which vary discontinuously with concentration*, Trans. Faraday Soc., 47 (1951), pp. 450–461.
[14] C. A. Finch, ed., *Chemistry and Technology of Water-Soluble Polymers*, Plenum Press, New York, 1983.
[15] W. R. Vieth, *Diffusion in and Through Polymers: Principles and Applications*, Oxford University Press, Oxford, 1991.
[16] J. S. Vrentas, C. M. Jorzelski, and J. L. Duda, *A Deborah number for diffusion in polymer-solvent systems*, AIChE J., 21 (1975), pp. 894–901.
[17] M. Dabral, *Solidification of Coatings: Theory and Modeling of Drying, Curing, and Microstructure Growth*. Ph.D. thesis, University of Minnesota, Minneapolis-St. Paul, MN, 1999.
[18] N. O. Ngu and S. K. Mallapragada, *Quantitative analysis of crystallization and skin formation during isothermal solvent removal from semicrystalline polymers*, Polymer, 40 (1999), pp. 5393–5400.
[19] I. H. Romdhane, P. E. Price, Jr., C. A. Miller, P. T. Benson, and S. Wang, *Drying of glassy polymer films*, Ind. Eng. Chem. Res., 40 (2001), pp. 3065–3075.
[20] H. L. Frisch, *Sorption and transport in glassy polymers—a review*, Polymer Engr. Sci., 20 (1980), pp. 2–13.
[21] D. R. Paul and W. J. Koros, *Effect of partially immobilizing sorption on permeability and diffusion time lag*, J. Polym. Sci., 14 (1976), pp. 675–685.
[22] W. R. Vieth and K. J. Sladek, *A model for diffusion in a glassy polymer*, J. Colloid Sci., 20 (1965), pp. 1014–1033.
[23] J. Crank, *The Mathematics of Diffusion*, 2nd ed., Clarendon Press, Oxford, 1976.
[24] D. A. Edwards and D. S. Cohen, *A mathematical model of a dissolving polymer*, AIChE. J., 41 (1995), pp. 2345–2355.
[25] D. A. Edwards and D. S. Cohen, *An unusual moving boundary condition arising in anomalous diffusion problems*, SIAM J. Appl. Math., 55 (1995), pp. 662–676.
[26] D. A. Edwards, *Skinning during desorption of polymers: An asymptotic analysis*, SIAM J. Appl. Math., 59 (1999), pp. 1134–1155.
[27] C. J. Durning, *Differential sorption in viscoelastic fluids*, J. Polym. Sci. B, 23 (1985), pp. 1831–1855.
[28] C. J. Durning and M. Tabor, *Mutual diffusion in concentrated polymer solutions under a small driving force*, Macromolecules, 19 (1986), pp. 2220–2232.
[29] D. A. Edwards, *A mathematical model for trapping skinning in polymers*, Stud. Appl. Math. (1997), pp. 49–80.
[30] J. C. Wu and N. A. Peppas, *Modeling of penetrant diffusion in glassy polymers with an integral sorption Deborah number*, J. Polym. Sci. B, 31 (1993), pp. 1503–1518.
[31] T. P. Witelski, *Traveling wave solutions for case* II *diffusion in polymers*, J. Polym. Sci. B, 34 (1996), pp. 141–150.
[32] H. L. Frisch, T. K. Kwei, and T. T. Wang, *Diffusion in glassy polymers*, II, J. Polym. Sci. A-2, 7 (1969), pp. 879–887.
[33] S. Mehdizadeh and C. J. Durning, *Predictions of differential sorption kinetics near T_g for benzene in polystyrene*, AICHE J., 36 (1990), pp. 877–884.
[34] C. Y. Hui, K. C. Wu, R. C. Lasky, and E. J. Kramer, *Case* II *diffusion in polymers.* I. *Transient swelling*, J. Appl. Phys., 61 (1987), pp. 5129–5136.
[35] C. Y. Hui, K. C. Wu, R. C. Lasky, and E. J. Kramer, *Case* II *diffusion in polymers.* II. *Steady state front motion*, J. Appl. Phys., 61 (1987), pp. 5137–5149.
[36] N. Thomas and A. H. Windle, *A deformation model for case* II *diffusion*, Polymer, 21 (1980), pp. 613–619.
[37] D. S. Cohen and A. B. White, Jr., *Sharp fronts due to diffusion and stress at the glass transition in polymers*, J. Polymer Sci. B, 27 (1989), pp. 1731–1747.
[38] B. A. Boley, *A method of heat conduction analysis of melting and solidification problems*, J. Math. Phys., 40 (1961), pp. 300–313.
[39] D. A. Edwards, *Constant front speed in weakly diffusive non-Fickian systems*, SIAM J. Appl. Math., 55 (1995), pp. 1039–1058.

[40] D. A. Edwards, *The effect of a varying diffusion coefficient in polymer-penetrant systems*, IMA J. Appl. Math., 55 (1995), pp. 49–66.

[41] N. S. Kalospiros, R. Ocone, G. Astarita, and J. H. Meldon, *Analysis of anomalous diffusion and relaxation in solid polymers*, Ind. Eng. Chem. Res., 30 (1991), pp. 851–864.

[42] L. R. Petzold, *A Description of DASSL: A Differential/Algebraic System Solver*, Sandia technical report 82-8637, Sandia National Laboratories, Livermore, CA, 1982.

[43] F. A. Long and D. Richman, *Concentration gradients for diffusion of vapors in glassy polymers and their relation to time dependent diffusion phenomena*, J. Am. Chem. Soc., 82 (1960), pp. 513–519.

SIAM J. APPL. MATH.
Vol. 63, No. 1, pp. 116–148

A SIMPLE PREDICTION ALGORITHM FOR THE LAGRANGIAN MOTION IN TWO-DIMENSIONAL TURBULENT FLOWS*

LEONID I. PITERBARG† AND TAMAY M. ÖZGÖKMEN‡

Abstract. A new algorithm is suggested for prediction of a Lagrangian particle position in a stochastic flow, given observations of other particles. The algorithm is based on linearization of the motion equations and appears to be efficient for an initial tight cluster and small prediction time. A theoretical error analysis is given for the Brownian flow and a stochastic flow with memory. The asymptotic formulas are compared with simulation results to establish their applicability limits. Monte Carlo simulations are carried out to compare the new algorithm with two others: the center-of-mass prediction and a Kalman filter–type method. The algorithm is also tested on real data in the tropical Pacific.

Key words. stochastic flow, Lagrangian motion, prediction, stochastic simulations, oceanographic applications

AMS subject classifications. 76F25, 76F55, 86A05, 62M20

PII. S003613990139194X

1. Introduction.

The problem discussed is motivated by applications to rescue and search operations in the sea. An important part of such operations is to properly narrow the search area based on the best possible prediction of the position of a lost object, given its approximate initial position (Schneider (1998)). It is hard to make a reasonable prediction based only on knowledge of the mean current, because of strong velocity fluctuations drifting the object away from the path indicated by the mean velocity field. One can expect more realistic help from other floating objects in the same area, like debris or drifters (human-made floats), which can be observed from planes or satellites. We consider here a simplified model of such a situation, as follows. Several current following floats are released simultaneously at different known positions in a stochastic flow. One of the floats, called the predictand, is unobservable, while the remaining floats, called the predictors, are observed. The problem is to predict the position of the unobservable float, given the above observations. In addition to practical needs, this problem is of great importance from a theoretical viewpoint since it addresses the predictability issue for the Lagrangian motion in turbulent flows. Here, by turbulent flows, we mean velocity fields with fluctuations described by stochastic differential equations. Thus, a kinematic approach is employed: given flow statistics one should conclude with the mean square error of a prediction algorithm. The mathematical framework we set up here is as follows.

Let $\mathbf{u}(t, \mathbf{r})$ be a random velocity field varying in time. By the method of applications we consider only the two-dimensional case: $\mathbf{u}, \mathbf{r} \in R^2$. Consider $M > 1$ Lagrangian (current following) particles starting at time $t = 0$ from different positions

*Received by the editors July 6, 2001; accepted for publication (in revised form) January 28, 2002; published electronically August 28, 2002. This research was supported by Office of Naval Research grants N00014-99-0042 and N00014-99-1-0049.

http://www.siam.org/journals/siap/63-1/39194.html

†Center for Applied Mathematical Sciences, University of Southern California, Los Angeles, CA 90089 (piter@math.usc.edu).

‡Division of Meteorology and Physical Oceanography, Rosenstiel School of Marine and Atmospheric Science, University of Miami, 4600 Rickenbacker Causeway, Miami, FL 33149.

$\mathbf{r}_1^0, \mathbf{r}_2^0, \ldots, \mathbf{r}_M^0$. Their motion is covered by the following system of $2M$ equations:

$$\frac{d\mathbf{r}_j}{dt} = \mathbf{u}(t, \mathbf{r}_j) \ , \qquad \mathbf{r}_j(0) = \mathbf{r}_j^0, \tag{1}$$

$j = 1, \ldots, M$. Assume that trajectories of the first $p = M-1$ particles $\mathbf{r}_1(t), \mathbf{r}_2(t), \ldots, \mathbf{r}_p(t)$ are completely observed during time interval $(0, T)$, while the trajectory of the last one, $\mathbf{r}_M(t)$, is not observed. The problem is to find a reasonable prediction of the position of the unobserved particle, given the above predictor observations and the initial predictand position. The optimal prediction in the mean square sense,

$$E|\widehat{\mathbf{r}}_M(T) - \mathbf{r}_M(T)|^2 \to \min,$$

is given by the conditional expectation (e.g., Liptser and Shiryaev (1978))

$$\hat{\mathbf{r}}_M(T) = E\left(\mathbf{r}_j(T) \mid \mathbf{r}_1(t), \mathbf{r}_2(t), \ldots, \mathbf{r}_p(t), \quad 0 \le t \le T\right),$$

based on all the observations. Alas, this general formula gives too little in the considered situation. Normally, this conditional expectation cannot be explicitly found even in the simplest case when the velocity fluctuations are delta-correlated in time. However, it can be approximated (with an uncertain accuracy) for some Markov models of stochastic flows (Piterbarg (2001b)). This approximation resulted in a Kalman filter–type prediction algorithm which was tested on synthetic (Özgökmen et al. (2000), Piterbarg (2001b)) and real (Castellari et al. (2001), Özgökmen et al. (2001)) data. In general, that algorithm, called henceforth the KF algorithm, performs well, but its essential drawback is that it requires knowledge of some statistics of the underlying stochastic flow such as the Lagrangian correlation time and the space correlation radius of the Eulerian velocity field.

The goal of this paper is to introduce and investigate a new model-independent prediction algorithm. At first glance, the suggested prediction method looks a little bit naive. Roughly, we linearize (1) in a vicinity of the initial cluster, obtain a linear regression model where regressors are the initial particle positions, then estimate the regression coefficients at any given moment based on the observations of the predictor positions, and, finally, use them for predicting the unknown particle. The idea for the new algorithm emerged when we found from real data that the position of the cluster center of mass is a not bad alternative to the KF algorithm. The trouble is that the center-of-mass algorithm (CM) performs poorly at the initial stage if the predictor is located far from the cluster center of mass. In fact, the suggested algorithm, called here the regression algorithm (RA), can be viewed as a CM algorithm adjusted to the initial position of the predictand. As it will be shown, the RA performs very well at the initial stage if the cluster diameter is essentially less than the space correlation radius of the velocity fluctuations and performs as well as the CM algorithm in the long term. The good predictive skill of RA demonstrated in real data processing has had an impact on development of theoretical and Monte Carlo error analysis for RA. Such an analysis is based on investigating the second moment $\rho(t, \mathbf{r}_0)$ of the difference between positions of two particles initially separated by $\mathbf{r}_0$, called the separation process. The quantity $\rho(t, \mathbf{r}_0)$ can be effectively studied for two important models: the well-known Brownian flow and a stochastic model with memory recently developed in (Piterbarg (2001a)).

The Brownian stochastic flow arises when the Eulerian velocity field is delta-correlated in time. In this case a closed partial differential equation for $\rho(t, \mathbf{r}_0)$ is

readily written and the asymptotical prediction error is obtained for a tight initial cluster by expanding the equation solution in $\mathbf{r}_0$. It is interesting that the prediction error is not determined by the second Lyapunov moment, but rather by the fourth order term in $\mathbf{r}_0$. Asymptotical behavior of the error for large t is determined by the top Lyapunov exponent of the stochastic flow. In the case of a positive Lyapunov exponent, the behavior of the mean square error is close to that of the dispersion, $\sqrt{2Dt}$, where D is an effective diffusivity. In fact, it is even a little bit worse due to the bias between the predictand initial position and the cluster center of mass. For a negative Lyapunov exponent the growth order is $\sqrt{t/\log t}$. It is worthwhile to notice that our algorithm is not designed for long-term prediction.

Being an important mathematical example, the Brownian stochastic flow is not a realistic model for upper ocean turbulence. We focus instead on the second model in which a joint vector of the positions and velocities form a Markov process. For this reason it is called the first-order Markov model to distinguish it from the zero-order model determined by the Brownian flow. The first-order model implies an additional important parameter, the Lagrangian correlation time τ, which can be estimated well from real data (Griffa et al. (1995)). In the framework of the model, the Lagrangian velocity of a single particle is an Ornstein–Uhlenbeck process with parameter τ. In this case the variance of the separation process $\rho(t, \mathbf{r}_0, \mathbf{v}_0)$ also depends on the difference of the initial velocities. As a consequence, the prediction error for close initial positions and velocities is determined by the expansion coefficient of ρ at $\mathbf{v}_0^2$. The equation for ρ in the first-order model is a standard Kolmogorov equation. An expansion of the solution in both $\mathbf{r}_0$ and $\mathbf{v}_0$ results in an approximate mean square error for the prediction. The proposed approximation is in good agreement with simulations. A special focus is on a linear shear mean flow determined by the stretch and rotation parameters γ and ω, respectively. It is shown that the relative prediction error decreases as γ and ω increase. Comparison with the KF algorithm shows that the RA performs essentially better in the presence of a deterministic linear shear flow, while for a pure stochastic flow they are equivalent or KF is better.

The main points of this work are (1) formulation of the new prediction algorithm (section 2); (2) formulas for the prediction error which are in very good agreement with stochastic simulations (sections 3–5); (3) comparative analysis of RA and KF performance based on synthetic and real data (sections 6 and 7).

The main investigation tools used are stochastic simulations, together with standard diffusion process analytic techniques. For the simulations we take real values of model parameters and show the error in dimension units to give an idea of the usefulness of the real prediction.

2. Prediction formula. Assume the following classical regression model for motion of M particles:

$$\mathbf{r}_i(t) = \mathbf{A}(t)\mathbf{r}_i(0) + \mathbf{b}(t) + \mathbf{y}_i(t), \tag{2}$$

where $\mathbf{A}(t)$ and $\mathbf{b}(t)$ are an unknown, random in general, 2×2 matrix and 2-vector, respectively, and $\mathbf{y}_j(t)$ are stochastic processes with zero mean uncorrelated for any fixed t. Notice that this model does not follow from the model (1) in the general case. Moreover, it even contradicts (1) for a nonlinear velocity field. The idea is to construct a prediction algorithm based on (2) and then forget (2) and investigate the algorithm performance for some important particular cases of the model (1). The reason to expect a good performance is that the system (1) can be linearized on short times, and then the obtained formula would be useful for the short-term prediction.

Recall that the first $p = M - 1$ particles (predictors) are supposed to be observed and the Mth one is to be predicted. To identify six unknown parameters at each time (four entries of $\mathbf{A}$ and two entries of $\mathbf{b}$) one should have $p \geq 3$. We accept this assumption for the rest of the paper. The underdetermined situation $p = 2$ is of practical interest as well, but it requires a special consideration which is outside the scope of this paper. The least square estimators of $\mathbf{A}(t)$ and $\mathbf{b}(t)$ based on the observed particles at the moment t are given by

$$\hat{\mathbf{A}}(t) = \mathbf{S}(t)\mathbf{S}(0)^{-1}, \quad \hat{\mathbf{b}}(t) = \mathbf{r}_c(t) - \hat{\mathbf{A}}(t)\mathbf{r}_c(0),$$

where

$$\mathbf{r}_c(t) = \frac{1}{p}\sum_{i=1}^{p} \mathbf{r}_i(t)$$

is the center of mass of the predictor cluster and

$$\mathbf{S}(t) = \sum_{i=1}^{p} (\mathbf{r}_i(t) - \mathbf{r}_c(t))(\mathbf{r}_i(0) - \mathbf{r}_c(0))^T,$$

the superscript T stands for transposition, the vectors mean column-vectors, and it is assumed that $p > 2$ to have nondegenerate matrix $S(0)$. The obtained estimators then are used to predict the unobservable particle

$$\hat{\mathbf{r}}_M(t) = \mathbf{r}_c(t) + \mathbf{S}(t)\mathbf{S}(0)^{-1}(\mathbf{r}_M(0) - \mathbf{r}_c(0)). \tag{3}$$

This prediction formula is optimal in the framework of the model (2) if $\mathbf{A}$ and $\mathbf{b}$ are supposed to be deterministic. Further we reject the regression model and study its performance for some specific models of the velocity field $\mathbf{u}(t, \mathbf{r})$ appearing in (1).

If the velocity field is smooth enough in time, then it is worthwhile to include the initial velocities as regressors as well. We do not do that in the present paper for two reasons: first, this does not make any sense when considering the Brownian flow since it implies infinite velocities, and second, determining initial velocities in practice is a very hard problem. However, a study of an initial velocity–based formula is of theoretical interest and will be considered in a further work.

Once again we underscore that the prediction formula (3) does not include any parameters except the initial particle positions. Of course, it is not always a strength. Including well-known parameters would probably improve prediction essentially, but the problem is that statistical estimates of mean currents and turbulence parameters are often not reliable in oceanic conditions. Therefore, apparently, sometimes it is better to use a rough prediction algorithm than fine algorithms with misspecified parameters. Further we try to evaluate limits of this "roughness."

3. General error analysis. Let

$$s^2(t) = E|\hat{\mathbf{r}}_M(t) - \mathbf{r}_M(t)|^2 \tag{4}$$

be the mean square error of the prediction (3). Introduce the variance of the separation process by

$$\rho_{ij}(t) = E|\mathbf{r}_i(t) - \mathbf{r}_j(t)|^2.$$

For the sake of brevity we call ρ_{ij} the separation. Then (see appendix)

(5)

$$s^2(t) = \frac{1}{p}\sum_{k=1}^{p}\rho_{kM}(t) - \frac{1}{2p^2}\sum_{k,l=1}^{p}\rho_{kl}(t) - \frac{1}{p}\sum_{k,l=1}^{p} b_k\rho_{kl}(t) + \sum_{k=1}^{p} b_k\rho_{kM}(t) - \frac{1}{2}\sum_{k,l=1}^{p} b_k b_l\rho_{kl}(t),$$

where coefficients

$$b_i = (\mathbf{r}_i(0) - \mathbf{r}_c(0))^T \mathbf{S}(0)^{-1}(\mathbf{r}_M(0) - \mathbf{r}_c(0)) \tag{6}$$

are defined by the initial positions only, while the behavior of $\rho_{ij}(t)$ also depends on the flow properties. Notice that the deviation $E|\mathbf{r}_c(t) - \mathbf{r}_M(t)|^2$ of the predictand from the cluster center of mass is determined by the first two terms in (5).

In the next two sections we will consider two different asymptotics of (4): first, small initial distances between particles and, second, the long time asymptotic. As one will see, in the first case the problem is reduced to a factorized separation

$$\rho_{ij}(t) \sim \rho_0(t)c_{ij},$$

where ρ_0 is independent on the initial configuration and c_{ij} are completely determined by the initial conditions. In the second case under common conditions, the separation is independent of the initial conditions:

$$\rho_{ij}(t) \sim \rho(t)(1 - \delta_{ij}),$$

where δ_{ij} is the Kronecker delta. Hence, in the first case (5) becomes

$$s^2(t) \sim C_0\rho_0(t), \tag{7}$$

where

$$C_0 = \frac{1}{p}\sum_{k=1}^{p} c_{kM} - \frac{1}{2p^2}\sum_{k,l=1}^{p} c_{kl} - \frac{1}{p}\sum_{k,l=1}^{p} b_k c_{kl} + \sum_{k=1}^{p} b_k c_{kM} - \frac{1}{2}\sum_{k,l=1}^{p} b_k b_l c_{kl} \tag{8}$$

is a function of the initial conditions only. A similar formula appears in the second case:

$$s^2(t) \sim C\rho(t), \tag{9}$$

with

$$C = \frac{1}{2} + \frac{1}{2p} + \frac{1}{2}(\mathbf{r}_M(0) - \mathbf{r}_c(0))^T \mathbf{S}(0)^{-1}(\mathbf{r}_M(0 - \mathbf{r}_c(0)). \tag{10}$$

Notice that this algorithm is not designed for long time prediction and the formula (9) is of theoretical interest only.

4. Brownian flow. Assume that the velocity field is decomposed into mean circulation and fluctuation:

$$\mathbf{u}(t, \mathbf{r}) = \mathbf{U}(t, \mathbf{r}) + \mathbf{u}'(t, \mathbf{r}), \tag{11}$$

where $\mathbf{U}_0(t,\mathbf{r})$ is a deterministic velocity field and $\mathbf{u}'(t,\mathbf{r})$ is a random vector field with zero mean, $E\mathbf{u}'(t,\mathbf{r})=0$. The Brownian flow is determined by the assumption that the velocity fluctuation $\mathbf{u}'(t,\mathbf{r})$ is a Gaussian white noise in t, i.e.,

$$E\mathbf{u}'(t,\mathbf{r})=0,\quad E\mathbf{u}'(t_1,\mathbf{r}_1)\mathbf{u}'(t_2,\mathbf{r}_2)^T=\delta(t_1-t_2)\mathbf{B}(\mathbf{r}_1,\mathbf{r}_2)\ , \tag{12}$$

where $\delta(t)$ is the Dirac delta-function and $\mathbf{B}(\mathbf{r}_1,\mathbf{r}_2)$ is the spatial covariance tensor of the velocity field. Introduce the state, drift, and noise vectors:

$$\mathbf{z}=\begin{pmatrix}\mathbf{r}_1\\ \mathbf{r}_2\\ \dots\\ \mathbf{r}_M\end{pmatrix},\quad \mathbf{A}(t,\mathbf{z})=\begin{pmatrix}\mathbf{U}(t,\mathbf{r}_1)\\ \mathbf{U}(t,\mathbf{r}_2)\\ \dots\\ \mathbf{U}(t,\mathbf{r}_M)\end{pmatrix},\quad \mathbf{W}(t)=\begin{pmatrix}\mathbf{w}_1\\ \mathbf{w}_2\\ \dots\\ \mathbf{w}_M\end{pmatrix},$$

where $\mathbf{r}_j(t)$ are the positions of M particles at time t starting from different locations and $\mathbf{w}_j(t)$ are independent standard Wiener processes. Then a rigorous interpretation of (11), (12) is as follows. The process $\mathbf{z}(t)$ is a $2M$-dimensional Markov process satisfying the stochastic Ito differential equation (Kunita (1990))

$$d\mathbf{z}=\mathbf{A}(t,\mathbf{z})dt+\mathbf{D}(\mathbf{z})^{1/2}d\mathbf{W}(t),$$

where the $2M\times 2M$ diffusion matrix is given by

$$\mathbf{D}(\mathbf{z})=(\mathbf{B}(\mathbf{r}_i,\mathbf{r}_j)).$$

Another equivalent formulation of this model is as follows: $\mathbf{z}(t)$ is a Markov process with the generator given by

$$L=\mathbf{U}(t,\mathbf{r}_i)\cdot\nabla_{\mathbf{r}_i}+\frac{1}{2}\nabla_{\mathbf{r}_i}\cdot\mathbf{B}(\mathbf{r}_i,\mathbf{r}_j)\nabla_{\mathbf{r}_j}.$$

In the homogeneous case characterized by the assumptions that the mean flow is constant $\mathbf{U}(t,\mathbf{r})\equiv\mathbf{U}$ and that the covariance is a function of the position difference $\mathbf{B}(\mathbf{r}_1,\mathbf{r}_2)=\mathbf{B}(\mathbf{r}_1-\mathbf{r}_2)$, the separation process $\mathbf{r}(t)=\mathbf{r}_1(t)-\mathbf{r}_2(t)$ (the difference between displacements of two different particles) is also a Markov process with the generator

$$L_s=\nabla_{\mathbf{r}}\cdot(\mathbf{B}(\mathbf{0})-\mathbf{B}(\mathbf{r})))\nabla_{\mathbf{r}}.$$

Further we assume that the velocity field is isotropic (that is, $\mathbf{U}\equiv 0$) and the entries $b_{ij}(\mathbf{r})$ of $\mathbf{B}(\mathbf{r})$ are given by (Monin and Yaglom (1975))

$$b_{ij}(\mathbf{r})=b_N(r)\delta_{ij}+\frac{x_ix_j}{r^2}(b_L(r)-b_N(r)),$$

where $r=|\mathbf{r}|$, $\mathbf{r}=(x_1,x_2)$. Assume that the longitudinal and normal covariances are four times continuously differentiable:

$$\begin{aligned}b_L(r)&=D-\tfrac{1}{2}\beta_Lr^2+\tfrac{1}{2}\gamma_Lr^4+O(r^6),\\ b_N(r)&=D-\tfrac{1}{2}\beta_Nr^2+\tfrac{1}{2}\gamma_Nr^4+O(r^6),\end{aligned} \tag{13}$$

where $D,\beta_L,\gamma_L,\beta_N,\gamma_N$ are positive parameters whose physical meaning is explained below. For the isotropic Brownian flow the squared dispersion is expressed as

$$d^2(t)\equiv E(\mathbf{r}(t)-\mathbf{r}(0))^2=2Dt; \tag{14}$$

hence, D means a diffusivity. Introduce the space correlation radius of the velocity field by

$$R^2 = D/\beta_L$$

and set $\mathbf{r}_{ij} = \mathbf{r}_i(0) - \mathbf{r}_j(0)$, $r_{ij} = |\mathbf{r}_{ij}|$. First, consider the "tight cluster" asymptotic characterized by

$$r_{ij} \ll R, \quad i, j = 1, M.$$

Under this approximation each separation $\rho_{ij}(t) = \rho(t, r_{ij})$ is expressible in form

$$\rho(t, r) = \rho_1(t) r^2 - \rho_0(t) r^4 + O(r^6), \tag{15}$$

where (see appendix)

$$\rho_1(t) = \exp(\overline{\beta} t), \quad \rho_0(t) = K(\exp(\beta_0 t) - \exp(\overline{\beta} t)), \tag{16}$$

where

$$\beta_0 = 2\beta_N + 6\beta_L, \quad \overline{\beta} = \beta_L + \beta_N, \quad K = \frac{\overline{\gamma}}{5\beta_L + \beta_N}, \quad \overline{\gamma} = \gamma_L + \gamma_N.$$

After substitution of the expansion (15) into (5) the terms quadratic in r_{ij} disappear, and we get

$$s^2(t) \sim C_0 \rho_0(t), \tag{17}$$

where C_0 is given by (8) with

$$c_{kl} = |\mathbf{r}_k(0) - \mathbf{r}_l(0)|^4$$

and $\rho_0(t)$ is given in (16)

For the incompressible Brownian flow, characterized by $b_N(r) = (d/dr)(r b_L(r))$, with the longitudinal covariance

$$b_L(r) = D \exp(-r^2/R^2),$$

we have

$$\rho_1(t) = \exp(8Dt/R^2), \quad \rho_0(t) = \frac{3}{8R^2}(\exp(24Dt/R^2) - \exp(8Dt/R^2)).$$

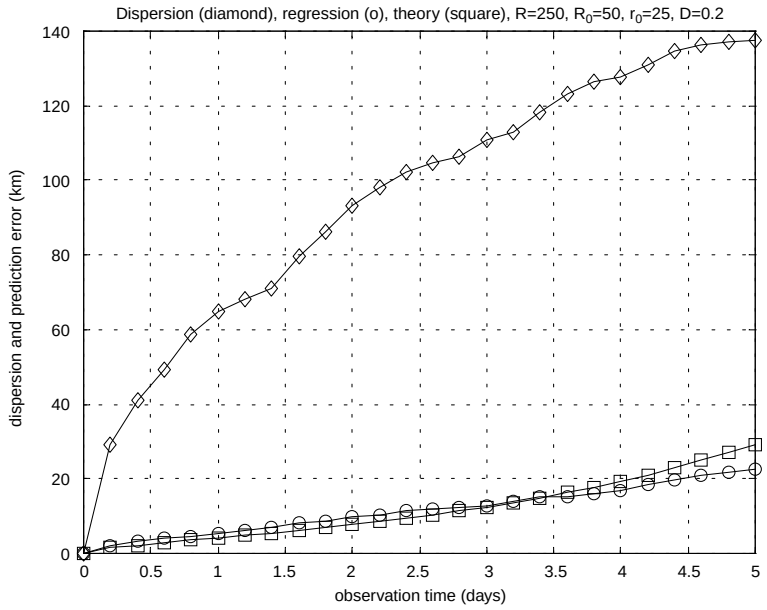
Assume that initially the predictors are located at the vertices of a right polygon at a distance R_0 from the center and the predictand is at a distance r_0 from the center. We call such an initial configuration perfect. In this case (17) becomes

$$s^2(t) \sim (3r_0^4 + 2R_0^4 - 4r_0^2 R_0^2)\rho_0(t) \tag{18}$$

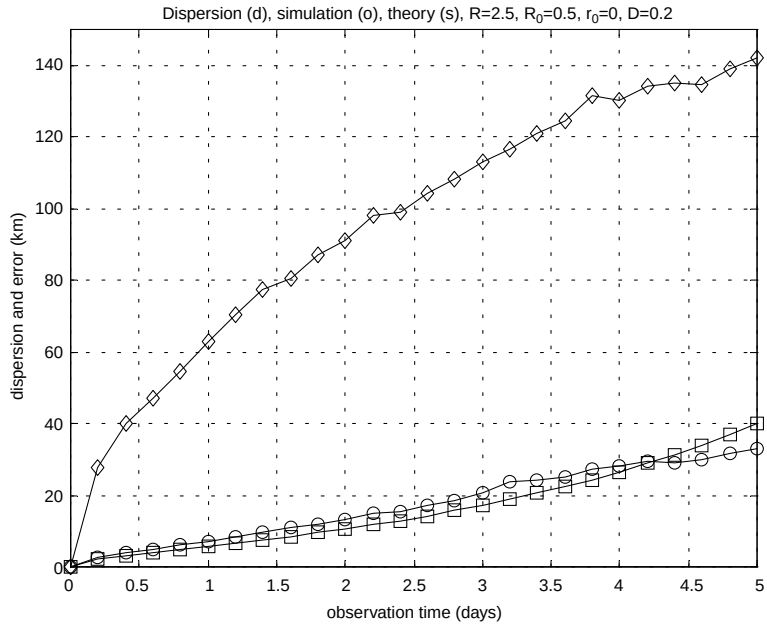
for $p > 3$ and

$$s^2(t) \sim (3r_0^4 + 2R_0^4 - 3r_0^2 R_0^2 - 2R_0 r_0^3 \cos(3\alpha))\rho_0(t)$$

for $p = 3$, where α is the angle between the directions from the center to the predictand and from the center to one of the predictors. The approximation (18) was checked via simulations, and the results are presented in Figure 1(a) and (b). For the simulation



(a)



(b)

FIG. 1. (a) *Dependence of the dispersion,* $d(t)$ (*diamonds*) *and prediction error,* $s(t)$, *obtained from simulations* (*circles*) *and from theory* ((18), *squares*) *on the observation time for the Brownian stochastic flow.* 100 *independent runs are used for* $d(t)$ *and* $s(t)$. *The diffusivity* $D = 2 \times 10^3\, km^2/day$, *number of predictors* $p = 6$, *initial hexagon radius* $R_0 = 50\, km$, *velocity space correlation radius* $R = 250\, km$, *distance of the predictand from the hexagon center* $r_0 = 25$ *km.* (b) *Same as in* (a) *with* $r_0 = 0$.

it was assumed that $R = 250$ km, $R_0 = 50$ km, and $D = 2000$ km^2/day. In the first case the predictor is distanced by $r_0 = 25$ km from the center, and in the second case it is initially located at the center. The dispersion $d(t)$ (diamonds) and experimental prediction error $s(t)$ (circles) are obtained from the simulations by averaging over 100 independent runs. A modest sample size is used on purpose to illustrate graphically how large the noise is under a moderate sample volume typical in oceanographic measurements. The line marked by squares expresses the suggested error formula (18). This approximation performs pretty well up to $T = 5$ days. After that the theoretical curve sharply diverges from the experimental one. The simulated dispersion is in good agreement with formula (14): it behaves as $\sqrt{t}$ and the value $d(5) \approx 139$ is very close to that of given by (14), $d(5) = 100\sqrt{2}$. In the second case (Figure 1(b)) the prediction is slightly worse in full agreement with (18). Notice that the relative prediction error for small t is approximately constant:

$$s_r(t) \equiv \frac{s(t)}{d(t)} \sim \frac{\sqrt{3(3r_0^4 + 2R_0^4 - 4r_0^2R_0^2)}}{2R^2}.$$

As for large t, we have two different situations depending of the sign of the Lyapunov exponent for the underlying flow. The Lyapunov exponent, λ, characterizes the exponential divergence (convergence) of initially close particles. It can be expressed in terms of the flow parameters (Baxendale and Harris (1986)):

$$\lambda = (\beta_N - \beta_L)/2.$$

If $\lambda > 0$, then for large t the difference between the positions of two particles goes to infinity with probability 1, and the mean square distance between them grows as

$$\rho(t) \sim 4Dt.$$

From (9), (10) it follows that the relative error is also approximately constant,

$$s_r(t) \sim \sqrt{1 + \frac{1}{p} + (\mathbf{r}_M(0) - \mathbf{r}_c(0))^T \mathbf{S}(0)^{-1} (\mathbf{r}_M(0 - \mathbf{r}_c(0))},$$

and greater than one.

In the opposite case $\lambda < 0$, the picture is more sophisticated: the difference goes to zero with probability 1; however, the mean square distance still grows, but at a lower rate (see Zirbel and Cinlar (1996)):

$$\rho(t) \sim \frac{ct}{\log t}$$

with constant c depending on the initial distance. Thus, the relative error goes to zero slowly as t goes to infinity:

$$s_r(t) \sim C\sqrt{\frac{1}{\log t}}$$

with constant C depending on the initial cluster configuration.

5. Stochastic flow with memory. As we already noticed, the Brownian flow is not an appropriate model for the upper ocean turbulence, since it is based on the white noise assumption for Lagrangian velocity. In fact, numerous observations clearly

demonstrate that Lagrangian velocity is well approximated by the first-order Markov process (Thomson (1986), Griffa (1996)). The following model of multiparticle motion suggested by Piterbarg (2001a), (2001b) generalizes the above experimental fact.

In addition to the decomposition (11) assume that there is a deterministic acceleration $\mathbf{a}(\mathbf{v},\mathbf{r})$ depending in general on the particle velocity and position such that the motion equations take the form

$$\begin{aligned} d\mathbf{r} &= (\mathbf{U}(t,\mathbf{r})+\mathbf{v})dt, \\ d\mathbf{v} &= \mathbf{a}(\mathbf{v},\mathbf{r})dt + d\mathbf{w}(t,\mathbf{r}), \end{aligned} \tag{19}$$

where

$$E\mathbf{w}(t,\mathbf{r}) = 0, \qquad E\mathbf{w}(t_1,\mathbf{r}_1)\mathbf{w}(t_2,\mathbf{r}_2)^T = \min(t_1,t_2)\mathbf{B}(\mathbf{r}_1,\mathbf{r}_2)\ .$$

In other words, now the velocity field is not a white noise itself but rather is driven by a white noise with a space covariance structure determined by tensor $\mathbf{B}(\mathbf{r}_1,\mathbf{r}_2)$. A rigorous formulation of (19) is given in the above-mentioned references. For a cluster of M particles, introduce a state vector containing the particle velocities as well as positions and a drift vector:

$$\mathbf{z} = \begin{pmatrix} \mathbf{v}_1 \\ \mathbf{r}_1 \\ \mathbf{v}_2 \\ \mathbf{r}_2 \\ \cdots \\ \mathbf{v}_M \\ \mathbf{r}_M \end{pmatrix}, \qquad \mathbf{A}(t,\mathbf{z}) = \begin{pmatrix} \mathbf{a}_1 \\ \mathbf{U}_1+\mathbf{v}_1 \\ \mathbf{a}_2 \\ \mathbf{U}_2+\mathbf{v}_2 \\ \cdots \\ \mathbf{a}_M \\ \mathbf{U}_M+\mathbf{v}_M \end{pmatrix},$$

where $\mathbf{U}_m = \mathbf{U}(t,\mathbf{r}_m)$, $\mathbf{a}_m = \mathbf{a}(\mathbf{v}_m,\mathbf{r}_m)$. The model (19) implies that the motion of any M particles is a Markov process in $4M$ dimensions described by a stochastic differential equation. Namely,

$$d\mathbf{z} = \mathbf{A}(t,\mathbf{z})dt + \mathbf{D}(\mathbf{z})^{1/2}d\mathbf{W},$$

where $\mathbf{W}(t)$ is a standard Wiener process in $4M$ dimensions and the diffusion matrix $\mathbf{D}(\mathbf{z})$ is given by

$$\mathbf{D}(\mathbf{z}) = (\mathbf{D}_{ij}(\mathbf{z}))$$

with 4×4 blocks

$$\mathbf{D}_{ij} = \begin{pmatrix} \mathbf{B}(\mathbf{r}_i,\mathbf{r}_j) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

Recall that now $\mathbf{B}(\mathbf{r}_1,\mathbf{r}_2)$ is the covariance tensor of the forcing, not the Eulerian velocity field itself. The equivalent formulation is given by the generator

$$L = (\mathbf{U}(t,\mathbf{r}_i)+\mathbf{v}_i)\cdot\nabla_{\mathbf{r}_i} + \mathbf{a}_i\cdot\nabla_{\mathbf{v}_i} + \frac{1}{2}\nabla_{\mathbf{v}_i}\cdot\mathbf{B}(\mathbf{r}_i,\mathbf{r}_j)\nabla_{\mathbf{v}_j}.$$

For our purposes the following homogeneous case is most important:

$$\mathbf{a}(\mathbf{v},\mathbf{r}) = -\tau^{-1}\mathbf{v}, \quad \mathbf{U}(\mathbf{r}) = \mathbf{U}+\mathbf{G}\mathbf{r}, \quad \mathbf{B}(\mathbf{r}_1,\mathbf{r}_2) = \mathbf{B}(\mathbf{r}_1-\mathbf{r}_2),$$

$$\mathbf{G} = \begin{pmatrix} \gamma & \omega \\ -\omega & -\gamma \end{pmatrix},$$

where τ is the Lagrangian correlation time and the mean velocity field is a divergence-free linear shear flow characterized by constant drift $\mathbf{U}$, stretching parameter γ, and rotation parameter ω. The one-particle motion in this case is described by the well-known Langevin equation for the Lagrangian velocity and the standard motion equation for the displacement

$$\begin{aligned} d\mathbf{v} &= -\tau^{-1}\mathbf{v}dt + \sigma_v\tau^{-1/2}d\mathbf{w}(t), \\ d\mathbf{r} &= (\mathbf{U} + \mathbf{G}\mathbf{r} + \mathbf{v})dt\ , \end{aligned} \tag{20}$$

where $\mathbf{w}(t)$ is a two-dimensional Brownian motion and $\sigma_v^2 = E\mathbf{v}^2$ is the velocity variance. To obtain (20) we assumed that $\mathbf{B}(\mathbf{0}) = (\sigma_v^2/2\tau)\mathbf{I}$. In particular, for the dispersion we get (see appendix)

(21)
$$d^2(t) \equiv\ E(\mathbf{r}(t) - \mathbf{r}(0))^2 = d_0^2(t) + \sigma_v^2 \int_0^t \int_0^t \cosh(\sqrt{\gamma^2(2t - s_1 - s_2)^2 - \omega^2(s_1 - s_2)^2}) \exp(-|s_1 - s_2|/\tau) ds_1 ds_2,$$

where $d_0(t)$ is determined by the mean flow and initial position $\mathbf{r}_0$ only. An explicit expression is given in the appendix. Further we assume that $\mathbf{U} = 0$ and $\mathbf{r}_0 = 0$, which results in $d_0^2(t) = 0$. Notice two partial cases of (21): if $\omega = 0$, then

$$d^2(t) = \frac{\sigma_v^2\tau}{1 - \gamma^2\tau^2}(\gamma^{-1}\ \sinh(2\gamma t) - \tau\cosh(2\gamma t) - \tau + 2\tau\cosh(\gamma t)\exp(-t/\tau)); \tag{22}$$

if $\gamma = 0$, then

$$d^2(t) = \frac{\sigma_v^2\tau}{1 + \omega^2\tau^2}\left(t - \frac{\tau(1 - \omega^2\tau^2)}{1 + \omega^2\tau^2}\ (1 - \exp(-t/\tau)\cos(\omega t)) - \frac{2\omega\tau^2}{1 + \omega^2\tau^2}\exp(-t/\tau)\sin(\omega t)\right).$$

Finally, for the zero shear (Zambianchi and Griffa (1994))

$$d^2(t) = \sigma_v^2\tau\,(t - \tau\ (1 - \exp(-t/\tau)))\,. \tag{23}$$

The stochastic equations for the separation process,

$$\mathbf{r} = \mathbf{r}_1 - \mathbf{r}_2, \quad \mathbf{v} = \mathbf{v}_1 - \mathbf{v}_2,$$

take the form

$$\begin{aligned} d\mathbf{v} &= -\tau^{-1}\mathbf{v}dt + (2(\mathbf{B}(\mathbf{0}) - \mathbf{B}(\mathbf{r})))^{1/2}d\mathbf{w}(t), \\ d\mathbf{r} &= (\mathbf{G}\mathbf{r} + \mathbf{v})dt. \end{aligned}$$

In other words, the generator of the separation process is

$$L_s = (\mathbf{G}\mathbf{r} + \mathbf{v}) \cdot \nabla_{\mathbf{r}} - \tau^{-1}\mathbf{v} \cdot \nabla_{\mathbf{v}} + \nabla_{\mathbf{v}} \cdot (\mathbf{B}(\mathbf{0}) - \mathbf{B}(\mathbf{r})))\nabla_{\mathbf{v}}. \tag{24}$$

Assume that the forcing is isotropic, i.e.,

$$b_{ij}(\mathbf{r}) = b_N(r)\delta_{ij} + \frac{x_i x_j}{r^2}(b_L(r) - b_N(r)),$$

with twice differentiable covariances

$$b_L(r) = \frac{\sigma_v^2}{2\tau} - \frac{1}{2}\beta_L r^2 + O(r^4), \quad b_N(r) = \frac{\sigma_v^2}{2\tau} - \frac{1}{2}\beta_N r^2 + O(r^4). \tag{25}$$

Let $\rho = \rho(t,u,v,x,y) = E\mathbf{r}(t)^2$, where $\mathbf{r}(0) = (x,y), \mathbf{v}(0) = (u,v)$. For small x, y, u, v expand

$$\rho = a_1x^2 + 2a_2xy + a_3y^2 + a_4u^2 + 2a_5uv + a_6v^2 + a_7xu + a_8xv + a_9yu + a_{10}yv \tag{26}$$

and set $\mathbf{a} = (a_1, a_2, \ldots, a_{10})$. It is shown in the appendix that

$$\frac{d\mathbf{a}}{dt} = \mathbf{A}\mathbf{a}, \quad \mathbf{a}|_{t=0} = \mathbf{a}_0, \tag{27}$$

where matrix $\mathbf{A}$ and vector $\mathbf{a}_0$ are given. In particular, for the zero shear

$$\rho(t) = \rho(t, \mathbf{r}, \mathbf{v}) = \rho_1(t)\mathbf{r}^2 + \rho_{10}(t)(\mathbf{r}\cdot\mathbf{v}) + \rho_0(t)\mathbf{v}^2, \tag{28}$$

where

$$\begin{aligned} \frac{d\rho_0(t)}{dt} &= -2\tau^{-1}\rho_0(t) + \rho_{10}(t), \\ \frac{d\rho_1(t)}{dt} &= 2\overline{\beta}\rho_0(t), \\ \frac{d\rho_{10}(t)}{dt} &= -\tau^{-1}\rho_{10}(t) + 2\rho_1(t), \end{aligned} \tag{29}$$

where $\overline{\beta} = \beta_L + \beta_N$. Assume that the initial velocities are Gaussian random values with zero mean and independent components with the same variance σ_0^2. Additionally assume that the velocities are independent for different particles and are independent of the forcing. Averaging (28) over the ensemble of initial values gives

$$\rho_{ij}(t) = \rho(t, \mathbf{r}_{ij}, \mathbf{v}_{ij}) = \rho_1(t)r_{ij}^2 + 2\rho_0(t)\sigma_0^2(1-\delta_{ij}). \tag{30}$$

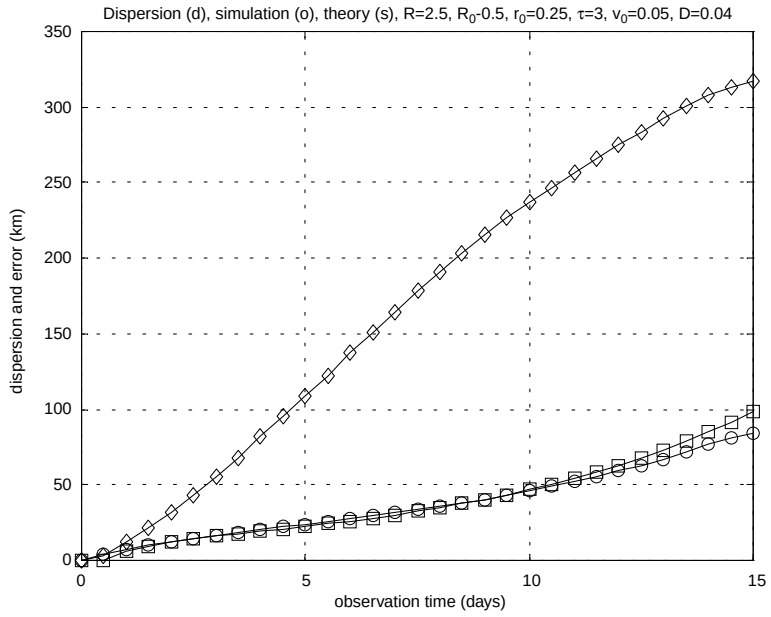
After substituting (30) into (5), the terms containing the distances disappear, and for the "tight cluster" asymptotic in the case of the perfect predictor, we get the initial configuration

$$s^2(t) \sim 2\sigma_0^2\left(1 + \frac{1}{p} + \frac{4r_0^2}{pR_0^2}\right)\rho_0(t), \tag{31}$$

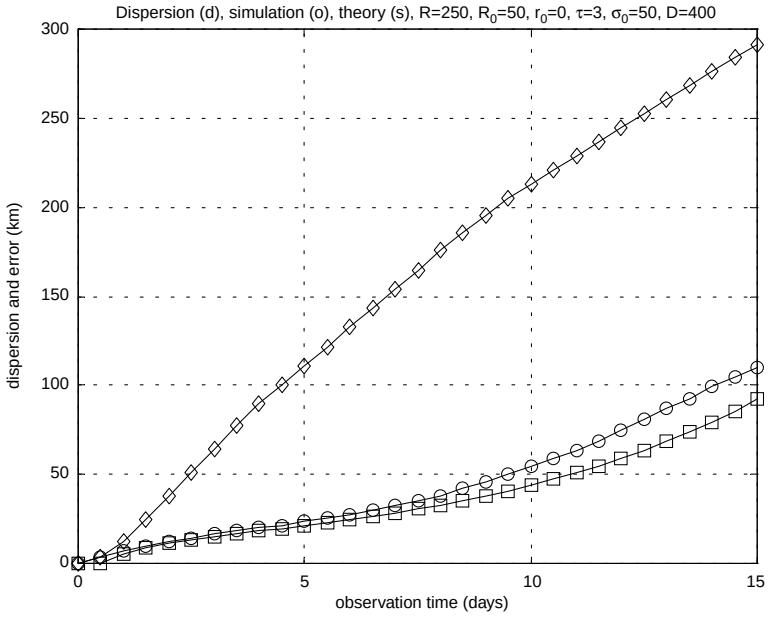
where $\rho_0(t)$ is obtained from (29). In the presence of the mean shear flow we get a similar formula:

$$s^2(t) \sim \sigma_0^2\left(1 + \frac{1}{p} + \frac{4r_0^2}{pR_0^2}\right)(a_4(t) + a_6(t)), \tag{32}$$

where $a_4(t), a_6(t)$ are obtained from (27). The asymptotic (31) is compared with simulations in Figure 2(a) and (b). For the simulations we used Lagrangian correlation time $\tau = 3$ days, the velocity variance $\sigma_v^2 = 0.12 \times 10^4$ km^2/day^2, initial velocity variance $\sigma_0^2 = 0.25 \times 10^2$ km^2/day^2, number of predictors $p = 6$, initial hexagon radius $R_0 = 50$ km, velocity space correlation radius $R = 250$ km. In Figure 2(a) the



(a)



(b)

FIG. 2. (a) *Dependence of the dispersion,* $d(t)$ (*diamonds*) *and prediction error,* $s(t)$*, obtained from simulations* (*circles*) *and from theory* ((31), *squares*) *on the observation time, for the stochastic flow with memory. The Lagrangian correlation time* $\tau = 3\,days$*, the velocity variance* $\sigma_v^2 = 12 \times 10^2\,km^2/day^2$*, initial velocity variance* $v_0^2 = 0.25 \times 10^2\,km^2/day^2$*, number of predictors* $p = 6$*, initial hexagon radius* $R_0 = 50\,km$*, velocity space correlation radius* $R = 250\,km$*, distance of the predictand from the hexagon center* $r_0 = 25$ *km.* (b) *Same as in* (a) *with* $r_0 = 0$.

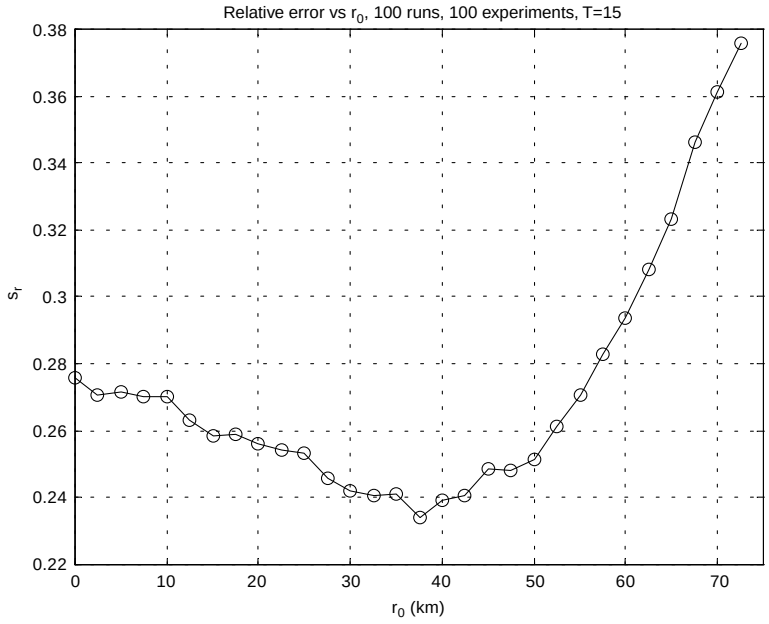


FIG. 3. *Dependence of the relative error s_r on r_0 for the observation time $T = 15$ days. The remaining parameters are the same as those as in Figure* 2(a).

predictand is distanced from the hexagon center by $r_0 = 25$ km, while in Figure 2(b) it is located exactly at the center. We take σ_0 much less than σ_v because the approximation (31) requires both small initial distances and small initial velocity differences. The dispersion $d(t)$ is shown as diamonds; the prediction error, $s(t)$, obtained from simulations, as circles; and the prediction error asymptotic (31) as squares. First, notice a good qualitative agreement of the simulated dispersion (100 runs) with the theoretical formula (23): for small t we have the ballistic regime ($d \sim t$), and for larger time the diffusion regime ($d \sim \sqrt{t}$). The quantitative agreement is also satisfactory. Then these figures show that the theoretical error formula works well for the ratio cluster radius/velocity correlation radius of less than 5 and for prediction periods of fewer than 15 days. The agreement is clearly better in the first case. In this regard, notice that unlike the Brownian flow case, the suggested approximation (31) does not give a correct dependence of the prediction error on the initial distance r_0 from the center. Indeed, in accordance with (31) the error in the second case should be less, but the simulations show the opposite. To get a correct dependence one should account for terms of higher order in the expansion of ρ. We do not do that here but instead study the dependence of the relative prediction error s_r on r_0 by the Monte Carlo method. Figure 3 demonstrates this dependence. The curve was obtained by averaging over 100 experiments with the same parameters $\tau = 3$ days, $\sigma_v^2 = 0.12 \times 10^4$ km^2/day^2, $\sigma_0^2 = 0.25 \times 10^2$ km^2/day^2, $p = 6$, $R_0 = 50$ km, $R = 250$ km, while each experiment included 100 runs to obtain s_r. This figure supports the previous observations that the error first decreases as r_0 increases, then assumes a minimum between 0 and R_0, and finally increases approaching R_0. For $r_0 > R_0$ the prediction worsens drastically. The obtained curve is affected by sampling variability, and the exact dependence s_r on r_0 is still to be investigated. It is interesting that the analytical dependence of s_r on the ratio $x = r_0^2/R_0^2$ obtained for the zero-order model

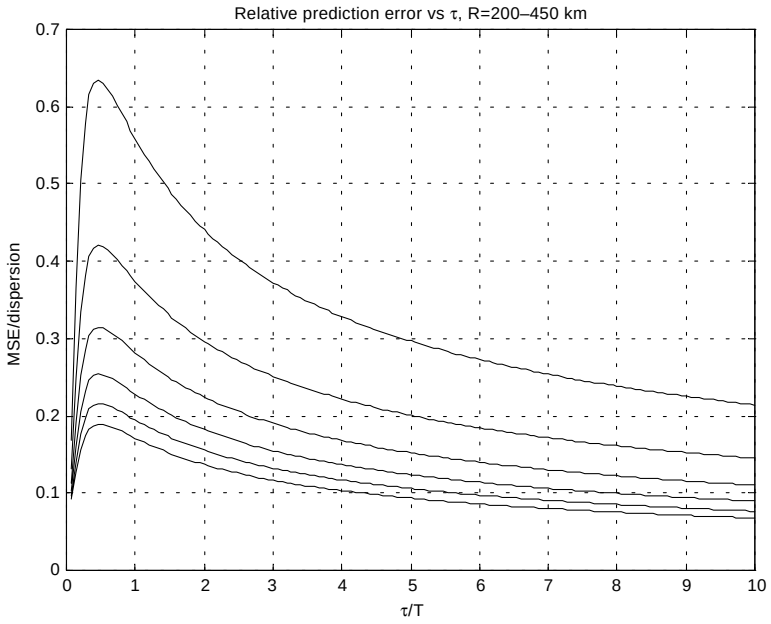
(19) describes pretty well the experimental curve for the first-order model. Indeed, from (19) $f(x) \equiv s(x)/s(0) = 1.5x^2 - 2x + 1$ assumes $f(0.5) = 0.375, f(2/3) = 1/3$ (minimum point), and $f(1) = 0.5$. Approximately the same values for that ratio follow from the curve in Figure 3. The reason is that for $T \gg \tau$ the first-order model approaches the Brownian flow.

Having sufficiently good agreement between (31) and the simulations, at least for values of r_0 close to $R_0/2$, we investigate the dependence of the prediction error on the model parameters using the analytical formulas (31) and (32). Figure 4(a) illustrates the dependence of the relative error $s_r(T)$ on τ for $R = 200, 250, 300, 350, 400, 450$ km with the zero mean flow and $T = 15$ days obtained from (21), (29), (31). The curves line up with R: the larger the R, the better the prediction. As for the Lagrangian correlation time, the error decreases with τ whenever $\tau/T > 0.5$. The effect of the error increasing for small value of τ is due to the regime changing in the dispersion behavior from ballistic to diffusive. As before, $R_0 = 50$ km and $p = 6$.

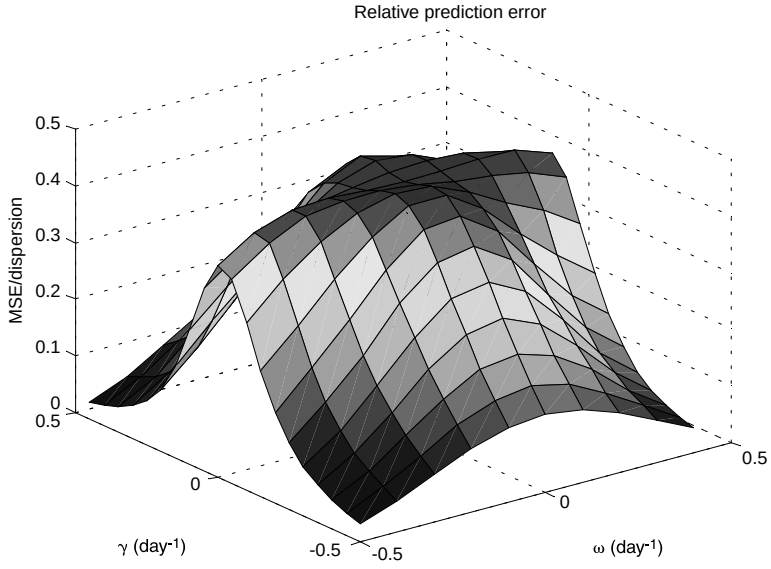
Figure 4(b) shows the dependence of s_r on γ and ω for fixed $\tau = 3$ days, $T = 15$ days, and $R = 250$ km, obtained from (21), (27), (32). The values of other parameters are indicated in the figure captions. Obviously, $s_r(\gamma, \omega)$ as a function of the shear parameters is even in both of them. The maximum relative error corresponds to $\gamma = 0$ because of a strong growth of the dispersion with γ (21), (22).

In the next series of experiments with zero mean flow we try a different initial configuration of predictors and a different velocity initialization to determine how the initialization affects the prediction skill. An eventual goal is to find an initial predictor configuration ensuring the best prediction. This problem is very complex and is beyond the scope of this paper. The goal of the present experiments with randomly distributed predictors is to understand the extent of the prediction error's dependence on the initial configuration and velocities. Namely, the alternative configuration we consider is a random initial configuration of predictors with the uniform distribution in a square with side of a. Now we compare four cases, the first two of which were discussed before: (1) perfect configuration ($R_0 = 50$ km) with the predictand $r_0 = 25$ km from the cluster center; (2) perfect configuration with the predictand at the center ($r_0 = 0$); (3) random configuration with $a = 2R_0$ and predictand $r_0 = 25$ km from the square center; (4) random configuration with $a = 2R_0$ and predictand at the square center ($r_0 = 0$). First, we consider the dependence of the error on the number of predictors for these four cases (Figure 5). As one can see there is not much difference in the algorithm performance when the number of predictors is 6 or more. It is worth noting that the random case is slightly worse than the perfect case for low p, but they quickly converge at $p = 6$ and do not change much as p increases. The error does not decrease significantly as p grows for all the initializations. This is in agreement with the theoretical formula (31).

Next we compare the statistical moments and histograms of the prediction error for 6 predictors and prediction time of 15 days (Table 1 and Figure 6(a)). The table gives the statistical moments of the relative deviation $\xi = |r_M - \widehat{r_M}|/d$, and the histograms are histograms of ξ. Thus, $s_r = \sqrt{E\xi^2}$, and it is also given in the table even though it can be found from the first two columns. First, it can be seen that for the perfect predictor configuration (series 1a, 2a) the initial location of the predictand is essential. Approaching the predictand to the predictors ($r_0 = 25$ km, series 1a) diminishes the mean from 0.275 to 0.249. This is in agreement with the previous experiments shown in Figures 2, 3, and 5. As for the random distribution of predictors, the initial position of the predictand almost does not make a difference,



(a)



(b)

FIG. 4. (a) *Dependence of the relative prediction error s_τ on the Lagrangian correlation time τ for different values of the Eulerian velocity correlation radius R obtained from the asymptotic formula* (31) *for observation time $T = 15$ days. Initial hexagon radius $R_0 = 50$ km, distance of the predictand from the hexagon center $r_0 = 25$ km, R varies from* 450 *km (lower curve) to* 200 *km (upper curve) with step* 50 *km.* (b) *Dependence of the relative prediction error s_τ on the shear parameters γ and ω obtained from the asymptotic formula* (32) *for observation time $T = 15$ days. Lagrangian correlation time $\tau = 3$ days, Eulerian velocity correlation radius $R = 250$ km, initial hexagon radius $R_0 = 50$ km, distance of the predictand from the hexagon center $r_0 = 25$ km.*

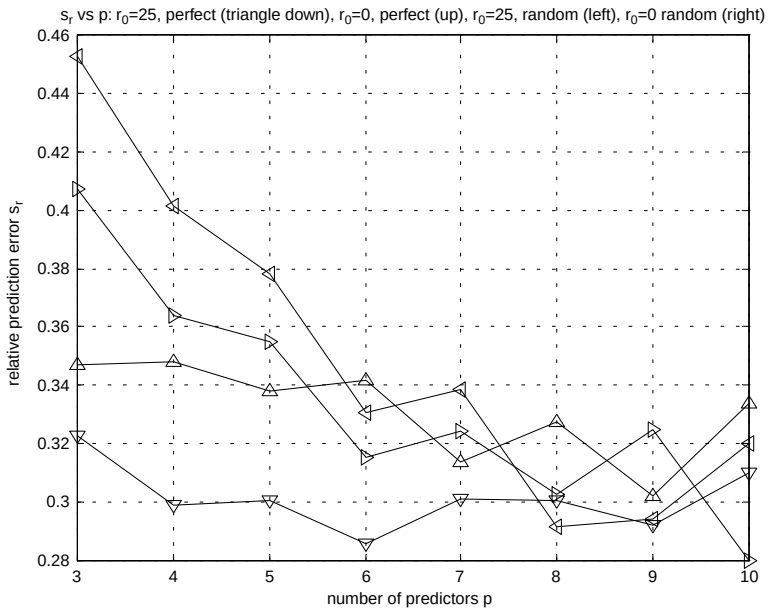


FIG. 5. *Dependence of the relative prediction error on the number of predictors (simulation) for different initial configurations.* (1) *Perfect configuration with the biased predictand (triangle down),* (2) *perfect configuration with the predictand at the center (triangle up),* (3) *uniformly distributed predictors with the biased predictand (triangle left),* (4) *uniformly distributed predictors with the predictand at the center (triangle right).*

TABLE 1

	Mean	STD	Median	Error
Series 1a	0.249	0.1947	0.203	0.3161
Series 2a	0.2751	0.2032	0.2294	0.342
Series 3a	0.2601	0.2088	0.2117	0.3335
Series 4a	0.2716	0.1981	0.2175	0.3362
Series 1b	0.2443	0.1923	0.198	0.3109
Series 2b	0.2792	0.2008	0.2329	0.3439
Series 3b	0.2525	0.2133	0.1998	0.3305
Series 4b	0.2601	0.202	0.2155	0.3293

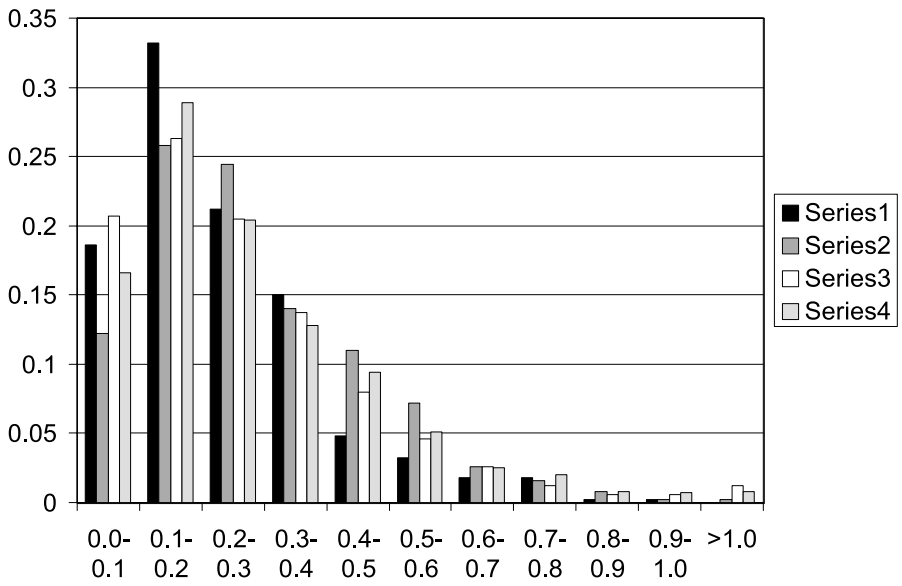
as can also be seen from the histogram (series 3 and 4). In contrast, the histograms in the case of perfect configuration (series 1 and series 2) look quite different. The mode of the first distribution is essentially higher, and the tail decays much faster.

Now introduce the alternative method of the velocity initialization as follows:

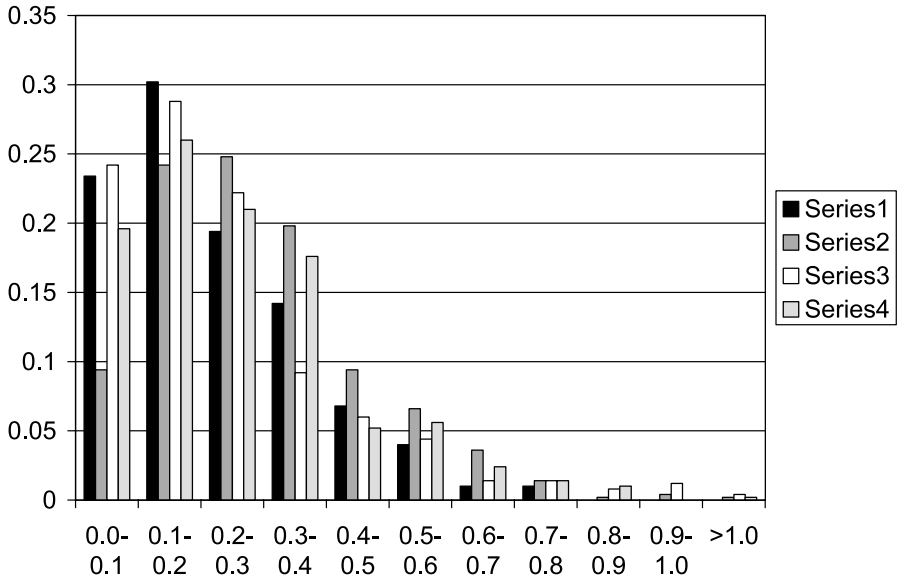
$$\mathbf{v}_j(0) = k\mathbf{r}_j(0), \tag{33}$$

where k is a constant independent of $j = 1, 2, \ldots, M$.

For the experiments we took $k = 0.1$ day^{-1} to have the same order of the initial velocities as in the case of the random initialization. Table 1 and Figure 6(b) demonstrate the statistical moments and histograms for the four cases discussed above, corresponding to the new velocity initialization (33). Regarding the predictand location, the conclusion is as before: for the perfect configuration of the predictors it is better to distance the predictor from the center, and for the random configuration it does not matter.



(a)



(b)

FIG. 6. (a) *Histograms of the relative prediction error for* 500 *runs with different initial configurations and random initial velocities:* (1) *perfect configuration with the biased predictand* (*series* 1), (2) *perfect configuration with the predictand at the center* (*series* 2), (3) *uniformly distributed predictors with the biased predictand* (*series* 3), (4) *uniformly distributed predictors with the predictand at the center* (*series* 4). (b) *Same as in* (a) *with the initial velocities proportional to the positions:* $\mathbf{v}_j(0) = k\mathbf{r}_j(0)$.

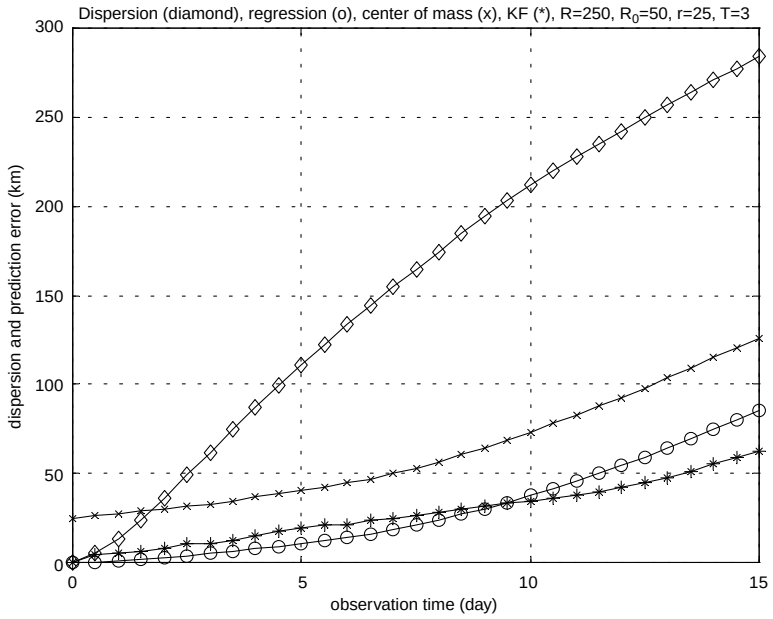
In general, there is not any visible difference in the prediction skill for the different velocity initializations. However, this conclusion is relevant only to 15 days' prediction. In practice we are more interested in a 3- to 5-day forecast, and in such time scales, the difference could be essential.

6. Comparison with KF and CM algorithms (simulations). The goal of the experiments discussed in this section is to compare the performance of RA and KF in both cases: the zero mean flow and a linear shear mean flow. The KF algorithm is based on the system of stochastic differential equations (19) for the M-particle motion. Since the diffusion matrix depends on the state variable $\mathbf{z}$, the classical Kalman filter cannot be applied to this system. What was proposed and studied in Özgökmen et al. (2000), (2001) and Piterbarg (2001b) is as follows. Pretend that the diffusion matrix is constant and write the KF equations for the optimal prediction of the unobserved particle velocity and position. Of course, these equations include the diffusion matrix. Then recall that it depends on the positions of all the particles and simply plug the observed positions for predictors and predictand forecast at the previous time step. We call this procedure a Kalman filter–type algorithm or, for short, the KF algorithm. An exact theoretical error analysis for the KF is very difficult. A Monte Carlo study showed that it gives a reasonable prediction if the model parameters are known (Piterbarg (2001b)). We follow the same approach here: the Lagrangian correlation time τ and the forcing covariance tensor $\mathbf{B}(\mathbf{r}_1, \mathbf{r}_2)$ are the same for generating Lagrangian trajectories and prediction formulas. However, the time steps are different: 1 hour for simulations and 12 hours for prediction. Thus, the KF has a big advantage over the RA, which does not use any information on the flow statistics.

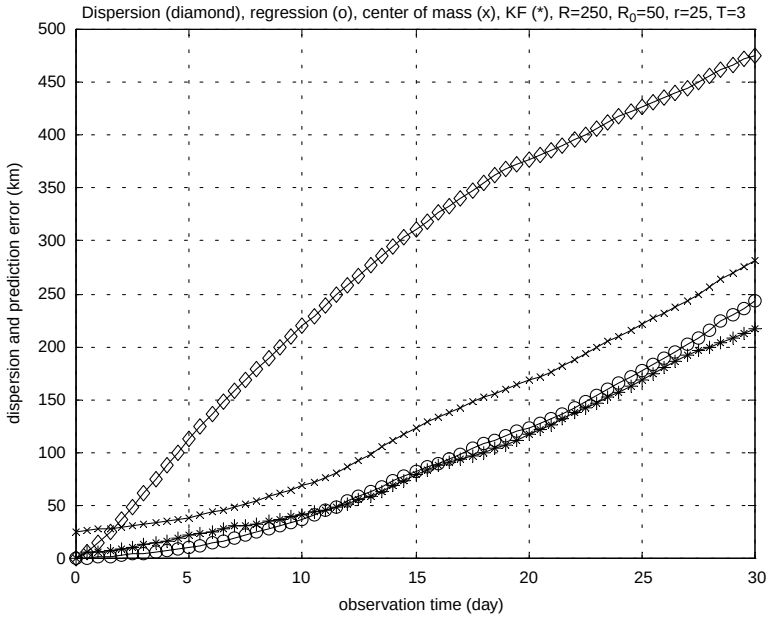
In the first series of experiments we considered the zero mean flow and fixed $\tau = 3$ days, $R = 250$ km, $R_0 = 50$ km, $M = 7$. Initially, the predictors are located in the vortices of the right hexagon, and the velocities are proportional to the position vectors; that is, the initialization (33) and the perfect configuration were used.

If the predictand is placed some distance from the center ($r_0 = 25$ km), then the mean square error of the regression algorithm is slightly lower than that of the KF (Figure 7(a)). Both algorithms are doing quite well compared with the dispersion (diamonds). This is because the initial cluster radius is 5 times less than the spatial correlation radius. The center of mass prediction (crosses) gives clearly worse prediction. After 22 days the performance of KF and regression is pretty much the same (Figure 7(b)).

If the predictand is placed at the center under the same experimental conditions, then the KF prediction turns out to be better (Figure 8(a)). As we mentioned before and which follows from the analytical formulas, the regression and center of mass methods give the same result in this case. For the midterm prediction (up to 30 days), this trend is confirmed (Figure 8(b)). For observation time $T = 30$ days the KF error is about 165 km, while the regression error is around 270 km under the dispersion 450 km. After changing τ to 2.5 days the general picture almost did not change (Figure 9(a)) and the conclusion is the same: KF performs better. However, if we introduce a mean flow, not very strong, the picture changes drastically. For a gyre given by $\omega = 0.1$ and $\gamma = 0$ the performance of KF is very poor (Figure 9(b)). The error reaches 160 km for a 15-day forecast, almost 80% of the dispersion (230 km), while the error of the regression algorithm is acceptable (60 km). Consider a different shear with no rotation: $\omega = 0$ and $\gamma = 0.05$. The conclusion is the same: the performance of the regression is clearly better (Figure 9(c)). The error of KF is

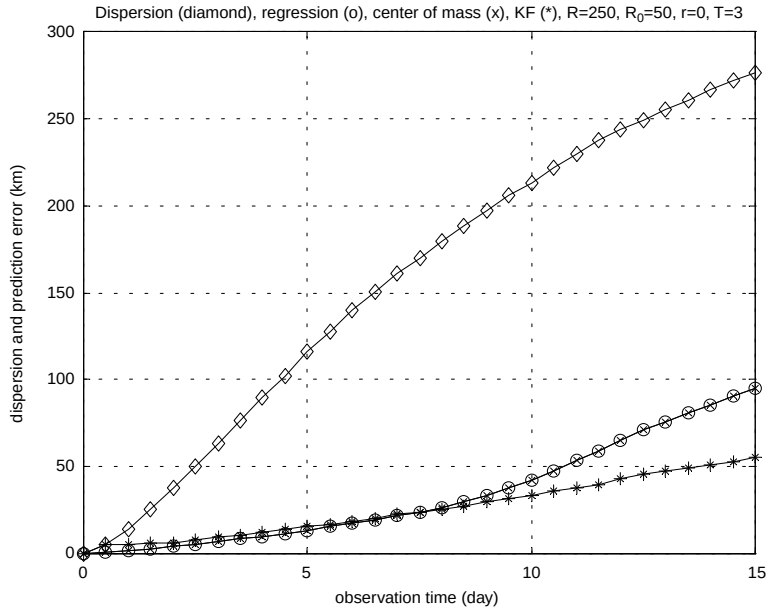


(a)

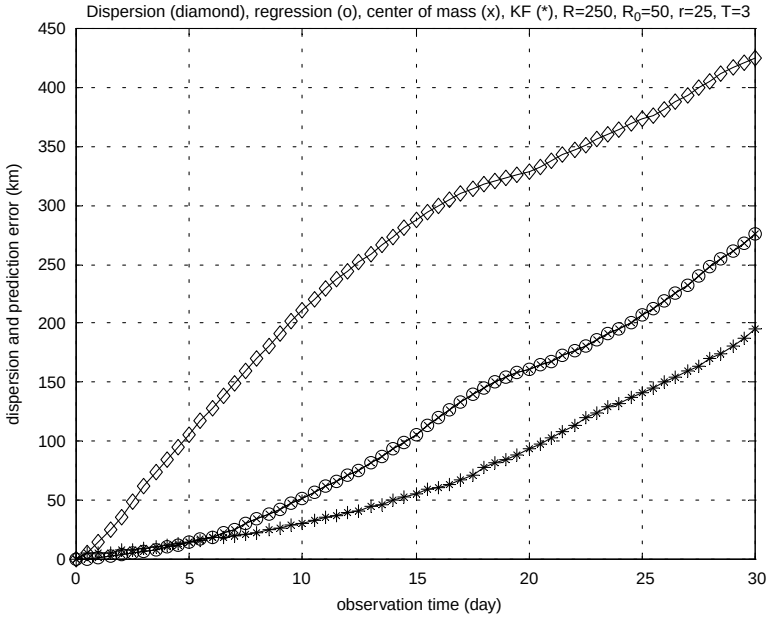


(b)

FIG. 7. (a) *Comparison of the dispersion (diamonds) and prediction error for the RA (circles), the CM (x), and the KF method for the maximum observation time* $T = 15$ *days and zero mean flow. The number of predictors* $p = 6$. *The predictors are located in vertices of a right hexagon. Lagrangian correlation time* $\tau = 3$ *days, Eulerian velocity correlation radius* $R = 250$ *km, initial hexagon radius* $R_0 = 50$ *km, distance of the predictand from the hexagon center* $r_0 = 25$ *km.* (b) *Same as in* (a) *for the maximum observation time* $T = 30$ *days. A different experiment.*

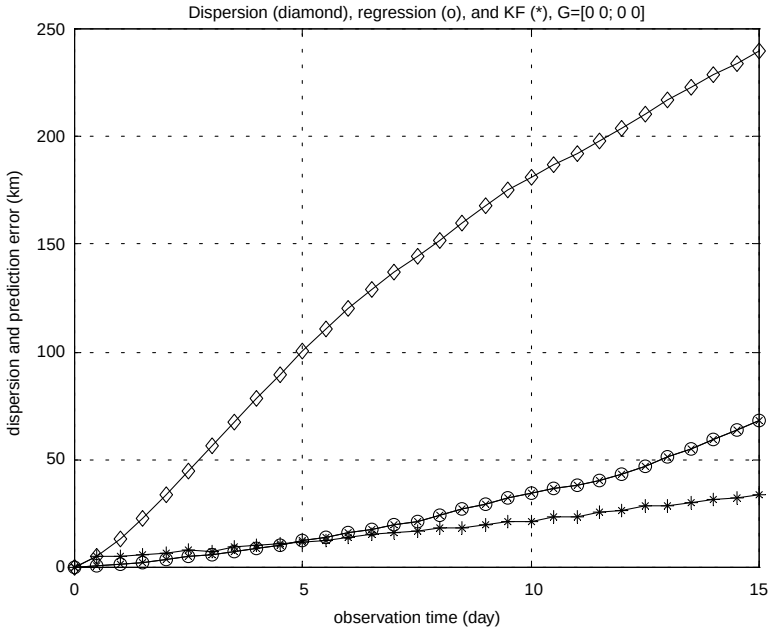


(a)

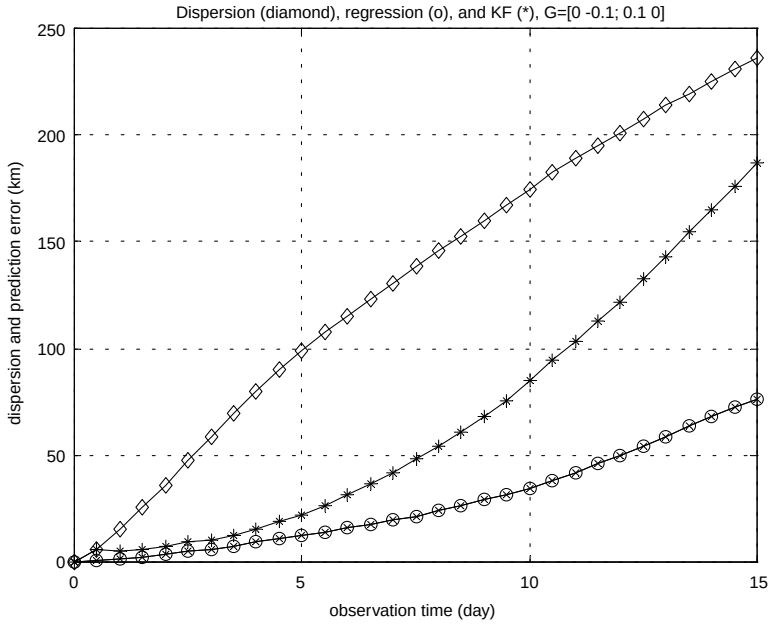


(b)

FIG. 8. (a) *Same as in Figure* 7(a) *with the predictor initially located at the center* ($r_0 = 0$). (b) *Same as in Figure* 8(a) *for the maximum observation time* $T = 30$ *days. A different experiment.*

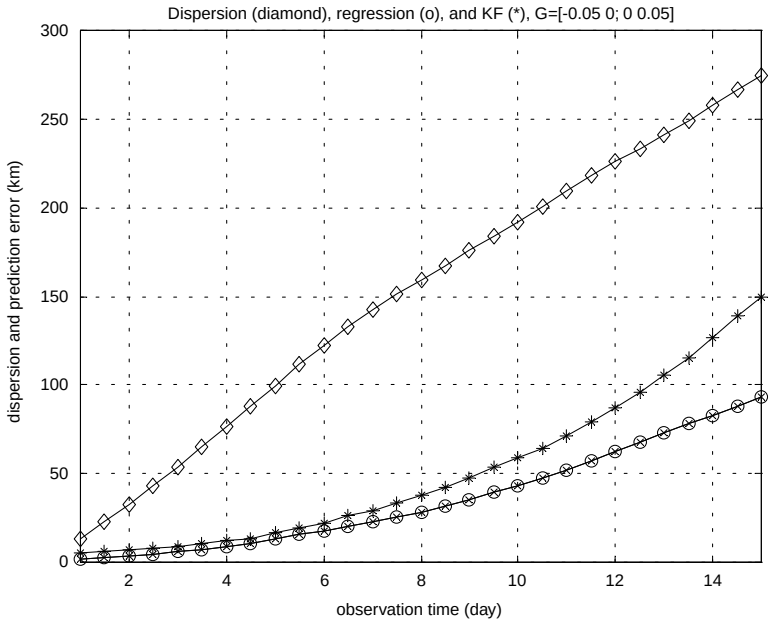


(a)

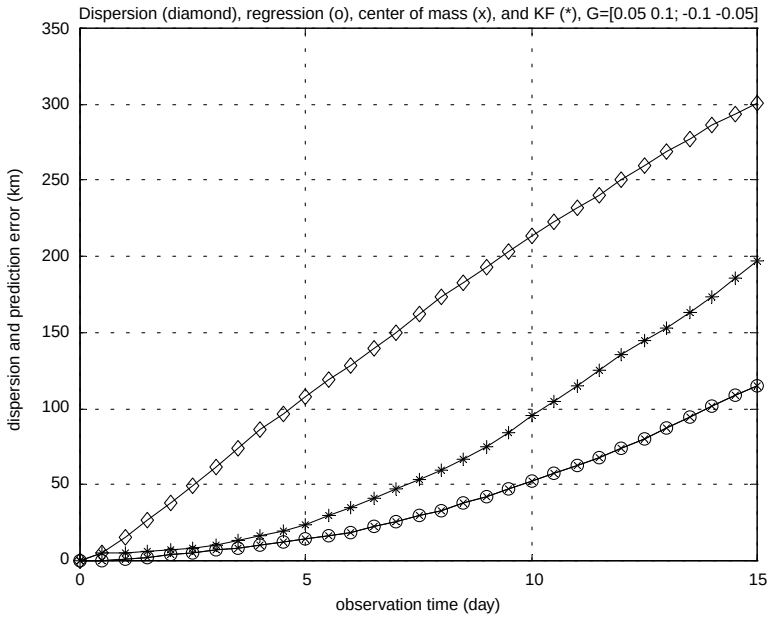


(b)

FIG. 9. (a) *Same as in Figure* 8(a) *with slightly different Lagrangian correlation time* 2.5 *days.* (b) *Same as in Figure* 9(a) *for a nonzero shear:* $\gamma = 0.05\,day^{-1}$, $\omega = 0$.



(c)



(d)

FIG. 9. (c) *Same as in Figure* 9(a) *for a nonzero shear:* $\gamma = 0$, $\omega = 0.1\,day^{-1}$. (d) *Same as in Figure* 9(a) *for a nonzero shear:* $\gamma = 0.05$, $\omega = 0.1\,day^{-1}$.

about 140 km, while for the regression it is the same: 60 km. Finally, combining both cases ($\omega = 0.1$ and $\gamma = 0.05$) we observe that the KF error is almost twice as much as the RA error (Figure 9(d)). Thus, with no doubt the regression algorithm performs better in the presence of a deterministic linear shear flow. This is because it is based on the assumption of linear dependence of the particle current position on the initial position. In fact, for purely linear flow the RA gives the exact prediction. Thus, the presence of the mean flow implies the better performance of the regression algorithm.

7. Comparison with KF and CM algorithms (real data). The final stage in this study is to apply RA to predict the motion of oceanic drifters released in a cluster and compare its performance with that of the simulations. It was found in Özgökmen et al. (2001) that during the period in which drifters remain close to one another as a tight cluster (quantified by the number of drifters within the velocity space correlation scale R), the CM method is a simple yet effective means of predicting the drifter location. However, how the prediction accuracy of RA compares to that of CM and the far more complicated technique KF for oceanic drifters needs to be investigated.

The drifter data are obtained from the NOAA Atlantic Oceanographic and Meteorological Laboratory, Global Drifter Center, by searching the entire 1988–1996 data set for a group of 5 or more drifters released within the velocity space correlation scale R. The drifter data are used as provided by the Global Drifter Center, which lists the drifter positions in six-hour intervals after standard quality control procedures (e.g., Hansen and Poulain (1996)) and no further processing has been applied. A total of 7 clusters, each consisting of 5–10 drifters, has been analyzed. In the following, we concisely present results from 3 of these clusters, since the main conclusions remain the same for others. These 3 drifter clusters have been released in the tropical Pacific Ocean, which is a region characterized by strong currents and shears and lacking the effect of coastlines or boundaries. The mean currents (Figure 10) are calculated using the technique described in Bauer et al. (1999) from the entire drifter data set collected under the World Ocean Circulation Experiment (WOCE) during 1988–1996. This figure depicts the general circulation pattern in this region, which is governed by the westward North Equatorial Current north of $10°N$, the eastward North Equatorial

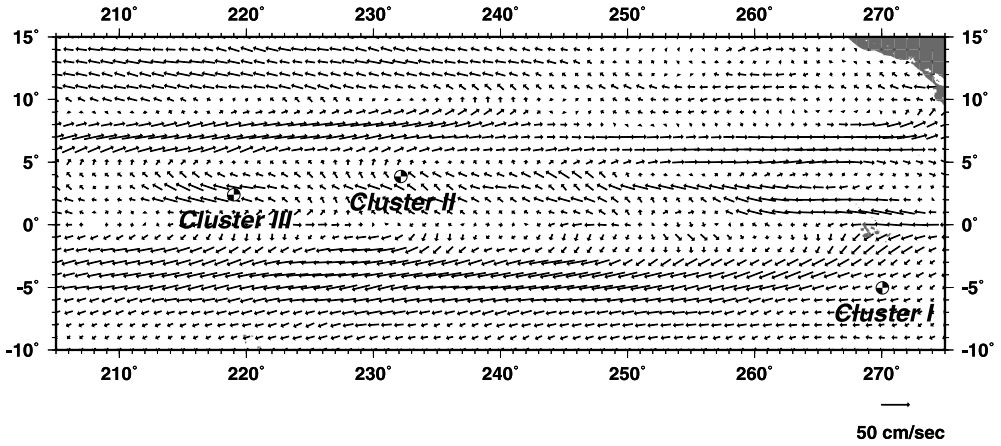


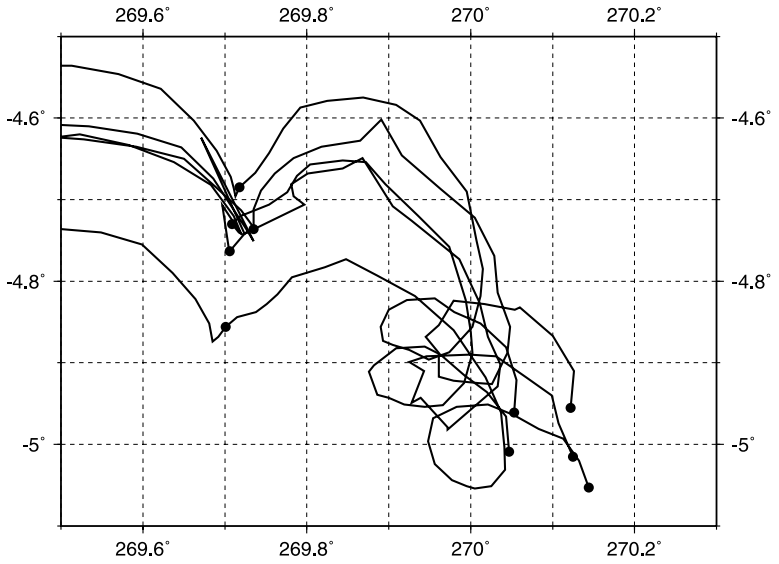
FIG. 10. *The climatological mean flow field depicting the major currents in the tropical Pacific Ocean and the initial release locations of clusters* I, II, *and* III.

Countercurrent between $4°N$ and $9°N$, and the westward South Equatorial Current, which extends across the equator to $10°S$. Drifters in the first cluster (cluster I) have been released in the South Equatorial Current, whereas the others in clusters II and III have been launched just south of the North Equatorial Countercurrent. The mean currents, however, are not a good indicator of drifter motion (Özgökmen et al. (2000), (2001)) and are discussed here only to provide the general surface flow characteristics of this oceanic region. It is also important to point out that in order to be able to deploy these drifters on tight grids, a real-time analysis of a variety of data sets, including current meter profilers and satellite data images, has been necessary for a detailed dynamical analysis due to strong currents in this region (e.g., Flament et al. (1996)). Finally, the space correlation radius of the velocity field in the tropical Pacific Ocean is taken as the Rossby deformation radius $R = 250$ km (Cushman-Roisin (1994)), and the Lagrangian correlation time is taken as $\tau = 3$ days, based on the analysis of drifter motion in the WOCE data set (Bauer et al. (1999)).

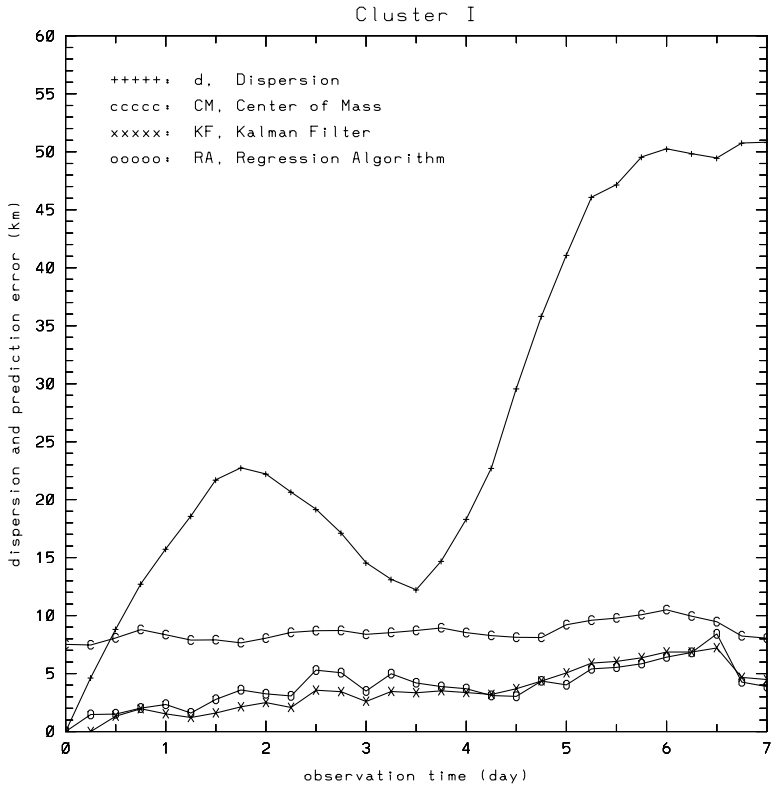
Clusters I–III are sorted according to the difficulty of prediction, quantified by the initial scale of the cluster, the velocity variance and prediction period. Cluster I consists of 5 drifters launched within a scale ($\sim R_0$) of approximately 10 km from each other (Figure 11(a)), and the prediction algorithms are applied for 7 days of particle motion, during which the velocity variance is approximately $\sigma_v^2 = 430$ km^2/day^2. During 7 days of motion, these drifters do not spread apart significantly. Given $R_0 \ll R$ and the low velocity variance, one can anticipate very good performance by RA based on the results from theory and stochastic simulations. Dispersion $d(t)$ and prediction errors $s(t)$ from RA, CM, and KM are calculated by sequentially selecting each drifter as predictand and the remaining others in the cluster as predictors, corresponding to the root mean square of that of all cluster particles. The results are shown in Figure 11(b) for cluster I for an observation time of 7 days. This figure shows that prediction errors of both KF and RA are less than that of CM during the observation period and that RA is as accurate as KF. More quantitatively, dispersion reaches approximately 51 km at $T = 7$ days, error from CM is 8 km ($s_r = 0.16$), and error from KF and RA is about 5 km ($s_r = 0.1$).

Cluster II consists of 7 drifters that are also released with a mean diameter of approximately 10 km, but disperses much faster than Cluster I due to a higher velocity variance of $\sigma_v^2 = 720$ km^2/day^2, and the mean cluster diameter reaches 25 km and 50 km after 7 and 14 days of observation time, respectively (Figure 12(a)). Dispersion and prediction errors for cluster II over an observation period of 14 days are shown in Figure 12(b), and the conclusion remains the same as for cluster I; prediction errors of both KF and RA are less than that of CM during the observation period, and RA is as accurate as KF. Dispersion reaches approximately 136 km at $T = 14$ days, error from CM is 44 km ($s_r = 0.32$), and errors from KF and RA are about 26 km ($s_r = 0.19$). The sensitivity of the prediction accuracy of RA to the number of predictors p is investigated by randomly eliminating drifters from cluster II. Figure 12(c) shows the dispersion curve based on the entire cluster and prediction errors calculated for $p = 6$ (same as in Figure 12(b)), $p = 5$, $p = 4$, and $p = 3$. When $p = 3$, a drastic reduction of prediction accuracy takes place, which is found to be independent of the combination of chosen predictors in this cluster. Otherwise, the prediction accuracy gradually decreases as the number of predictors is decreased from 6 to 4, but the accuracy of the method using 4 to 6 predictors remains essentially constant for $T \leq \tau$ or for $T \leq 3$ days.

The motion of cluster III, consisting of 10 drifters, is investigated for 21 days

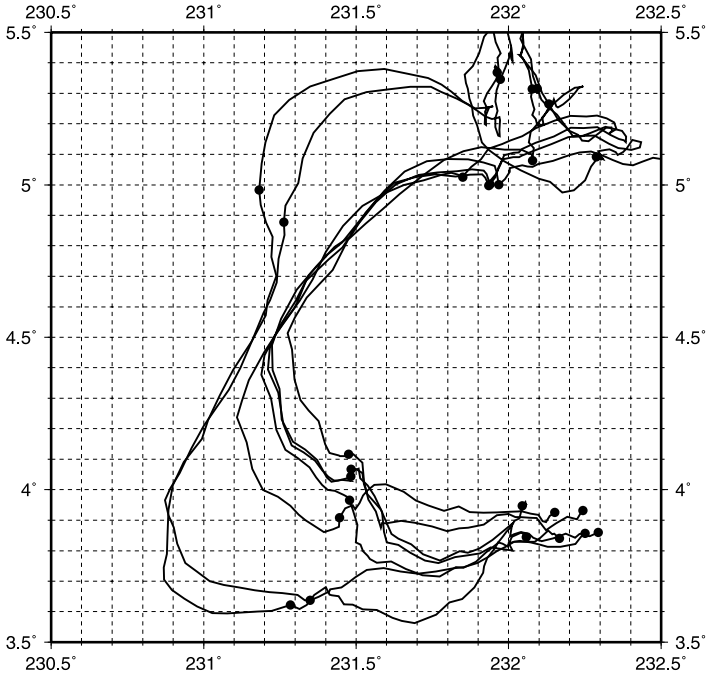


(a)

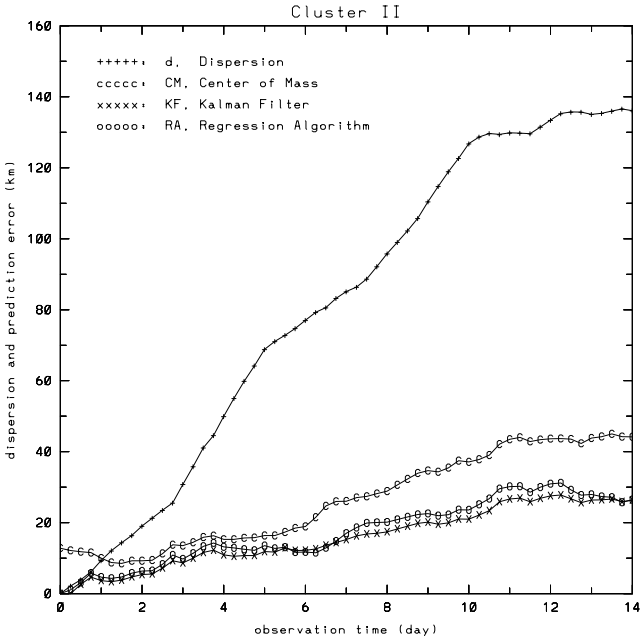


(b)

FIG. 11. (a) *Drifters trajectories in cluster* I. *The circles mark* 7*-day intervals.* (b) *Comparison of the dispersion,* $d(t)$*, and prediction errors,* $s(t)$*, of RA, CM, and KM for an observation time of* 7 *days for cluster* I.

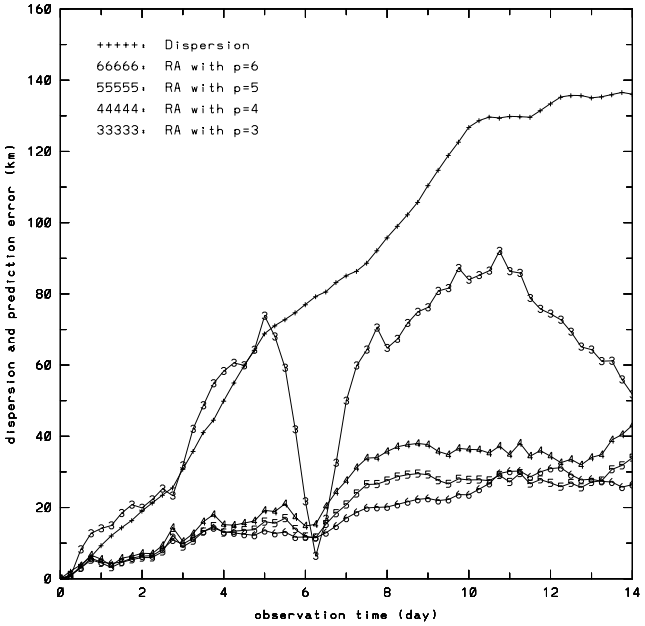


(a)



(b)

FIG. 12. (a) *Drifters trajectories in cluster* II. *The circles mark* 7*-day intervals.* (b) *Comparison of the dispersion,* $d(t)$*, and prediction errors,* $s(t)$*, of RA, CM, KM for an observation time of* 14 *days for cluster* II.



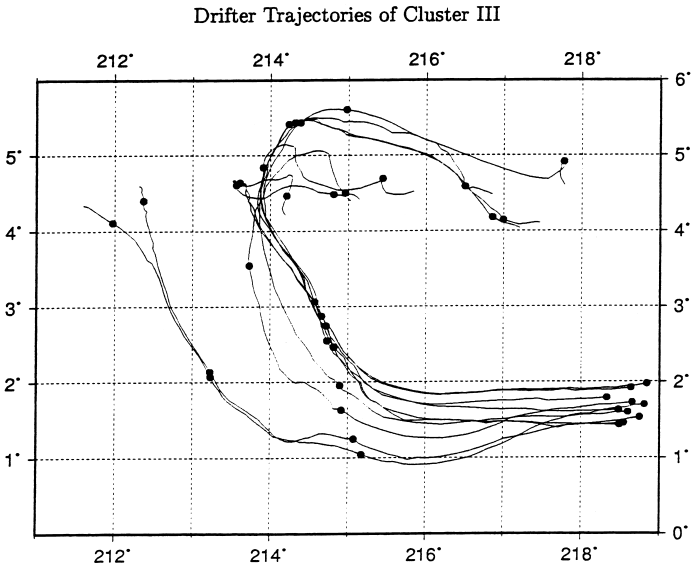
(c)

FIG. 12. (c) *Sensitivity of the prediction error of RA to the number of predictors in cluster* II.

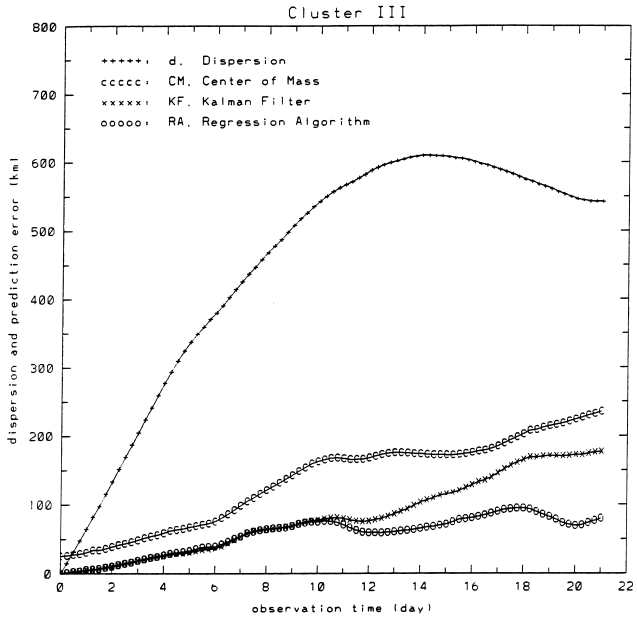
during which velocity variance is $\sigma_v^2 = 2240$ km^2/day^2, or the highest of the three clusters. These drifters were released over an area with an approximate diameter of 30 km, but this scale increases to approximately 100, 180, and 250 km after 7, 14, and 21 days, respectively (Figure 13(a)). Dispersion and prediction errors for cluster III are shown in Figure 13(b). During the first 10 days, prediction errors of both KF and RA are approximately the same and less than that of CM, but during the second half of the observation period, the error of KF increases faster than that of RA. This increase appears to be related to the inability of the KF algorithm to follow the bifurcation of some drifters in a larger group as effectively as the RA technique. Dispersion is 426 km (543 km), error from CM is 99 km (235 km) or $s_r = 0.23$ ($s_r = 0.43$), error from KF is 55 km (176 km) or $s_r = 0.13$ ($s_r = 0.32$), and error from RA is 54 km (80 km) or $s_r = 0.13$ ($s_r = 0.15$) at $T = 7$ days ($T = 21$ days).

All in all, the real data comparison of different prediction algorithms is in good qualitative agreement with the simulation results. Even the prediction error values are of the same order, as our simple error theory concludes. Deviations are related to oversimplifications accepted in the considered stochastic model such as the shear flow linearity and fluctuations isotropy.

In summary, the algorithm described in this study presents several important simplifications with respect to the KF method developed and investigated by Piterbarg (2001b) and Özgökmen et al. (2000), (2001): (i) This algorithm does not require any parameters, such as the Lagrangian parameters describing the characteristics of the underlying flow, the velocity correlation space scale R, and the Lagrangian correlation time scale τ. (ii) RA does not utilize the mean flow field, the calculation of which requires large data sets and the associated subgrid scale interpolation introduces further errors. (iii) RA does not need to be initialized with turbulent velocity fluctuations at



(a)



(b)

FIG. 13. (a) *Drifter trajectorie in cluster* III. *The circles mark* 7*-day intervals.* (b) *Comparison of the dispersion,* $d(t)$*, and prediction errors,* $s(t)$*, of RA, CM, KM for an observation time of* 21 *days for cluster* III.

the launch location. (iv) RA is not based on the integration of velocity field to estimate the particle position, which necessarily leads to accumulation of velocity errors as errors of drifter location. (v) Consequently RA is computationally far simpler than KF. Despite these simplifications, it is found on the basis of several oceanic clusters that RA outperforms CM and that RA is as accurate as KF. Also, predictions from RA appear to remain applicable over a time scale of $T \gg \tau$, or much longer than one would anticipate. In future studies, it will be investigated theoretically and numerically how this method performs when $R_0 \approx R$, which is likely to be the case in mid- and high-latitude oceans and for bifurcating clusters.

Appendix.

A.1. Prediction error in terms of separation. Using the definition (4) and expression (6) for b_k, obtain

$$\begin{aligned} s^2 &= E|\hat{\mathbf{r}}_M(t) - \mathbf{r}_M(t)|^2 \\ &= E\{(\mathbf{r}_c(t) + \mathbf{S}(t)\mathbf{S}(0)^{-1}(\mathbf{r}_M(0) - \mathbf{r}_c(0)) - \mathbf{r}_M(t))^T(\mathbf{r}_c(t) \\ &\quad + \mathbf{S}(t)\mathbf{S}(0)^{-1}(\mathbf{r}_M(0) - \mathbf{r}_c(0)) - \mathbf{r}_M(t))\} \\ &= E\{(\mathbf{r}_c(t) - \mathbf{r}_M(t))^T(\mathbf{r}_c(t) - \mathbf{r}_M(t))\} \\ &\quad + 2E\{(\mathbf{r}_c(t) - \mathbf{r}_M(t))^T\mathbf{S}(t)\}\mathbf{S}(0)^{-1}(\mathbf{r}_M(0) - \mathbf{r}_c(0)) \\ &\quad + (\mathbf{r}_M(0) - \mathbf{r}_c(0))^T\mathbf{S}(0)^{-1}E\{\mathbf{S}(t)^T\mathbf{S}(t)\}\mathbf{S}(0)^{-1}(\mathbf{r}_M(0) - \mathbf{r}_c(0)) \\ &= E\{(\mathbf{r}_c(t) - \mathbf{r}_M(t))^T(\mathbf{r}_c(t) - \mathbf{r}_M(t))\} \\ &\quad + 2\textstyle\sum_{k=1}^{p} b_k E\{(\mathbf{r}_c(t) - \mathbf{r}_M(t))^T(\mathbf{r}_k(t) - \mathbf{r}_c(t))\} \\ &\quad + \textstyle\sum_{k,l=1}^{p} b_k b_l E\{(\mathbf{r}_k(t) - \mathbf{r}_c(t))^T(\mathbf{r}_l(t) - \mathbf{r}_c(t))\}. \end{aligned} \tag{A1}$$

Then, dropping the time argument for brevity,

$$\begin{aligned} E\{(\mathbf{r}_k - \mathbf{r}_c)^T(\mathbf{r}_l - \mathbf{r}_c)\} &= \tfrac{1}{p^2}\textstyle\sum_{i,j=1}^{p} E\{(\mathbf{r}_k - \mathbf{r}_i)^T(\mathbf{r}_l - \mathbf{r}_j)\} \\ &= -\tfrac{1}{2p^2}\textstyle\sum_{i,j=1}^{p}(\rho_{kl} + \rho_{ij} - \rho_{ki} - \rho_{lj}) \\ &= -\tfrac{1}{2}\rho_{kl} + \tfrac{1}{2p}\textstyle\sum_{i=1}^{p}(\rho_{ki} + \rho_{li}) - \tfrac{1}{2p^2}\textstyle\sum_{i,j=1}^{p}\rho_{ij}. \end{aligned} \tag{A2}$$

In the latter we used the relation

$$E\{\mathbf{r}_k{}^T\mathbf{r}_l\} = \frac{1}{2}(-\rho_{kl} + E\{\mathbf{r}_k{}^T\mathbf{r}_k\} + E\{\mathbf{r}_l{}^T\mathbf{r}_l\}).$$

By substituting (A2) into (A1), using the obvious relation

$$\sum_{k=1}^{p} b_k = 0,$$

and changing the summation indexes, we obtain (5).

A.2. Separation for Brownian flow (close initial positions). The function

$$\rho(t, r) = E\{(\mathbf{r}_1(t) - \mathbf{r}_2(t))^2\},$$

where $|\mathbf{r}_1(0) - \mathbf{r}_2(0)| = r$, satisfies

$$\frac{\partial \rho}{\partial t} = L_s\rho\ , \qquad \rho(0, r) = r^2, \tag{A3}$$

where in the isotropic case the generator for the separation process is

$$L_s = \frac{b_0 - b_N(r)}{r}\frac{\partial}{\partial r} + (b_0 - b_L(r))\frac{\partial^2}{\partial r^2}\ .$$

Substitute expansions (13), (15) into (A3). The result is

$$\frac{d\rho_1}{dt} = \overline{\beta}\rho_1, \quad \rho_1(0) = 1, \qquad \frac{d\rho_0}{dt} = \beta_0\rho_0 + \overline{\gamma}\rho_1, \quad \rho_0(0) = 0,$$

where

$$\overline{\beta} = \beta_N + \beta_L, \quad \beta_0 = 2\beta_N + 6\beta_L, \qquad \overline{\gamma} = \gamma_N + \gamma_L.$$

Solving the latter equations we obtain (16).

A.3. Dispersion for the flow with memory in presence of linear shear flow. For simplicity assume $\mathbf{U} = 0$. From (20) it follows that

$$\mathbf{r}(t) = \bar{\mathbf{r}}(t) + \int_0^t \exp(\mathbf{G}(t-s))\mathbf{v}(s)ds,$$

where $\mathbf{v}(t)$ is a two-dimensional Ornstein–Uhlenbeck process with the covariance $\sigma_v^2\exp(-t/\tau)\mathbf{I}$ and

$$\bar{\mathbf{r}} = \exp(\mathbf{G}t)\mathbf{r}_0 + (\exp(\mathbf{G}t) - \mathbf{I})\mathbf{G}\mathbf{U}.$$

Set

$$d_0^2 = (\bar{\mathbf{r}} - \mathbf{r}_0)^2;$$

then

$$d^2 = d_0^2 + \sigma_v^2\int_0^t\int_0^t Sp(\exp(\mathbf{G}(t-s_1))\exp(\mathbf{G}^T(t-s_2)))\exp(-|s_1-s_2|/\tau)ds_1ds_2,$$

where $Sp(\mathbf{A})$ means the race of matrix A. Then we use the following relations:

$$Sp(\mathbf{A}) = \lambda_1(\mathbf{A}) + \lambda_2(\mathbf{A}), \quad \lambda(\exp(\mathbf{A})) = \exp\lambda((\mathbf{A})),$$

where $\lambda(\mathbf{A})$ is an eigenvalue of $\mathbf{A}$ and arrive at (21).

A.4. Separation for the flow with memory (close initial positions and velocities). In the case considered, the separation satisfies

$$\frac{\partial\rho}{\partial t} = L_s\rho, \quad \rho|_{t=0} = x^2 + y^2, \tag{A4}$$

with the generator (24) written in the coordinatewise form

$$\begin{aligned} L_s = {} & (\gamma y + \omega x + u)\frac{\partial}{\partial x} - (\gamma y + \omega x + v)\frac{\partial}{\partial y} - \frac{1}{\tau}u\frac{\partial}{\partial u} - \frac{1}{\tau}v\frac{\partial}{\partial v} \\ & + (b_0 - b_N(r) - \frac{x^2}{r^2}(b_L(r) - b_N(r))\frac{\partial^2}{\partial u^2} + (b_0 - b_N(r) - \frac{y^2}{r^2}(b_L(r) - b_N(r))\frac{\partial^2}{\partial v^2} \\ & - \frac{2xy}{r^2}(b_L(r) - b_N(r))\frac{\partial^2}{\partial u\partial v}. \end{aligned}$$

Substituting expansions (25), (27) into (A4) we obtain (27), where

$$\mathbf{A} = \begin{pmatrix} 2\gamma & -2\omega & 0 & 2\beta_L & 0 & 2\beta_N & 0 & 0 & 0 & 0 \\ \omega & 0 & -\omega & 0 & -2\beta & 0 & 0 & 0 & 0 & 0 \\ 0 & 2\omega & -2\gamma & 2\beta_N & 0 & 2\beta_L & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -2/\tau & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -2/\tau & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 0 & -2/\tau & 0 & 0 & 0 & 1 \\ 2 & 0 & 0 & 0 & 0 & 0 & \gamma - 1/\tau & 0 & -\omega & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 & 0 & \gamma - 1/\tau & 0 & -\omega \\ 0 & 2 & 0 & 0 & 0 & 0 & \omega & 0 & -\gamma - 1/\tau & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 & 0 & \omega & 0 & -\gamma - 1/\tau \end{pmatrix},$$

where $\beta = \beta_N - \beta_L$,

$$\mathbf{a}_0 = (1\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0)^T.$$

For the zero mean flow $\gamma = 0$, $\omega = 0$, and (27) reduces to (29) for $\rho_0 = a_4 + a_6$, $\rho_1 = a_1 + a_3$, and $\rho_{01} = a_7 + a_{10}$.

REFERENCES

S. Bauer, M. S. Swenson, A. Griffa, A. J. Mariano, and K. Owens (1999), *Eddy-mean flow decomposition and eddy-diffusivity estimates in the tropical Pacific Ocean*, J. Geophys. Res., 103, pp. 30855–30871.

P. Baxendale and T. Harris (1986), *Isotropic stochastic flows*, Ann. Probab., 14, pp. 1155–1179.

S. Castellari, A. Griffa, T. M. Özgökmen and P.-M. Poulain (2001), *Prediction of particle trajectories in the Adriatic Sea using Lagrangian data assimilation*, J. Marine Sys., 29, pp. 33–50.

B. Cushman-Roisin (1994), *Introduction to Geophysical Fluid Dynamics*, Prentice-Hall, Englewood Cliffs, NJ, p. 285.

R. E. Davis (1991a), *Lagrangian ocean studies*, Ann. Rev. Fluid Mech., 23, pp. 43–64.

R. E. Davis (1991b), *Observing the general circulation with floats*, Deep-Sea Research, 38, pp. 5531–5571.

P. J. Flament, S. C. Kennan, R. A. Knox, P. P. Niiler, and R. L. Bernstein (1996), *The three-dimensional structure of an upper ocean vortex in the tropical Pacific Ocean*, Nature, 383, pp. 610–613.

A. Griffa (1996), *Applications of stochastic particle models to oceanographic problems*, in Stochastic Modelling in Physical Oceanography, R. Adler, P. Muller, and B. Rozovskii, eds., Birkhäuser, Boston, pp. 113–128.

A. Griffa, K. Owens, L. Piterbarg, and B. Rozovskii (1995), *Estimates of turbulence parameters from Lagrangian data using a stochastic particle model*, J. Marine Res., 53, pp. 371–401.

D. V. Hansen and P.-M. Poulain (1996), *Quality control and interpolations of WOCE/TOGA drifter data*, J. Atmos. Ocean Techn., 13, pp. 900–909.

H. Kunita (1990), *Stochastic Flows and Stochastic Differential Equations*, Cambridge University Press, Cambridge, UK.

R. S. Liptser and A. N. Shiryaev (1978), *Statistics of Random Processes*, Springer-Verlag, Berlin.

A. S. Monin and A. M. Yaglom (1975), *Statistical Fluid Mechanics: Mechanics of Turbulence*, MIT Press, Cambridge, MA.

T. M. Özgökmen, A. Griffa, L. I. Piterbarg, and A. Mariano (2000), *On the predictability of Lagrangian trajectories in the ocean*, J. Atmos. Ocean Techn., 17, pp. 366–383.

T. M. Özgökmen, L. I. Piterbarg, A. J. Mariano, and E. H. Ryan (2001), *Predictability of drifter trajectories in the tropical Pacific Ocean*, J. Phys. Oceanogr., 31, pp. 2691–2720.

L. I. Piterbarg (1998), *Drift estimation for Brownian flows*, Stochastic Process. Appl., 79, pp. 132–149.

L. I. Piterbarg (2001a), *The top Lyapunov exponent for a stochastic flow modeling the upper ocean turbulence*, SIAM J. Appl. Math., 62, pp. 777–800.

L. I. Piterbarg (2001b), *Short-term prediction of Lagrangian trajectories*, J. Atmos. Ocean Techn., 18, pp. 1398–1410.

T. Schneider (1998), *Lagrangian Drifter Models as Search and Rescue Tools*, M.S. thesis, Dept. of Meteorology and Physical Oceanography, University of Miami, Miami, FL.

D. J. Thomson (1986), *A random walk model of dispersion in turbulent flows and its application to dispersion in a valley*, Quart. J. R. Met. Soc., 112, pp. 511–529.

E. Zambianchi and A. Griffa (1994), *Effects of finite scales of turbulence on dispersion estimates*, J. Marine Res., 52, pp. 129–148.

C. L. Zirbel and E. Cinlar (1996), *Dispersion of particle systems in Brownian flows*, Adv. in Appl. Probab., 28, pp. 53–74.

SIAM J. APPL. MATH.
Vol. 63, No. 1, pp. 149–168

A RIGOROUS TREATMENT OF A FOLLOW-THE-LEADER TRAFFIC MODEL WITH TRAFFIC LIGHTS PRESENT*

BRENNA ARGALL†, EUGENE CHELESHKIN†, J. M. GREENBERG†, COLIN HINDE†, AND PEI-JEN LIN†

Abstract. Traffic flow on a unidirectional roadway in the presence of traffic lights is modeled. Individual car responses to green, yellow, and red lights are postulated and these result in rules governing the acceleration and deceleration of individual cars. The essence of the model is that only specific cars are directly affected by the lights. The other cars behave according to simple follow-the-leader rules which limit their speed by the spacing between them and the car directly ahead. The model has a number of desirable properties; namely, cars do not run red lights, cars do not smash into one another, and cars exhibit no velocity reversals. In a situation with multiple lights operating in-phase, we get, after an initial start-up period, a constant number of cars through each light during any green-yellow period. Moreover, this flux is less by one or two cars per period than the flux obtained in discretized versions of the idealized Lighthill–Whitham–Richards model which allows for infinite accelerations.

Key words. traffic flow, follow-the-leader, relaxation models, conservation laws

AMS subject classification. 35

PII. S0036139901391215

1. Introduction, model description, and statement of results. In this note we examine the behavior of traffic on a unidirectional highway when multiple traffic lights are present. For simplicity we assume the lights operate in-phase.

The model postulates the dynamics of individual cars but may also be thought of as a coarse discretization of a continuum model introduced recently by Greenberg [1], Aw and Rascle [2], Aw, Klar, Materne, and Rascle [3], and Zhang [9] (details of this correspondence may be found in section 4, (4.6)–(4.8)).

We assume we are presented with an empirically determined function $s \to \mathcal{V}(s)$ on $L \leq s$ which satisfies

$$\mathcal{V}(L^+) = 0, \tag{1.1}$$

$$\frac{d\mathcal{V}}{ds}(s) > 0 \text{ and } \frac{d^2\mathcal{V}}{ds^2}(s) < 0, \qquad L \leq s < \infty, \tag{1.2}$$

and

$$\lim_{s\to\infty} \left(\mathcal{V}(s),\ \frac{d\mathcal{V}}{ds}(s),\ \frac{d^2\mathcal{V}}{ds^2}(s) \right) = (\mathcal{V}_\infty > 0, 0, 0). \tag{1.3}$$

The independent variable s is interpreted as the spacing between cars, L is the minimum car spacing (a lower bound for L is the length of typical car), and $\mathcal{V}_\infty > 0$

*Received by the editors June 22, 2001; accepted for publication (in revised form) March 1, 2002; published electronically August 28, 2002.
http://www.siam.org/journals/siap/63-1/39121.html
†Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA 15213 (greenber@andrew.cmu.edu). The research of the third author was partially supported by the Applied Mathematical Sciences Program, U.S. Department of Energy and by the U.S. National Science Foundation.

is the maximum allowable speed of a car. A typical function, and one we shall use in simulations, is

$$\mathcal{V}(s) = \mathcal{V}_\infty \left(1 - \frac{L}{s}\right), \qquad L \le s < \infty. \tag{1.4}$$

In this classic Lighthill–Whitham–Richards model [4, 5, 6, 7] the function $\mathcal{V}(\cdot)$ gives the velocity of individual cars; in ours it provides an upper bound for the velocity of an individual car. An extensive discussion of suitable functions $\mathcal{V}(\cdot)$ may be found in [8, Chapter 4] and the references contained therein. Suffice it to say that the functions $\mathcal{V}(\cdot)$ in our model are consistent with those used in practice.

In this model $x_k(t), 1 \le k \le N$, denotes the position of the kth car at time t, and $0 \le u_k(t)$ is the velocity of the kth car. Throughout,

$$\frac{dx_k}{dt} = u_k, \qquad 1 \le k \le N, \tag{1.5}$$

and the cars are ordered so that $(x_{k+1} - x_k)(t) \ge L$, $1 \le k \le N-1$. During time intervals where the lights are green we assume that

$$u_k = \mathcal{V}((x_{k+1} - x_k)(t)) + \alpha_k, \qquad 1 \le k \le N, \tag{1.6}$$[1]

where $\alpha_k(t) \le 0$ satisfies

$$\epsilon \frac{d\alpha_k}{dt} = -\alpha_k, \qquad 1 \le k \le N. \tag{1.7}$$

The parameter $\epsilon > 0$ may be thought of as a relaxation time. Equations (1.6) and (1.7) imply that during the green light periods the velocities, u_k, satisfy

$$\frac{du_k}{dt} = \mathcal{V}'(x_{k+1} - x_k)(u_{k+1} - u_k) + \frac{(\mathcal{V}(x_{k+1} - x_k) - u_k)}{\epsilon}, \qquad 1 \le k \le N-1, \tag{1.7a}$$

and

$$\frac{du_N}{dt} = \frac{(\mathcal{V}_\infty - u_N)}{\epsilon}. \tag{1.7b}$$

The interesting feature of our model is how yellow or red lights effect the dynamics of an individual car. Our traffic lights cycle from green to yellow to red, and the numbers $0 < TG$, $0 < TY$, and $0 < TR$ denote the duration of the green, yellow, and red lights. At time $t = 0$ we assume we have a sequence of N cars located at

$$x_k(0) = (k - k_0)L_1, \qquad 1 \le k \le N, \tag{1.8}$$

where $L_1 \ge L$ (again L is the minimum allowable auto spacing), and we assume these cars are all at rest; i.e.,

$$u_k(0) = 0, \qquad 1 \le k \le N. \tag{1.9}$$

Finally, we assume they are at traffic lights located at $x = l_I, 1 \le I \le M$, where

$$(N - k_0)L_1 < l_1 < l_2 < \cdots < l_M. \tag{1.10}$$

[1] When $k = N$, $u_N = \mathcal{V}_\infty + \alpha_N$.

We further assume that each intersection is of width $w > 0$ and let

$$t_m = (m-1)(TG + TY + TR), \qquad m = 1, 2, \ldots, \tag{1.11}$$

denote the start of the mth light cycle.

During the time interval $t_m \le t \le t_m + TG$ all cars satisfy (1.5)–(1.7). At time $t_y \stackrel{def}{=} t_m + TG$, the green lights turn yellow, and this will have an effect on the traffic flow.

We start by describing what happens to the lead car, the one indexed by N, when it encounters a light at $x = l$. We assume that

$$x_N(t_y) < l. \tag{1.12}$$

If

$$x_N(t_y) + u_N(t_y)TY \ge l + w + L, \tag{1.13}$$

then the lead car will be able to completely clear the intersection if it travels at its current speed $u_N(t_y)$. We allow it to clear the intersection by following its standard dynamics; that is, over the time interval $t_y \le t \le t_{m+1}$ the Nth car satisfies

$$\frac{dx_N}{dt} = u_N, \tag{1.14}$$

where

$$u_N = \mathcal{V}_\infty + \alpha_N \tag{1.15}$$

and $\alpha_N \le 0$ satisfies

$$\epsilon \frac{d\alpha_N}{dt} = -\alpha_N. \tag{1.16}$$

Following these dynamics the lead car accelerates through the intersection.

On the other hand, if

$$x_N(t_y) + u_N(t_y)TY < l + w + L, \tag{1.17}$$

then it will be impossible for the Nth car to clear the intersection during the yellow phase if it continues to travel at its current speed. If

$$x_N(t_y) + u_N(t_y)(TY + TR) \le l, \tag{1.18}$$

then over the time interval $t_y \le t \le t_{m+1}$ we require it satisfies the modified dynamics

$$\frac{dx_N}{dt} = u_N \quad \text{and} \quad \frac{du_N}{dt} = 0; \tag{1.19}$$

i.e., we insist that it travels at its current speed. This strategy avoids the Nth car accelerating and then possibly having to decelerate as it nears the light.

If (1.17) holds and (1.18) is violated, the lead car will have to slow down and possibly stop. When it satisfies the additional inequality

$$x_N(t_y) + u_N(t_y)(TY + TR)/2 > l, \tag{1.20}$$

the lead car is mandated to satisfy

$$\frac{dx_N}{dt} = u_N \quad \text{and} \quad \frac{du_N}{dt} = \begin{cases} \dfrac{-u_N^2(t_y)}{2(l - x_N(t_y))}, t_y \le t \le t_y + \dfrac{2(l - x_N(t_y))}{u_N(t_y)} \\ \\ 0, t_y + \dfrac{2(l - x_N(t_y))}{u_N(t_y)} \le t \le t_{m+1}.^2 \end{cases} \tag{1.21}$$

This constant deceleration strategy brings the Nth car to rest at $x = l$ at $t = t_y + \frac{2(l - x_N(t_y))}{u_N(t_y)} \le t_{m+1}$, and it then sits at the light until $t = t_{m+1}$.

Finally, when

$$x_N(t_y) + u_N(t_y)(TY + TR) > l \quad \text{and} \quad x_N(t_y) + u_N(t_y)(TY + TR)/2 \le l, \tag{1.22}$$

the lead car is mandated to satisfy

$$\frac{dx_N}{dt} = u_N(t) \quad \text{and} \quad \frac{du_N}{dt} = \frac{-2(x_N(t_y) + u_N(t_y)(TY + TG) - l)}{(TY + TG)^2}$$

over the whole interval $t_y \le t \le t_{m+1}$. This strategy brings the car to the light at $x = l$ at t_{m+1} with velocity

$$u_N(t_{m+1}) = \frac{2(l - x_N(t_y))}{(TY + TR)} - u_N(t_y) > 0. \tag{1.23}$$

We note that if the lead car satisfies (1.17), then the cars with indices $k \le N - 1$ follow their standard dynamics (1.5)–(1.7) over $[t_y, t_{m+1}]$ unless they happen to be influenced by some other light at $x = l' < l$.

Having described what happens when the lead car encounters a yellow light at $x = l$, we turn our attention to what happens when other cars encounter the same light. We let $k_l \le N - 1$ be the largest integer so that

$$x_{k_l}(t_y) < l, \tag{1.24}$$

and we let $p_l \le k_l$ be the largest integer so that

$$x_{p_l}(t_y) + \min_{p_l \le j \le k_l} u_j(t_y)TY < l + w + L. \tag{1.25}$$

The p_lth car will be the first one that does not get through the light at $x = l$.

We first consider the situation when $p_l < k_l$. We assume the existence of a number $\lambda \ge 1$ such that cars travelling with the maximum speed $\mathcal{V}_\infty$ can safely brake at a constant deceleration rate $a = \frac{-\mathcal{V}_\infty^2}{2\lambda L}$ over a road segment of length λL.

We first focus our attention on the situation in which

$$x_{p_l}(t_y) < l - \lambda L. \tag{1.26}$$

[2]The dynamics described by (1.21) are equivalent to

$$\frac{dx_N}{dt} = \frac{u_N(t_y)(l - x_N(t))^{1/2}}{2(l - x_N(t_y))^{1/2}}, \ t_y \le t \le t_y + \frac{2(l - x_N(t_y))}{u_N(t_y)}$$

and

$$\frac{dx_N}{dt} = 0, t_y + \frac{2(l - x_N(t_y))}{u_N(t_y)} \le t \le t_{m+1}.$$

Our basic strategy is to let cars with indices $k \geq p_l+1$ follow their standard dynamics (1.5)–(1.7) over $t_y \leq t \leq t_{m+1}$. The cars with indices $p_l + 1 \leq k \leq k_l$ will clear the intersection by $t_m + TG + TY \stackrel{def}{=} t_r$; i.e., they will satisfy $x_k(t_r) \geq l + w + L$. This follows from the observation that local spatial minima in the velocity are nondecreasing in t (for details see (2.79)–(2.81)).

Rules for the p_lth car. So long as $t_y \leq t \leq t_r$ and $x_{p_l}(t) < l - \lambda L$ we let the p_lth car follow its standard dynamics (1.5)–(1.7). If there is a first $t_{p_l} < t_r$ so that $x_{p_l}(t_{p_l}) = l - \lambda L$, then the driver must decide what to do. In the unlikely event that

$$u_{p_l}(t_{p_l})(t_{m+1} - t_{p_l}) \leq \lambda L, \tag{1.27}$$

then over the interval $[t_{p_l}, t_{m+1}]$ the p_lth car is required to satisfy

$$\frac{dx_{p_l}}{dt} = \min\left(u_{p_l}(t_y),\ U_{p_l}(t)\right) \stackrel{def}{=} u_{p_l}(t)$$

and

$$\begin{aligned}\frac{dU_{p_l}}{dt} &= \mathcal{V}'\left(x_{p_l+1} - x_{p_l}\right)\left(u_{p_l+1} - U_{p_l}\right) \\ &\quad + \frac{\left(\mathcal{V}\left(x_{p_l+1} - x_{p_l}\right) - U_{p_l}\right)}{\epsilon}\end{aligned} \tag{1.28}$$

and

$$U_{p_l}(t_y) = u_{p_l}(t_y).$$

On the other hand, if

$$u_{p_l}(t_{p_l})(t_{m+1} - t_{p_l}) > \lambda L, \tag{1.29}$$

then the p_lth car will have to slow down and possibly stop.

When the p_lth car satisfies the additional inequality

$$u_{p_l}(t_{p_l})(t_{m+1} - t_{p_l})/2 > \lambda L, \tag{1.30}$$

the p_lth car is required to satisfy

$$\frac{dx_{p_l}}{dt} = \min\left(\frac{u_{p_l}(t_{p_l})(l - x_{p_l})^{1/2}}{2(\lambda L)^{1/2}},\ U_{p_l}\right) \stackrel{def}{=} u_{p_l}, \tag{1.31}$$

where

$$\frac{dU_{p_l}}{dt} = \mathcal{V}'(x_{p_l+1} - x_{p_l})(u_{p_l+1} - U_{p_l}) + \frac{(\mathcal{V}(x_{p_l+1} - x_{p_l}) - U_{p_l})}{\epsilon} \tag{1.32}$$

and

$$x_{p_l}(t_{p_l}) = l - \lambda L \ \text{ and } \ U_{p_l}(t_{p_l}) = u_{p_l}(t_{p_l}). \tag{1.33}$$

When (1.31) reduces to

$$\frac{dx_{p_l}}{dt} = \frac{u_{p_l}(t_{p_l})(l - x_{p_l})^{1/2}}{2(\lambda L)^{1/2}} \stackrel{def}{=} v_{p_l}, \tag{1.34}$$

we see that

$$\frac{dv_{p_l}}{dt} = -\frac{u_{p_l}^2(t_{p_l})}{2\lambda L} \leq -\frac{\mathcal{V}_\infty^2}{2\lambda L}, \tag{1.35}$$

and thus we apply this constant braking strategy over $t_{p_l} \leq t \leq t_{p_l} + \frac{2\lambda L}{u_{p_l}(t_{p_l})}$ and the strategy $x_{p_l}(t) = l$ over $t_{p_l} + \frac{2\lambda L}{u_{p_l}(t_{p_l})} \leq t \leq t_{m+1}$.

If instead of (1.30) the p_lth car satisfies

$$u_{p_l}(t_{p_l})(t_{m+1} - t_{p_l})/2 \leq \lambda L, \tag{1.36}$$

the p_lth car is required to satisfy

$$\begin{aligned} \tfrac{dx_{p_l}}{dt} &= \min\left(u_{p_l}(t_{p_l}) + \tfrac{2\lambda L - u_{p_l}(t_{p_l})(t_{m+1}-t_{p_l})}{(t_{m+1}-t_{p_l})^2}(t - t_{p_l}), U_{p_l}\right) \\ &\overset{def}{=} u_{p_l},\ t_{p_l} \leq t \leq t_{m+1} \end{aligned} \tag{1.37}$$

and (1.33), and again U_{p_l} satisfies (1.32) and $(1.33)_2$.

The dynamics for U_{p_l} postulated in (1.28) and (1.32) might seem a bit strange. What we are insisting is that the p_lth car must travel no faster than the minimum of its braking speed and the speed that it would travel at if it disregarded the light and allowed its velocity to be determined by the car ahead. The latter speed U_{p_l} is computed from the standard dynamics equation (see (1.6), (1.7), (1.7a), and (1.7b)).

If there is no such time $t_{p_l} < t_r$ so that $x_{p_l}(t_{p_l}) = l - \lambda L$, then we know that $x_{p_l}(t_r) \leq l - \lambda L$. In this situation we replace t_{p_l} in (1.27)–(1.37) by t_r and the terms λL in all inequalities and identities by $l - x_{p_l}(t_r)$.

Finally, if (1.26) does not hold, i.e., if

$$l - \lambda L \leq x_{p_l}(t_y) < l, \tag{1.38}$$

we set t_{p_l} to t_y in (1.27)–(1.37) and replace λL in these formulas by $l - x_{p_l}(t_y)$.

The rules when $p_l = k_l$ are similarly amended.

The cars with indices $p_{l-1} \leq k \leq p_l - 1$ are required to satisfy their standard dynamics over $[t_y, t_{m+1}]$.

Our first result deals with the model's consistency; we shall show that for all $t \geq 0$ and all indices, $L \leq (x_{k+1} - x_k)(t)$ and $0 \leq u_k(t) < \mathcal{V}((x_{k+1} - x_k)(t))$. We also have the theorem that no cars run any red lights. With two in-phase lights, the number of cars through an intersection during the green and yellow phases is, after a start-up period, a constant. This constant is less than the constant obtained with models which allow for infinite accelerations, i.e., discrete Lagrangian versions of the Lighthill–Whitham–Richards model [4, 5, 6, 7].

One surprising observation about the model just described is that the largest decelerations are not necessarily associated with the cars indexed by p_l but rather cars with indices $k \leq p_l - 1$ which are forced to slow down because the p_lth car has stopped. Equation (1.7a) implies that the latter cars' decelerations are determined by the negative velocity gradients $u_{k+1} - u_k$.

Finally, we note that though we have been quite specific in postulating our stopping rules for the p_lth car, it would have sufficed to have chosen any rule of the form

$$\frac{dx_{p_l}}{dt} = \min\left(v_{p_l}, U_{p_l}\right) \overset{def}{=} u_{p_l}, \qquad t_{p_l} \leq t \leq t_{m+1},$$

where U_{p_l} satisfies

$$\frac{dU_{p_l}}{dt} = \mathcal{V}'(x_{p_{l+1}} - x_{p_l})(u_{p_{l+1}} - U_{p_l}) + (\mathcal{V}(x_{p_{l+1}} - x_{p_l}) - U_{p_l})\epsilon$$

and $U_{p_l}(t_y) = u_{p_l}(t_y)$ if $p_l \leq N-1$, and

$$\frac{dU_N}{dt} = \frac{(\mathcal{V}_\infty - U_N)}{\epsilon} \text{ and } U_N(t_y) = u_{p_l}(t_y)$$

if $p_l = N$, and where $v_{p_l} \geq 0$ is chosen so that if

$$\frac{dx_{p_l}}{dt} = v_{p_l}, \qquad t_y \leq t \leq t_{m+1} \text{ and } x_{p_l}(t_y) < l,$$

then $x_{p_l}(t) \leq l$, $t_y \leq t \leq t_{m+1}$.

2. Model consistency. In this section we turn our attention to the issue of model consistency. The central issue before us is to show that for $1 \leq k \leq N-1$ and $0 \leq t$

$$L \leq (x_{k+1} - x_k)(t) \text{ and } 0 \leq u_k(t) < \mathcal{V}((x_{k+1} - x_k)(t)) \tag{2.1}$$

and that for $k = N$ and $0 \leq t$

$$0 \leq u_N(t) \leq \mathcal{V}_\infty. \tag{2.2}$$

We are also interested in knowing that the distinguished cars indexed by p_l do not run the red lights over the intervals $t_r \overset{def}{=} (m-1)(TG+TY+TR)+TG+TY \leq t \leq m(TG+TY+TR) \overset{def}{=} t_{m+1}$ and that the (p_l+1)st car clears the intersection by t_r, i.e., satisfies

$$x_{p_l+1}(t_r) \geq l + w + L. \tag{2.3}$$

Once again $x = l$ is supposed to be the leading edge of the intersection, w the width of the intersection, and L the length of a car.

There are two natural approaches that one can take to establish the above claims. The first is to show that the desired conclusions follow directly from the governing differential equations and initial and constraining conditions while the second is to show that approximate solutions, generated by numerical discretization, satisfy the desired consistency results. Noting then that these consistency results are sufficient to guarantee that the approximate solutions converge (as $\Delta t \to 0$) to solutions of the original model, we are guaranteed that these limiting solutions satisfy the same consistency results. We adopt the latter procedure here since in the next section we shall perform computations with the discrete approximating system.

Throughout, Δt will denote our time step and the quantities $(x_k^n, u_k^n, \alpha_k^n)$ will denote the values of the approximate solutions at $t_n = n\Delta t$. To keep matters simple we shall assume that the numbers $TG/\Delta t$, $TY/\Delta t$, $TR/\Delta t$, and $\epsilon/\Delta t$ are all integers and we shall assume that $\Delta t \leq \min(\epsilon, (\mathcal{V}'(L) = \max_{L \leq s} \mathcal{V}'(s))^{-1})$.

Our first result deals with the traffic flow over the time intervals

$$t_m \overset{def}{=} (m-1)(TG+TY+TR) \leq t_n = n\Delta t \leq t_y \overset{def}{=} t_m + TG \tag{2.4}$$

when all lights are green. Over such intervals we replace (1.5) by

$$x_k^{n+1} = x_k^n + u_k^n \Delta t, \qquad 1 \le k \le N, \tag{2.5}$$

and this yields

$$s_k^{n+1} = s_k^n + (u_{k+1}^n - u_k^n)\Delta t, \qquad 1 \le k \le N-1, \tag{2.6}$$

where

$$s_k^n = (x_{k+1}^n - x_k^n) \ \text{ and } \ s_k^{n+1} = (x_{k+1}^{n+1} - x_k^{n+1}). \tag{2.7}$$

The u's and s's are related by

$$u_k^n = \mathcal{V}(s_k^n) + \alpha_k^n \tag{2.8}$$

and

$$u_k^{n+1} = \mathcal{V}(s_k^{n+1}) + \left(1 - \frac{\Delta t}{\epsilon}\right) \alpha_k^n. \tag{2.9}$$

These updates hold for indices n satisfying

$$(m-1)(TG + TY + TR)/\Delta t \stackrel{def}{=} n_m \le n \le n_m + TG/\Delta t - 1. \tag{2.10}$$

THEOREM 1. *Suppose that*

$$L \le s_k^{n_m} \quad \textit{and} \quad 0 \le u_k^{n_m} \le \mathcal{V}(s_k^{n_m}), \qquad 1 \le k \le N-1, \tag{2.11}$$

and

$$0 \le u_N^{n_m} \le \mathcal{V}_\infty = \lim_{s \to \infty} \mathcal{V}(s). \tag{2.12}$$

Then, the same inequalities hold for

$$n_m \le n \le n_m + TG/\Delta t \stackrel{def}{=} n_y. \tag{2.13}$$

Proof. The identity (2.6) implies that if $s_k^n \ge L$ and $u_{k+1}^n - u_k^n \ge 0$, then $s_k^{n+1} \ge s_k^n \ge L$. In the situation in which $u_{k+1}^n - u_k^n < 0$, (2.6) implies that

$$s_k^{n+1} = s_k^n + (u_{k+1}^n - \alpha_k^n - \mathcal{V}(s_k^n))\Delta t \tag{2.14}$$

and the natural induction hypotheses $\alpha_k^n \le 0$, $0 \le u_k^n \le \mathcal{V}(s_k^n)$, and $s_k^n \ge L$ imply that $u_{k+1}^n - \alpha_k^n \ge 0$. In the situation in which $0 \le u_{k+1}^n - \alpha_k^n < \mathcal{V}_\infty$ we are guaranteed a unique $\bar{s}_{k+1}^n \in [L, \infty)$ satisfying

$$u_{k+1}^n - \alpha_k^n = \mathcal{V}(\bar{s}_{k+1}^n), \tag{2.15}$$

and here (2.14) reduces to

$$s_k^{n+1} = s_k^n + \left(\mathcal{V}(\bar{s}_{k+1}^n) - \mathcal{V}(s_k^n)\right) \Delta t \tag{2.16}$$

or

$$s_k^{n+1} = (1 - \mathcal{V}'(s_*)\Delta t) s_k^n + \mathcal{V}'(s_*)\Delta t \bar{s}_k^n \tag{2.17}$$

for some $s_* \in (\min(s_k^n, \bar{s}_{k+1}^n),\ \max(s_k^n, \bar{s}_{k+1}^n))$. The latter identity, together with

$$\Delta t \mathcal{V}'(L) \le 1 \quad \text{and} \quad \min(s_k^n, \bar{s}_{k+1}^n) \ge L, \tag{2.18}$$

yields $s_k^{n+1} \ge L$. When $u_{k+1}^n - u_k^n < 0$ and $u_{k+1}^n - \alpha_k^n \ge \mathcal{V}_\infty$, the identity (2.14) implies that

$$s_k^{n+1} \ge s_k^n + (\mathcal{V}_\infty - \mathcal{V}(s_k^n))\Delta t. \tag{2.19}$$

The inequality $(2.18)_1$ guarantees that $s \to s + (\mathcal{V}_\infty - \mathcal{V}(s))\Delta t$ is strictly increasing on $[L, \infty)$ and thus (2.19) implies that $s_k^{n+1} \ge L + \mathcal{V}_\infty \Delta t \ge L$ as desired.

The induction hypothesis $\alpha_k^n \le 0$ together with $\Delta t/\epsilon \le 1$ and (2.9) guarantees that $u_k^{n+1} \le \mathcal{V}(s_k^{n+1})$. What remains to be shown is that $u_k^{n+1} \ge 0$. To establish this assertion we combine (2.8) and (2.9) to obtain

$$u_k^{n+1} = \mathcal{V}(s_k^n + (u_{k+1}^n - u_k^n)\Delta t) + \left(1 - \frac{\Delta t}{\epsilon}\right)(u_k^n - \mathcal{V}(s_k^n)).$$

Noting that

$$\mathcal{V}(s_k^n + (u_{k+1}^n - u_k^n)\Delta t) = \mathcal{V}(s_k^n) + \mathcal{V}'(s_\#)(u_{k+1}^n - u_k^n)\Delta t$$

for some $s_\# \ge L$, we find that

$$u_k^{n+1} = \mathcal{V}'(s_\#)\Delta t u_{k+1}^n + \frac{\Delta t}{\epsilon}(\mathcal{V}(s_k^n) - u_k^n) + (1 - \mathcal{V}'(s_\#)\Delta t)u_k^n.$$

The last identity, when combined with

$$\Delta t \mathcal{V}'(s_\#) \le 1, \quad \Delta t/\epsilon \le 1, \quad u_k^n \ge 0, \quad u_{k+1}^n \ge 0, \quad \text{and} \quad \mathcal{V}(s_k^n) - u_k^n \ge 0,$$

yields $u_k^{n+1} \ge \min(u_k^n, u_{k+1}^n) \ge 0$ as desired. □

We now turn our attention to what happens over the yellow and red phases, i.e., when

$$t_y \overset{def}{=} (m-1)(TG + TY + TR) + TG \le t_n = n\Delta t < t_{m+1} \overset{def}{=} m(TG + TY + TR). \tag{2.20}$$

The results of Theorem 1 imply that when $n = n_y \overset{def}{=} (m-1)(TG+TY+TR)+TG/\Delta t$ the following inequalities are valid:

$$L \le s_k^{n_y} \quad \text{and} \quad 0 < u_k^{n_y} \le \mathcal{V}(s_k^{n_y}), \qquad 1 \le k \le N-1, \tag{2.21}$$

and

$$0 \le u_N^{n_y} \le \mathcal{V}_\infty = \lim_{s \to \infty} \mathcal{V}(s). \tag{2.22}$$

Our next goal is to show that (2.21) and (2.22) hold for indices

$$n_y \le n \le n_{m+1} \overset{def}{=} m(TG + TY + TR). \tag{2.23}$$

For definiteness we assume the lights are located at $l_1 < l_2 < \cdots < l_M$ where $M << N$ and that $L << l_{I+1} - l_I$, $1 \le I \le M-1$. For $1 \le I \le M$, k_I will be the largest integer less than or equal to N, so that

$$x_{k_I}^{n_y} < l_I \tag{2.24}$$

and p_I will be the largest integer less than or equal to k_I so that

$$x_{p_I}^{n_y} + \left(\min_{p_I \le j \le k_I} u_j^{n_y} \right) TY < l_I + w + L. \tag{2.25}$$

It can and does happen that for some $I < M$

$$p_I = p_{I+1} = \cdots = p_M = N. \tag{2.26}$$

Our first task is to establish the desired inequalities for indices $(p_{I-1}+1) \le k \le p_I = N$ for $n_y \le n \le n_{m+1}$. This is the situation that is obtained when the lead car, indexed by N, has passed the $(I-1)$st light but not the Ith light.

The rules laid out in (1.17)–(1.23) imply that $x_N(\cdot)$ satisfies

$$\frac{dx_N}{dt} = \min\,(v_N, U_N) \stackrel{def}{=} u_N, \qquad t_y \le t \le t_{m+1}, \tag{2.27}$$

where U_N satisfies

$$\frac{dU_N}{dt} = \frac{(\mathcal{V}_\infty - U_N)}{\epsilon} \;\text{ and }\; U_N(t_y) = u_N(t_y), \tag{2.28}$$

and $v_N(\cdot) \ge 0$ is chosen so that if $x_N(\cdot)$ satisfies

$$\frac{dx_N}{dt} = v_N \;\text{ and }\; x_N(t_y) < l_I, \tag{2.29}$$

then $x_N(t_{m+1}) \le l_I$. We replace this system with its discrete analogue,

$$x_N^{n+1} = x_N^n + u_N^n \Delta t, \qquad n_y \le n \le n_{m+1} - 1, \tag{2.30}$$

$$U_N^{n+1} = \mathcal{V}_\infty + \left(1 - \frac{\Delta t}{\epsilon}\right)(U_N^n - \mathcal{V}_\infty), \qquad n_y \le n \le n_{m+1} - 1, \tag{2.31}$$

and these are solved subject to the initial conditions

$$x_N^{n_y} < l_I \;\text{ and }\; 0 \le u_N^{n_y} \le U_N^{n_y} \le \mathcal{V}_\infty. \tag{2.32}$$

The discrete velocity u_N^n is given by

$$u_N^n = \min\,(v_N^n, U_N^n), \tag{2.33}$$

and $v_N^n \ge 0$ is a discretization of v_N with the property that if

$$x_N^{n+1} = x_N^n + v_N^n \Delta t \;\text{ and }\; x_N^{n_y} < l_I \tag{2.34}$$

for $n_y \le n \le n_{m+1} - 1$, then

$$x_N^{n_{m+1}} \le l_I. \tag{2.35}$$

The identities (2.31), $(2.32)_2$, and (2.33) guarantee that

$$0 \le u_N^n \le \mathcal{V}_\infty, \qquad n_y \le n \le n_{m+1}. \tag{2.36}$$

If we assume that $(p_{I-1}+1) \le N-1$, then the $(N-1)$st car will follow the standard dynamics (1.5)–(1.7) on $t_y \le t \le t_{m+1}$, and thus for $n_y \le n \le n_{m+1}-1$ we have the approximating discrete system:

$$\begin{aligned} x_{N-1}^{n+1} &= x_{N-1}^n + u_{N-1}^n \Delta t, \qquad u_{N-1}^n = \mathcal{V}(s_{N-1}^n) + \alpha_{N-1}^n, \\ \text{and} \quad u_{N-1}^{n+1} &= \mathcal{V}(s_{N-1}^{n+1}) + \left(1 - \tfrac{\Delta t}{\epsilon}\right) \alpha_{N-1}^n, \end{aligned} \tag{2.37}$$

where

$$s_{N-1}^n = x_N^n - x_{N-1}^n \quad \text{and} \quad s_{N-1}^{n+1} = x_N^{n+1} - x_{N-1}^{n+1} = s_{N-1}^n + (u_N^n - u_{N-1}^n)\Delta t. \tag{2.38}$$

The inequalities (2.21) and (2.22) imply that $\alpha_{N-1}^{n_y} \le 0$, $\alpha_N^{n_y} \le 0$, and $s_{N-1}^{n_y} \ge L$. The identities (2.37) and (2.38) imply that

$$s_{N-1}^{n+1} = s_{N-1}^n + \left(u_N^n - \left(1 - \frac{\Delta t}{\epsilon}\right)^n \alpha_{N-1}^{n_y} - \mathcal{V}(s_{N-1}^n) \right) \Delta t, \tag{2.39}$$

and $(2.37)_2$ and (2.39), together with

$$L \le s_{N-1}^{n_y}, \quad \alpha_{N-1}^{n_y} \le 0, \quad u_N^n \ge 0, \quad \Delta t \mathcal{V}'(L) \le 1, \quad \text{and} \quad \Delta t \le \epsilon \tag{2.40}$$

and the arguments used to establish Theorem 1, imply that

$$L \le s_{N-1}^n, \qquad n_y \le n \le n_{m+1}. \tag{2.41}$$

The arguments used to establish Theorem 1 along with (2.40) and (2.41) also yield $0 \le u_{N-1}^n \le \mathcal{V}(s_{N-1}^n)$, $n_y \le n \le n_{m+1}$. An induction on k for indices $(p_{I-1}+1) \le k$ then yields

$$L \le s_k^n = (x_{k+1}^n - x_k^n) \quad \text{and} \quad 0 \le u_k^n \le \mathcal{V}(s_k^n), \qquad n_y \le n \le n_{m+1}. \tag{2.42}$$

This situation when $p_{I-1} = N-1$ is handled similarly, provided that one adopts the proper first order integration scheme for U_{N-1}. The governing equation for U_{N-1} is

$$\frac{dU_{N-1}}{dt} = \mathcal{V}'(x_N - x_{N-1})(u_N - U_{N-1}) + \frac{(\mathcal{V}(x_N - x_{N-1}) - U_{N-1})}{\epsilon}, \tag{2.43}$$

where

$$\frac{d(x_N - x_{N-1})}{dt} = u_N - u_{N-1}, \tag{2.44}$$

and $v_{N-1} \ge 0$ is chosen so that if

$$\frac{dx_N}{dt} = v_{N-1} \quad \text{and} \quad x_{N-1}(t_y) < l_I, \tag{2.45}$$

then

$$x_{N-1}(t_{m+1}) \le l_I. \tag{2.46}$$

Additionally

$$u_{N-1} \stackrel{def}{=} \min(v_{N-1}, U_{N-1}). \tag{2.47}$$

The integration scheme we use is

$$U_{N-1}^{n+1} = \mathcal{V}(s_{N-1}^n + (u_N^n - U_{N-1}^n)\Delta t) + \left(1 - \frac{\Delta t}{\epsilon}\right)(U_{N-1}^n - \mathcal{V}(s_{N-1}^n)),^3 \tag{2.48}$$

where

$$s_{N-1}^n = x_N^n - x_{N-1}^n. \tag{2.49}$$

To complete the proof one does an induction on the index I, first replacing I by $I-1$. One knows that the car with index $(p_{I-1}+1)$ has a velocity $u_{(p_{I-1}+1)}^n$ satisfying

$$0 \le u_{(p_{I-1}+1)}^n \le \mathcal{V}(s_{p_{I-1}+1}^n), \qquad n_y \le n \le n_{m+1}. \tag{2.50}$$

We first focus on the p_{I-1}st car and note that

$$\frac{dx_{p_{I-1}}}{dt} = \min\,(v_{p_{I-1}}, U_{p_{I-1}}) \stackrel{def}{=} u_{p_{I-1}}, \tag{2.51}$$

and

$$\frac{ds_{p_{I-1}}}{dt} = (u_{(p_{I-1}+1)} - u_{p_{I-1}}). \tag{2.52}$$

The rules laid out in (1.7)–(1.23) imply that

$$\frac{dU_{p_{I-1}}}{dt} = \mathcal{V}'(s_{(p_{I-1})})(u_{(p_{I-1}+1)} - U_{p_{I-1}}) + \frac{(\mathcal{V}(s_{(p_{I-1}+1)}) - U_{p_{I-1}})}{\epsilon} \tag{2.53}$$

and that the velocity field $0 \le v_{p_{I-1}}$ is chosen so that if $x_{p_{I-1}}$ evolves as

$$\frac{dx_{p_{I-1}}}{dt} = v_{p_{I-1}} \quad \text{and} \quad x_{p_{I-1}}(t_y) < l_I, \tag{2.54}$$

then

$$x_{p_{I-1}}(t_{m+1}) \le l_{I-1}. \tag{2.55}$$

The discretization we apply to the p_{I-1}st car is

$$x_{p_{I-1}}^{n+1} = x_{p_{I-1}}^n + u_{p_{I-1}}^n \Delta t \quad \text{and} \quad s_{p_{I-1}}^{n+1} = s_{p_{I-1}}^n + \left(u_{(p_{I-1}+1)}^n - u_{p_{I-1}}^n\right)\Delta t \tag{2.56}$$

for $n_y \le n \le n_{m+1} - 1$. Moreover, for some $n_y \le n_0 \le n_y + TY/\Delta t - 1$

$$u_{p_{I-1}}^{n+1} = \mathcal{V}(s_{p_{I-1}}^{n+1}) + \left(1 - \frac{\Delta t}{\epsilon}\right)(u_{p_{I-1}}^n - \mathcal{V}(s_{p_{I-1}}^n)) \tag{2.57}$$

and

$$U_{p_{I-1}}^{n+1} = u_{p_{I-1}}^{n+1}, \tag{2.58}$$

[3]This scheme is essentially a first order Euler scheme applied to (2.43). The scheme implies that

$$U_{N-1}^{n+1} = U_{N-1}^n + \Delta\, t\mathcal{V}'\left(s_{N-1}^n\right)\left(u_N^n - U_{N-1}^n\right) + \frac{\Delta t}{\epsilon}\left(\mathcal{V}\left(s_{N-1}^n\right) - U_{N-1}^n\right) + 0(\Delta t)^2.$$

whereas for $n_0 \le n \le n_{m+1} - 1$

$$u^n_{p_{I-1}} = \min\,(v^n_{p_{I-1}}, U^n_{p_{I-1}}), \tag{2.59}$$

$$U^{n+1}_{p_{I-1}} = \mathcal{V}(s^n_{p_{I-1}} + (u^n_{(p_{I-1}+1)} - U^n_{p_{I-1}})\Delta t) + \left(1 - \frac{\Delta t}{\epsilon}\right)(U^n_{p_{I-1}} - \mathcal{V}(s^n_{p_{I-1}})),^4 \tag{2.60}$$

and

$$U^{n_0}_{p_{I-1}} = u^{n_0}_{p_{I-1}} \quad \text{and} \quad x^{n_0}_{p_{I-1}} < l_I. \tag{2.61}$$

Finally $v^n_{p_{I-1}}$ is chosen so that if

$$x^{n+1}_{p_{I-1}} = x^n_{p_{I-1}} + v^n_{p_{I-1}}\Delta t, \qquad n_0 \le n \le n_{m+1} - 1, \tag{2.62}$$

then

$$x^{n_{m+1}}_{p_{I-1}} \le l_I. \tag{2.63}$$

The arguments employed to establish Theorem 1 guarantee that for $n_y \le n \le n_0$

$$L \le s^n_{p_{I-1}} \quad \text{and} \quad 0 \le u^n_{p_{I-1}} \le \mathcal{V}(s^n_{p_{I-1}}) \tag{2.64}$$

and that for $n = n_0$

$$0 \le u^{n_0}_{p_{I-1}} \le U^{n_0}_{p_{I-1}} \le \mathcal{V}(s^{n_0}_{p_{I-1}}). \tag{2.65}$$

LEMMA 1. *For* $n_0 \le n \le n_{m+1}$

$$L \le s^n_{p_{I-1}} \quad \textit{and} \quad 0 \le u^n_{p_{I-1}} \le U^n_{p_{I-1}} \le \mathcal{V}(s^n_{p_{I-1}}). \tag{2.66}$$

Proof. The identities (2.56) and (2.60) imply that

$$\begin{aligned}
\mathcal{V}(s^{n+1}_{p_{I-1}}) - U^{n+1}_{p_{I-1}} &= \mathcal{V}(s^n_{p_{I-1}} + (u^n_{(p_{I-1}+1)} - u^n_{p_{I-1}})\Delta t) \\
&\quad - \mathcal{V}(s^n_{p_{I-1}} + (u^n_{(p_{I-1}+1)} - U^n_{p_{I-1}})\Delta t) \\
&\quad + \left(1 - \frac{\Delta t}{\epsilon}\right)(\mathcal{V}(s^n_{p_{I-1}}) - U^n_{p_{I-1}}) \\
&= \Delta t \mathcal{V}'(s_\#)\left(U^n_{p_{I-1}} - u^n_{p_{I-1}}\right) + \left(1 - \frac{\Delta t}{\epsilon}\right)(\mathcal{V}(s^n_{p_{I-1}}) - U^n_{p_{I-1}})
\end{aligned} \tag{2.67}$$

for some $s_\# \ge \min\,(s^n_{p_{I-1}} + (u^n_{(p_{I-1}+1)} - u^n_{p_{I-1}})\Delta t, s^n_{p_{I-1}} + (u^n_{(p_{I-1}+1)} - U^n_{p_{I-1}})\Delta t)$. If we now make the induction hypotheses that

$$L \le s^n_{p_{I-1}} \quad \text{and} \quad 0 \le U^n_{p_{I-1}} \le \mathcal{V}(s^n_{p_{I-1}}), \tag{2.68}$$

then (2.59) implies that

$$0 \le u^n_{p_{I-1}} \le U^n_{p_{I-1}} \le \mathcal{V}(s^n_{p_{I-1}}) \tag{2.69}$$

[4] See Footnote 3.

and (2.69) and (2.42) with $k = p_{I-1} + 1$ implies that

$$\begin{aligned} &\min\left(s^n_{p_{I-1}} + (u^n_{(p_{I-1}+1)} - u^n_{p_{I-1}})\Delta t,\ s^n_{p_{I-1}} + (u^n_{(p_{I-1}+1)} - U^n_{p_{I-1}})\Delta t\right) \\ &\geq s^n_{p_{I-1}} - \mathcal{V}(s^n_{p_{I-1}})\Delta t \stackrel{def}{=} \mathcal{F}(s^n_{p_{I-1}}). \end{aligned} \tag{2.70}$$

This constraint $\Delta t\mathcal{V}'(s) \leq 1$, $L \leq s$ guarantees $\mathcal{F}(\cdot)$ in nondecreasing on $L \leq s$, and this fact, together with $\mathcal{F}(L) = L$, guarantees that $s^{n+1}_{p_{I-1}}$ and $s_\#$ are both greater than or equal to L. Moreover, (2.67) also yields $U^{n+1}_{p_{I-1}} \leq \mathcal{V}(s^n_{p_{I-1}})$. The defining relation (2.60) and (2.70) and $u^n_{(p_{I-1}+1)} \geq 0$ also implies that

$$U^{n+1}_{p_{I-1}} = \Delta t\mathcal{V}'(s_*)u^n_{(p_{I-1}+1)} + (1 - \Delta t\mathcal{V}'(s_*))U^n_{p_{I-1}} + \frac{\Delta t}{\epsilon}(\mathcal{V}(s^n_{p_{I-1}}) - U^n_{p_{I-1}}) \tag{2.71}$$

for some $s_* \geq L$ and (2.71) guarantees that $U^{n+1}_{p_{I-1}} \geq 0$. The last inequality and (2.59), with $n+1$, guarantees that

$$0 \leq u^{n+1}_{p_{I-1}} \leq U^{n+1}_{p_{I-1}} \leq \mathcal{V}(s^{n+1}_{p_{I-1}}), \tag{2.72}$$

and this completes the proof of Lemma 1. ∎

Once again an induction on k for indices $(p_{I-2} + 1) \leq k$ yields

$$L \leq s^n_k = (x^n_{k+1} - x^n_k) \ \text{ and } \ 0 \leq u^n_k \leq \mathcal{V}(s^n_k) \tag{2.73}$$

and additionally yields the following theorem.

THEOREM 2. *For $n_y \leq n \leq n_{m+1} = m(TG + TY + TR)$*

$$L \leq s^n_k \ \textit{and} \ \ 0 \leq u^n_k \leq \mathcal{V}(s^n_k), \qquad 1 \leq k \leq N-1, \tag{2.74}$$

and

$$0 \leq u^n_N \leq \mathcal{V}_\infty = \lim_{s\to\infty} \mathcal{V}(s). \tag{2.75}$$

Moreover, for $1 \leq I \leq M$

$$x^{n_{m+1}}_{p_I} \leq l_I. \qquad \square \tag{2.76}$$

Theorems 1 and 2 go a long way towards establishing the consistency of our model. What remains to be shown is that cars with index $p_I + 1$ clear the light, i.e., that they satisfy

$$x^{n_y + \frac{TY}{\Delta t}}_{(p_I+1)} \geq l_I + w + L. \tag{2.77}$$

The reader should recall that the cars with these indices satisfy

$$x^{n_y}_{(p_{I+1})} < l_I \ \text{ and } \ x^{n_y}_{(p_I+1)} + \left(\min_{(p_I+1)\leq j\leq k_I} u^{n_y}_j\right) TY \geq l_I + w + L \tag{2.78}$$

and that cars with indices $(p_I + 1) \leq k \leq k_I$ evolve by the standard discrete dynamics for $n_y \leq n \leq n_y + TY/\Delta t - 1$; i.e.,

$$x^{n+1}_k = x^n_k + u^n_k\Delta t \ \text{ and } \ u^n_k = \mathcal{V}(s^n_k) + \left(1 - \frac{\Delta t}{\epsilon}\right)^{n-n_y}(u^{n_y}_k - \mathcal{V}(s^{n_y})),$$

where

$$0 \le u_k^{n_y} \le \mathcal{V}(s_k^{n_y}) \quad \text{and} \quad L \le s_k^n.$$

It is a straightforward calculation to show that cars with these indices also satisfy

$$u_k^{n+1} = \mathcal{V}(s_k^n + (u_{k+1}^n - u_k^n)\Delta t) + \left(1 - \frac{\Delta t}{\epsilon}\right)(u_k^n - \mathcal{V}(s_k^n))$$
$$= \Delta t \mathcal{V}'(s_\#) u_{k+1}^n + (1 - \Delta t \mathcal{V}'(s_\#)) u_k^n + \frac{\Delta t}{\epsilon}(\mathcal{V}(s_k^n) - u_k^n)$$

from some $s_\# \ge L$, and that this identity, along with

$$\Delta t \mathcal{V}'(L) \le 1, \qquad \Delta t \le \epsilon, \quad \text{and} \quad 0 \le \mathcal{V}(s_k^n) - u_k^n$$

implies

$$u_k^{n+1} \ge \min\,(u_k^n, u_{k+1}^n). \tag{2.79}$$

We now note that at $t = t_y$ (equivalently $n = n_y$) the cars with indices $p_I \le k$ typically satisfy

$$\min_{p_I \le j \le k_I} u_j^{n_y} = u_{k_0}^{n_y}, \quad \text{where} \quad (p_I + 1) \le k_0 \le k_I, \tag{2.80}$$

and

$$u_{k+1}^{n_y} - u_k^{n_y} \ge 0, \qquad k_0 \le k \le k_\#, \tag{2.81}$$

where $k_\#$ is greater than k_I. Moreover, if the spacing of the lights is sufficiently large, then the spatial monotonicity of the velocities is preserved for $n_y \le n \le n_y + TY/\Delta t$ and $k_0 \le k \le k_\#$. When this is the case, the inequalities (2.78)–(2.81) guarantee (2.77).

3. Simulations. In this section we present some simulations of the system outlined in section 1. We chose

$$\mathcal{V}_\infty = 50f/s, \quad L = 20f, \quad L_1 = 25f, \quad \lambda = 5, \quad \epsilon = 5s \text{ and } N = 600.$$

Our maximal velocity was given by

$$\mathcal{V}(s) = \mathcal{V}_\infty \left(1 - \frac{L}{s}\right), \qquad L \le s.$$

We restrict our attention to a roadway with two in-phase lights located at

$$l_1 = 1 \text{ mile } = 5280f \quad \text{and} \quad l_2 = 2 \text{ miles } = 10,560f,$$

and we assume that the width of each intersection is

$$w = 20f.$$

Finally, the durations of the green, yellow, and red lights were chosen to be

$$TG = 25s, \qquad TY = 5s, \quad \text{and} \quad TR = 30s.$$

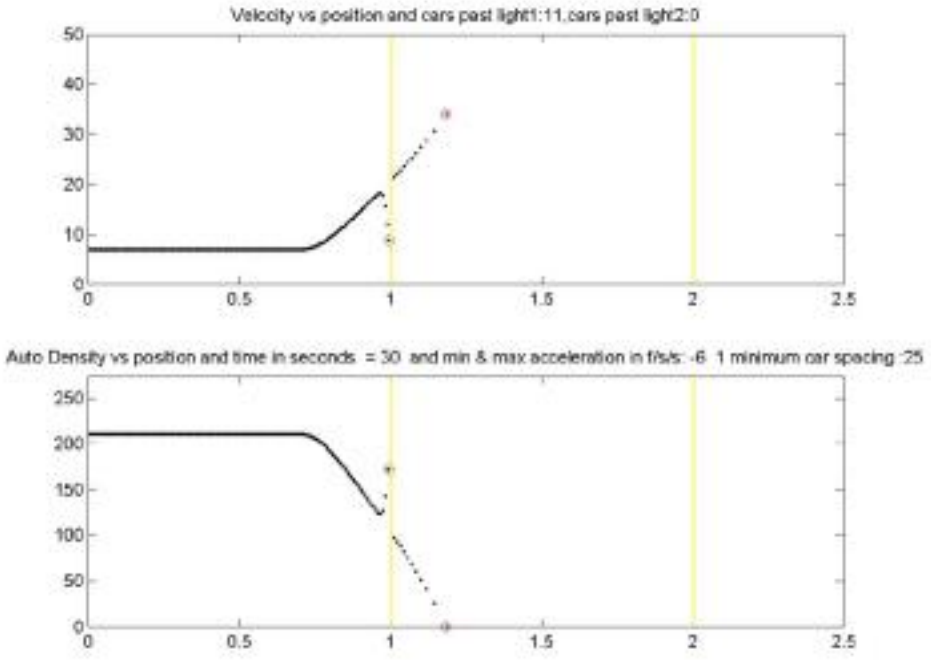


FIG. 1.

Our initial data are taken to be

$$x_k(0) = 25(k - 400) \quad \text{and} \quad u_k(0) = 0,\ 1 \le k \le 600.$$

Snapshots of the solution are shown at times 30, 147, 151, 179, and 191 seconds in Figures 1–5, respectively, and a film may be seen at http://www.math.cmu.edu/users/plin/21380/traffic.html.

In the first frame of each snapshot we plot the auto velocity u_k (in miles/hour) versus current auto position x_k (in miles), and in the second frame we plot the empirical density $\rho_k = \frac{1}{x_{k+1}-x_k}$ (in cars/mile) versus current auto position x_k (in miles).

After an initial startup period we are able to get 18 cars through each light during each green-yellow-red cycle. This number should be contrasted with what one obtains in the singular limit, where $\epsilon = 0^+$, $TY = 0s$, $TG = 30s$, $w = 0f$, and $\lambda = 5$. In this limit

$$u_k \equiv \mathcal{V}_\infty \left(1 - \frac{L}{x_{k+1} - x_k}\right),$$

and if, perchance, we have a car satisfying

$$x_k((t_m + TG)^-) = l_I, \qquad I = 1 \text{ or } 2,$$

and

$$u_k((t_m + TG)^-) > 0,$$

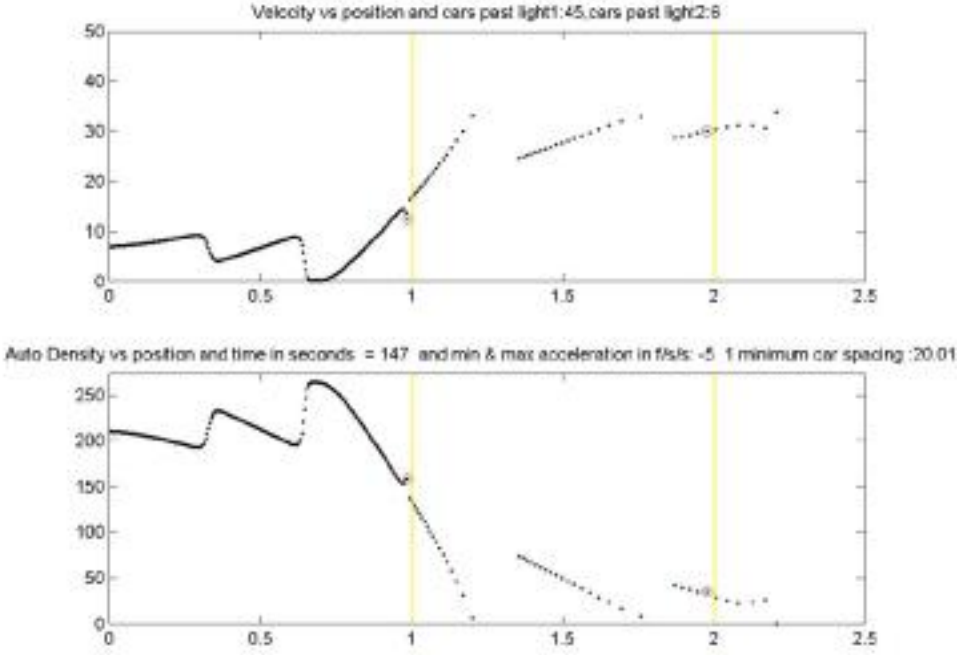

Velocity vs position and cars past light1:45.cars past light2:6
Auto Density vs position and time in seconds = 147 and min & max acceleration in ft/s/s: -5 1 minimum car spacing :20.01

FIG. 2.

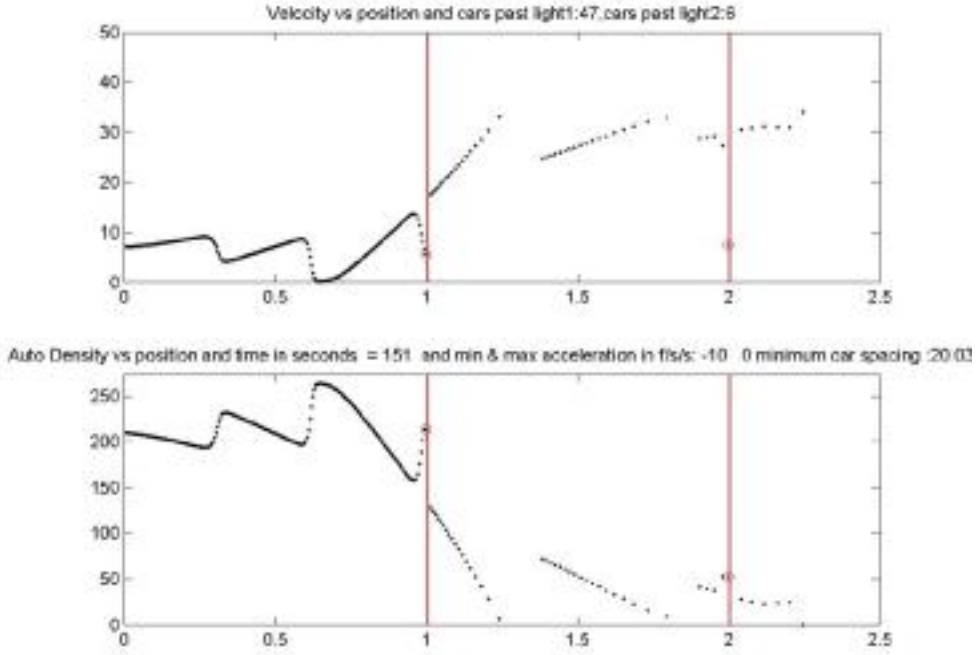

Velocity vs position and cars past light1:47.cars past light2:6
Auto Density vs position and time in seconds = 151 and min & max acceleration in ft/s/s: -10 0 minimum car spacing :20.03

FIG. 3.

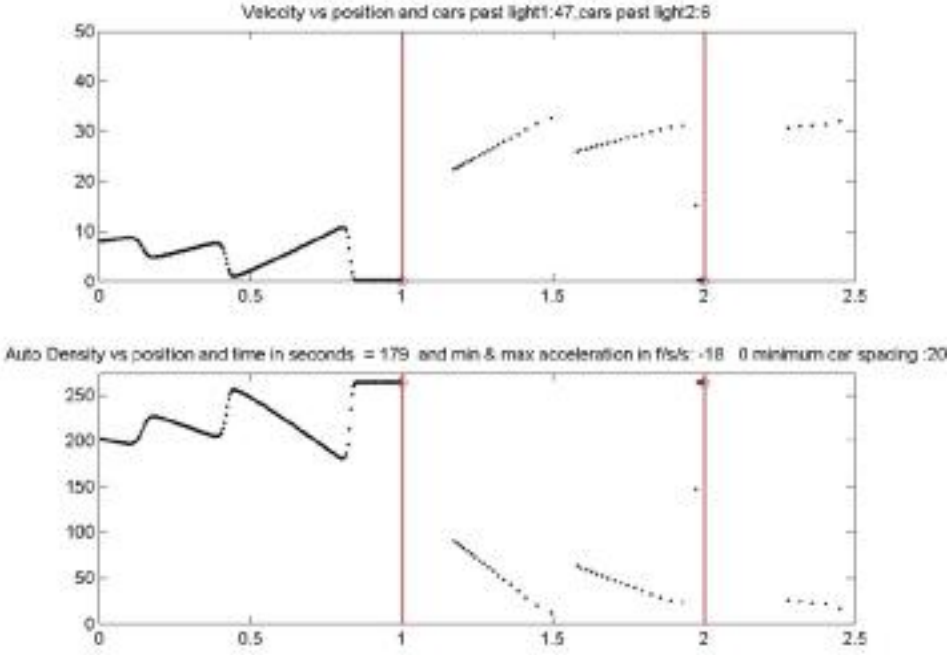

Velocity vs position and cars past light1:47,cars past light2:6
Auto Density vs position and time in seconds = 179 and min & max acceleration in f/s/s: -18 0 minimum car spacing :20

FIG. 4.

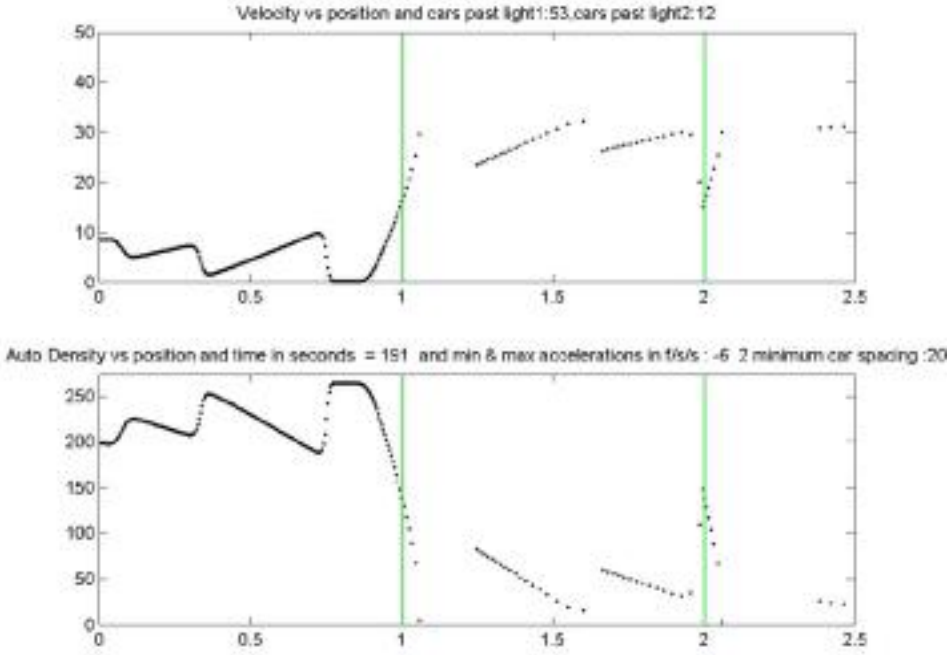

Velocity vs position and cars past light1:53,cars past light2:12
Auto Density vs position and time in seconds = 151 and min & max accelerations in f/s/s : -6 2 minimum car spacing :20

FIG. 5.

then for times $t_m + TG < t \le t_m + TG + TR$,

$$x_k(t) = l \text{ and } u_k(t) \equiv 0.$$

For this singular model we declare a car through the light at l if $x_k(t_y) > l$. The singular model has the potential for infinite accelerations. In steady state the singular model allows us to get 20 cars through an intersection during each green-red cycle.

We note that our choice of which car must stop is made at times $t_y = t_m + TG$ (when a green light turns yellow) and is conservative when the car chosen to stop satisfies $x_{p_l}(t_y) < l - \lambda L$. A more aggressive strategy would have been to allow the p_lth car to follow its standard dynamics until time $t_{p_l} < t_y + TY$, where $x_{p_l}(t_{p_l}) = l - \lambda L$, and then reevaluate whether the p_lth car can get through the light in the remaining time $t_y + TY - t_{p_l}$, i.e., check whether

$$x_{p_l}(t_{p_l}) + \min_{p_l \le k \le k_l(t_{p_l})} u_k(t_{p_l})(t_y + TY - t_{p_l}) \ge l + w + L.$$

If the latter inequality holds, the aggressive strategy would allow the p_lth car through and stop the $(p_l - 1)$st car. We avoided this strategy because it did not seem to be worth the effort to get one more car through the intersection during the green-yellow-red cycle.

The attentive reader will by now realize that once we have determined which car will slow down or stop at a given light the particular braking strategy adopted is immaterial; all that is required is that the velocity associated with the braking strategy, v_{p_l}, be such that if x_{p_l} satisfies

$$\frac{dx_{p_l}}{dt} = v_{p_l} \quad \text{and} \quad x_{p_l}(t_y) < l,$$

then $x_{p_l}(t_{m+1}) \le l$. We adopted constant braking strategies here because they were simple and realistic.

4. Concluding remarks. There are some obvious connections between the discrete model studied in this paper and the continuum or macroscopic models of Aw, Klar, Materne, and Rascle [3].

If one assumes that the maximal velocity $\mathcal{V}(\cdot)$ introduced in (1.1)–(1.3) is actually a function of $\gamma = \frac{s}{L}$ defined on $\gamma = \frac{s}{L} \ge 1$, i.e.,

$$\mathcal{V}(s) = W\left(\frac{s}{L}\right), \tag{4.1}$$

then (1.1) and (1.7) take the form

$$\frac{dx_k}{dt} = u_k \quad \text{and} \quad \frac{du_k}{dt} = W'(\gamma_k)\left(\frac{u_{k+1} - u_k}{L}\right) + \frac{(W(\gamma_k) - u_k)}{\epsilon}, \tag{4.2}$$

where again

$$\gamma_k = \frac{(x_{k+1} - x_k)}{L} \tag{4.3}$$

and

$$\frac{d\gamma_k}{dt} = \frac{u_{k+1} - u_k}{L}. \tag{4.4}$$

The connection between the follow-the-leader system (4.1)–(4.4) is now clear. One introduces reference coordinates

$$X_k = kL, \tag{4.5}$$

lets

$$\mathcal{X}(X_k, t) = x_k(t) \quad \text{and} \quad u(X_k, t) = u_k(t), \tag{4.6}$$

and interprets γ_k and $\frac{u_{k+1}-u_k}{L}$ as the downwind finite difference approximations to $\frac{\partial \mathcal{X}}{\partial X}$ and $\frac{\partial u}{\partial X}$ at the reference point X_k; i.e.,

$$\frac{\partial \mathcal{X}}{\partial X}(X_k, t) = \gamma_k = \frac{x_{k+1} - x_k}{L} \quad \text{and} \quad \frac{\partial U}{\partial X}(X_k, t) = \frac{u_{k+1} - u_k}{L}. \tag{4.7}$$

With these identifications one obtains, at least formally, the Lagrangian traffic equations

$$\frac{\partial \mathcal{X}}{\partial t}(X, t) = u(X, t) \quad \text{and} \quad \frac{\partial \mathcal{X}}{\partial X} = \gamma(X, t), \tag{4.8}$$

where

$$\frac{\partial \gamma}{\partial t} = \frac{\partial u}{\partial X} \quad \text{and} \quad \frac{\partial u}{\partial t} = W'(\gamma)\frac{\partial u}{\partial X} + \frac{(W(\gamma) - u)}{\epsilon}. \tag{4.9}$$

This correspondence is faithful if one restricts one's attention to initial value problems exclusively. We have not seen how to incorporate the traffic light problem into a continuum format.

REFERENCES

[1] J.M. Greenberg, *Extensions and amplifications of a traffic model of Aw and Rascle*, SIAM J. Appl. Math., 62 (2001), pp. 729–745.

[2] A. Aw and M. Rascle, *Resurrection of "second order" models of traffic flow*, SIAM J. Appl. Math., 60 (2000), pp. 916–938.

[3] A. Aw, A. Klar, T. Materne, and M. Rascle, *Derivation of continuum traffic flow models from microscopic follow-the-leader models*, SIAM J. Appl. Math., to appear.

[4] M.J. Lighthill and G.B. Whitham, *On kinematic waves.* I: *Flood movement in long rivers*, Proc. Roy. Soc. London Ser. A, 229 (1955), pp. 281–316.

[5] M.J. Lighthill and G.B. Whitham, *On kinematic waves.* II: *A theory of traffic flow on long crowded roads*, Proc. Roy. Soc. London Ser. A, 229 (1955), pp. 317–345.

[6] P.I. Richards, *Shock waves on the highway*, Oper. Res., 4 (1956), pp. 42–51.

[7] G.B. Whitham, *Linear and Nonlinear Waves*, Pure and Applied Mathematics, Wiley-Interscience, New York, 1974.

[8] R. Rotherby, *Car following models*, in Traffic Flow Theory - A State of the Art Report, Special Report 165, Transportation Research Board, Washington, DC, 1975, Chapter 4; also available online from http://www.tfhrc.gov/its/tft/chap4.html.

[9] H.M. Zhang, *A non-equilibrium traffic model devoid of gas-like behavior*, Transportation Res. B, 36 (2002), pp. 275–290.

SIAM J. APPL. MATH.
Vol. 63, No. 1, pp. 169–194

AN ASYMPTOTIC FINITE DEFORMATION ANALYSIS FOR AN ISOTROPIC COMPRESSIBLE HYPERELASTIC HALF-SPACE SUBJECTED TO A TENSILE POINT LOAD*

DEBRA POLIGNONE WARNE† AND PAUL G. WARNE†

Abstract. The nonlinearly elastic Boussinesq problem is to find the deformation produced in a homogeneous, isotropic, elastic half-space by a point force normal to the undeformed boundary, using the exact equations of elasticity. For this core problem of elasticity and engineering, the 1885 linear elasticity solution of Boussinesq is still used in a variety of applications. In [*SIAM J. Appl. Math.*, 62 (2001), pp. 107–128], we addressed the case of a tensile point load under the constraint of incompressible finite elasticity. Here we consider an analogous asymptotic analysis of this problem within the context of compressible finite elasticity. Asymptotic tests are developed to determine whether an isotropic hyperelastic material can support a finite deflection under a tensile point load. The results are then applied to a variety of particular constitutive models for compressible nonlinearly elastic materials. It is found that, for many of the well-known strain energy models for compressible hyperelastic materials proposed in the literature, a tensile point load cannot be supported. For models which may sustain a tensile point load, we determine the remaining equations and conditions for the asymptotic solution, and numerically compute this solution for a particular case.

Key words. point load, concentrated load, Boussinesq, asymptotic analysis, compressible hyperelasticity, material formulation of equilibrium, conservation laws

AMS subject classifications. 73G05, 73C50, 73V99, 35Q72, 35B40

PII. S0036139901394955

1. Introduction.

In this paper, we continue to study the axisymmetric deformation of an isotropic, nonlinearly elastic half-space subjected to a *tensile* point force normal to its undeformed boundary. The present authors have considered this problem in the context of incompressible hyperelasticity [1] and here treat the unconstrained problem. In 1885, Boussinesq solved the analogous problem within the linear theory of elasticity, determining the following nondimensional solution to the linearized equations in the case of a unit point load (see, e.g., [2] for a discussion of the linear Boussinesq problem and solution):

$$r(R,Z) = R - \frac{(1-2\nu)}{4\pi R}\left[\frac{Z}{\sqrt{R^2+Z^2}} - 1 + \frac{1}{(1-2\nu)}\frac{R^2 Z}{\sqrt{(R^2+Z^2)^3}}\right], \tag{1}$$

$$z(R,Z) = Z - \frac{1}{4\pi}\left[\frac{Z^2}{\sqrt{(R^2+Z^2)^3}} + \frac{2(1-\nu)}{\sqrt{R^2+Z^2}}\right]. \tag{2}$$

As discussed in [3], the classical Boussinesq solution is deficient in that it not only predicts an infinite displacement under the point load, thereby violating the basic premise of linear elasticity, but also implies, in the case of compressive loads, that some particles on the line of action of the load pass through one another. To rectify these fundamental physical defects, Simmonds and Warne [3] treat the concentrated

*Received by the editors September 10, 2001; accepted for publication (in revised form) November 29, 2001; published electronically August 28, 2002. This research was supported by the James Madison University Program of Grants for Faculty Assistance.
http://www.siam.org/journals/siap/63-1/39495.html
†Department of Mathematics and Statistics, James Madison University, Harrisonburg, VA 22807 (warneda@math.jmu.edu, warnepg@math.jmu.edu).

tensile load problem posed within the nonlinear theory of elastostatics and also further study additional interesting aspects of this core problem. In [3], the authors invoke two hypotheses (H1 and H2), which describe the role played by (1), (2) in the context of the fully nonlinear, exact formulation and provide for unique asymptotic solutions near the point load. We shall again adopt these hypotheses as we did in [1], and list them below:

H1: As the dimensionless distance from the point load grows, the solution(s) of the nonlinear Boussinesq problem approaches the solution of the linear Boussinesq problem.

H2: The strain-energy density is bounded everywhere except at the point load; the displacements are bounded everywhere.

In [3], Simmonds and Warne use the principle of stationary potential energy to derive the associated Euler equations from a variational formulation, and then employ known conservation laws of elastostatics [4] and a third hypothesis (H3), successfully used by Knowles and Sternberg [5, 6] for asymptotic analysis of the finite-deformation elastostatic field near a crack tip. In [3] it is shown that the special Blatz–Ko material cannot support a point load and that the generalized neo-Hookean material considered by Knowles [7] with dimensionless stiffening parameter k can support a finite deflection under a point load provided that the material is sufficiently stiffer ($k > \frac{3}{2}$) than the conventional neo-Hookean material ($k = 1$).

In [1], we address the nonlinearly elastic Boussinesq problem using a material formulation of the governing equations in terms of nominal stresses. Upon invoking the hypotheses which proved useful in [3], simple criteria to determine if a hyperelastic material can support a point load are developed in [1] for incompressible isotropic materials, and the results are then applied to a variety of models in the literature. There have been very few works considering the nonlinear Boussinesq problem. The related problem of a tensile concentrated load acting on a cone composed of a particular compressible hyperelastic material proposed by Gao has been treated asymptotically by Gao and Liu [8]. In [3], a case-by-case asymptotic analysis of the governing equations was carried out, and there it was noted that much remains to be done for this problem, including determination of criteria of the type established in [1]. Thus, in this paper, we consider the corresponding problem to that considered in [1] for compressible finite elasticity and establish, in this context, asymptotic tests which allow for a much simpler way to determine whether a compressible hyperelastic material can support a finite deflection under a point load. In subsequent work, we shall similarly address the even more complex problem of an inward (compressive) point load.

In section 2, we briefly review our material formulation of this traction boundary-value problem (developed in [1]) which, in the authors' view, sets up this problem in a rather tractable and explicable form, and then give the governing equations of equilibrium in terms of nominal stress components. Section 3 considers hyperelasticity and restates the basic boundary-value problem for compressible nonlinearly elastic materials upon using a representation for the strain energy which proves more convenient than the standard principal isotropic invariant expression. In section 4, we exploit our material formulation of this problem together with several conservation laws for nonlinear elastostatics, and upon linearizing and invoking H1 as in [1], implications of the linear solution (1), (2) for the conservation integrals are obtained. In section 5, the development and use of simple asymptotic tests for compressible materials is presented, and many of the strain-energy density functions proposed in the literature for compressible hyperelastic material models are tested for the ability to support a finite deflection under tensile point load.

2. Problem formulation. We consider, from the outset, nondimensionalized quantities (see, e.g., [3]), where the position vector $\mathbf{x}$ in the deformed configuration of the body has cylindrical polar coordinates (r, θ, z) and the position vector $\mathbf{X}$ in the reference configuration, the half-space defined by $Z \geq 0$, can be described in terms of cylindrical polar coordinates (R, Θ, Z) or spherical polars (ξ, Φ, Θ). Our interest is thus in axisymmetric deformation fields given by

$$r = r(R, Z) \quad \text{or} \quad r = r(\xi, \Phi), \tag{3}$$

$$\theta = \Theta, \tag{4}$$

$$z = z(R, Z) \quad \text{or} \quad z = z(\xi, \Phi), \tag{5}$$

where the orthonormal bases $(\mathbf{E}_R, \mathbf{E}_\Theta, \mathbf{E}_Z)$ or $(\mathbf{E}_\xi, \mathbf{E}_\Phi, \mathbf{E}_\Theta)$ in the reference configuration are such that

$$\mathbf{E}_R = \cos\Theta\mathbf{E}_X + \sin\Theta\mathbf{E}_Y, \qquad \mathbf{E}_\Theta = -\sin\Theta\mathbf{E}_X + \cos\Theta\mathbf{E}_Y$$

and

$$\mathbf{E}_\xi = \sin\Phi\mathbf{E}_R + \cos\Phi\mathbf{E}_Z, \qquad \mathbf{E}_\Phi = \cos\Phi\mathbf{E}_R - \sin\Phi\mathbf{E}_Z,$$

respectively, and $R = \xi\sin\Phi$, $Z = \xi\cos\Phi$. In the deformed configuration, we thus have orthonormal bases $(\mathbf{e}_r, \mathbf{e}_\theta, \mathbf{e}_z)$ with

$$\mathbf{e}_r = \cos\theta\mathbf{e}_x + \sin\theta\mathbf{e}_y, \qquad \mathbf{e}_\theta = -\sin\theta\mathbf{e}_x + \cos\theta\mathbf{e}_y,$$

and $\mathbf{x} = r\mathbf{e}_r + z\mathbf{e}_z$. In the above, $(\mathbf{E}_X, \mathbf{E}_Y, \mathbf{E}_Z)$ and $(\mathbf{e}_x, \mathbf{e}_y, \mathbf{e}_z)$ are orthonormal Cartesian bases for the reference and deformed configurations, respectively.

The corresponding deformation gradient tensor $\mathbf{F}$ referred to the appropriate right-handed coordinate system is then given by either

$$\mathbf{F} = \begin{bmatrix} r_{,R} & 0 & r_{,Z} \\ 0 & \dfrac{r}{R} & 0 \\ z_{,R} & 0 & z_{,Z} \end{bmatrix} \quad \text{or} \quad \mathbf{F} = \begin{bmatrix} r_{,\xi} & z_{,\xi} & 0 \\ 0 & 0 & \dfrac{r}{\xi\sin\Phi} \\ \dfrac{1}{\xi}r_{,\Phi} & \dfrac{1}{\xi}z_{,\Phi} & 0 \end{bmatrix}, \tag{6}$$

where here and elsewhere a comma denotes differentiation.

The point load is applied via a dimensionless tensile concentrated load normal to the surface $Z = 0$, and thus the dimensionless applied surface traction vector in a cylindrical basis is given by

$$\mathbf{s}(R) = \frac{-\delta(R)}{2\pi R}\mathbf{e}_z \quad \text{on } Z = 0. \tag{7}$$

This implies a boundary condition for the *nominal stress tensor* $\mathbf{S}$ $(= (\det\mathbf{F})\mathbf{F}^{-1}\mathbf{T}$; see, e.g., Ogden [9]) of the form

$$\begin{aligned} \mathbf{s} = \mathbf{S}^T\mathbf{N} &= \mathbf{S}^T(-\mathbf{E}_Z) \\ &= -S_{Zr}\mathbf{e}_r - S_{Zz}\mathbf{e}_z, \end{aligned}$$

as the component forms of $\mathbf{S}$ are similar to those of $\mathbf{F}$ in that they contain the same null components. Thus, the boundary conditions at $Z = 0$ in terms of components of the nominal stress tensor referred to a cylindrical polar basis are

$$S_{Zr} = 0 \quad \text{and} \quad S_{Zz} = \frac{\delta(R)}{2\pi R} \quad \text{on } Z = 0. \tag{8}$$

As determined in [1], the equations of equilibrium $\operatorname{Div}\mathbf{S} \equiv \nabla \cdot \mathbf{S} = \mathbf{0}$ governing deformation fields of the type $(3)_2$, (4), $(5)_2$, where $\mathbf{S}$, represented in a mixed basis, takes the form

$$\mathbf{S} = S_{\xi r}\mathbf{E}_\xi \otimes \mathbf{e}_r + S_{\xi z}\mathbf{E}_\xi \otimes \mathbf{e}_z + S_{\Phi r}\mathbf{E}_\Phi \otimes \mathbf{e}_r + S_{\Phi z}\mathbf{E}_\Phi \otimes \mathbf{e}_z + S_{\Theta\theta}\mathbf{E}_\Theta \otimes \mathbf{e}_\theta$$

and ∇ is the usual spherical gradient operator, reduce to the two equations

$$\sin\Phi(\xi^2 S_{\xi r})_{,\xi} + \xi(\sin\Phi S_{\Phi r})_{,\Phi} - \xi S_{\Theta\theta} = 0 \tag{9}$$

and

$$\sin\Phi(\xi^2 S_{\xi z})_{,\xi} + \xi(\sin\Phi S_{\Phi z})_{,\Phi} = 0. \tag{10}$$

Equation (10) can be written in divergence form and implies the existence of a function $v(\xi, \Phi)$ such that

$$v_{,\Phi} = \xi^2 \sin\Phi S_{\xi z} \quad \text{and} \quad -v_{,\xi} = \xi \sin\Phi S_{\Phi z}, \tag{11}$$

while regularity requirements at the origin $\xi = 0$ and along the axis of symmetry $\Phi = 0$ give

$$v(0^+, \Phi) = 0 \quad \text{and} \quad v(\xi, 0^+) = 0. \tag{12}$$

Also, converting the boundary condition $(8)_1$, we obtain

$$\cos\Phi S_{\xi r} - \sin\Phi S_{\Phi r} = 0 \quad \text{at } \Phi = \frac{\pi}{2}, \tag{13}$$

$$\Rightarrow \quad S_{\Phi r} = 0 \quad \text{at } \Phi = \frac{\pi}{2}, \quad \xi > 0. \tag{14}$$

The implications of the boundary condition $(8)_2$ will be considered in section 4.

3. Hyperelasticity, equilibrium equations, and boundary conditions. We now consider the implications of hyperelasticity and thus assume the existence of a (dimensionless) stored-energy density function $W(\mathbf{F})$ such that

$$\mathbf{S} = \frac{\partial W}{\partial \mathbf{F}} \quad \text{and} \quad S_{ij} = \frac{\partial W}{\partial F_{ji}}. \tag{15}$$

Then the equilibrium equations (9) and (11) become

$$\sin\Phi(\xi^2 W_{,(r,\xi)})_{,\xi} + \xi^2(\sin\Phi W_{,(r,\Phi)})_{,\Phi} - \xi^2 \sin\Phi W_{,r} = 0 \tag{16}$$

and

$$v_{,\Phi} = \xi^2 \sin\Phi W_{,(z,\xi)} \quad \text{and} \quad -v_{,\xi} = \xi^2 \sin\Phi W_{,(z,\Phi)}, \tag{17}$$

recovering the Euler equations derived in [3] (see [3, (18) and (24)]).

For isotropic hyperelastic materials, the stored-energy function is such that

$$W = W(I_1, I_2, I_3),$$

where I_1, I_2, I_3 are the standard principal isotropic invariants of the left and right Cauchy–Green deformation tensors $\mathbf{F}\mathbf{F}^T$ and $\mathbf{F}^T\mathbf{F}$, respectively. We note that the usual normalization conditions,

$$W(3,3,1) = 0 \quad \text{and} \quad \left(\frac{\partial W}{\partial I_1} + \frac{2\partial W}{\partial I_2} + \frac{\partial W}{\partial I_3}\right)\bigg|_{I_1=I_2=3, I_3=1} = 0,$$

require that both the strain energy and the stress, respectively, vanish in the reference configuration. Upon computing the invariants I_1, I_2, I_3 with $\mathbf{F}$ given by $(6)_2$, it is convenient, as in [1] and [3], to use the representation

$$I_1 = A + C, \tag{18}$$

$$I_2 = AC + B^2, \tag{19}$$

$$I_3 = B^2 C, \tag{20}$$

where

$$A = r_{,\xi}^2 + z_{,\xi}^2 + \xi^{-2}(r_{,\Phi}^2 + z_{,\Phi}^2), \tag{21}$$

$$B = \xi^{-1}(r_{,\Phi}\, z_{,\xi} - r_{,\xi}\, z_{,\Phi}), \tag{22}$$

$$C = (\xi \sin\Phi)^{-2} r^2. \tag{23}$$

We consider a general nondimensionalized representation of the strain-energy density function for compressible isotropic hyperelastic materials given by

$$W = W(A, B, C), \tag{24}$$

where here and elsewhere we shall neglect to introduce a new symbol for W on the right-hand side for ease of notation. Then the equilibrium equations (16) and (17) and the boundary condition (14) become

$$\sin\Phi(2\xi^2 r_{,\xi}\, W_A - \xi z_{,\Phi}\, W_B)_{,\xi} + [(2r_{,\Phi}\, W_A + \xi z_{,\xi}\, W_B)\sin\Phi]_{,\Phi} - 2r\csc\Phi W_C = 0, \tag{25}$$

$$v_{,\Phi} = \sin\Phi(2\xi^2 z_{,\xi}\, W_A + \xi r_{,\Phi}\, W_B), \qquad -v_{,\xi} = \sin\Phi(2z_{,\Phi}\, W_A - \xi r_{,\xi}\, W_B), \tag{26}$$

and

$$2r_{,\Phi}\, W_A + \xi z_{,\xi}\, W_B = 0 \quad \text{at } \Phi = \tfrac{\pi}{2}, \quad \xi > 0, \tag{27}$$

respectively, where here and subsequently $W_A \equiv \frac{\partial W}{\partial A}$, and similarly for W_B, W_C.

4. Conservation laws and implications of the linearly elastic Boussinesq problem. We next consider three integral identities which have proven useful in the asymptotic analyses carried out in [1] and [3] for this problem. The first, a direct application of the divergence theorem

$$\int_{\partial\Omega} \mathbf{S}^T \mathbf{N}\, dA = \int_{\Omega} \operatorname{Div} \mathbf{S} dV$$

and equilibrium ($\operatorname{Div}\mathbf{S} = \mathbf{0}$), states that

$$\int_{\partial\Omega} \mathbf{S}^T \mathbf{N} dA = \mathbf{0}, \tag{28}$$

where here and subsequently we take Ω to be the hemisphere of radius $\xi = a$ which is centered at the source, and thus the unit outward normal to the top surface of the hemisphere is $\mathbf{N} = -\mathbf{E}_Z$, while the unit outward normal to the lateral surface is $\mathbf{N} = \mathbf{E}_\xi = \sin\Phi\mathbf{E}_R + \cos\Phi\mathbf{E}_Z$. Thus, as in [1], the above along with $(8)_2$ yields

$$1 = 2\pi a^2 \int_0^{\frac{\pi}{2}} S_{\xi z} \sin\Phi d\Phi, \tag{29}$$

which expresses, in dimensionless form, overall force equilibrium. As before, $(11)_1$, $(12)_2$, and (29) further imply that

$$v\left(\xi, \frac{\pi}{2}\right) = \frac{1}{2\pi} \quad \text{at } \xi = a. \tag{30}$$

The second conservation law, related to Eshelby's energy-momentum tensor, is derived from an application of the divergence theorem to the tensor $W\mathbf{I} - \mathbf{SF}$. As the divergence of this tensor also vanishes (see Chadwick [4]), we obtain

$$\int_{\partial\Omega} (W\mathbf{I} - \mathbf{SF})\mathbf{N}\, dA = \mathbf{0}, \tag{31}$$

which, as shown in [1], results in

$$\begin{aligned} &\int_0^a WR dR - \frac{1}{2\pi} z_{,Z}\,(R = 0, Z = 0) \\ &\quad = \int_0^{\frac{\pi}{2}} \left[\cos\Phi W + S_{\xi r}\left(\frac{\sin\Phi}{a} r_{,\Phi} - \cos\Phi r_{,\xi}\right)\right. \\ &\qquad \left. + S_{\xi z}\left(\frac{\sin\Phi}{a} z_{,\Phi} - \cos\Phi z_{,\xi}\right)\right] a^2 \sin\Phi d\Phi. \end{aligned} \tag{32}$$

The third relation we shall use is the integral identity

$$3\int_\Omega W dV = \int_{\partial\Omega} [W\mathbf{X}\cdot\mathbf{N} + (\mathbf{S}^T\mathbf{N})\cdot(\mathbf{x} - \mathbf{FX})] dA, \tag{33}$$

which is contained among conservation laws of elastostatics derived by Chadwick [4] from the energy-momentum tensor used above. Specializing to the problem under consideration here, we recall from [1] that this determines the relation

$$\begin{aligned} &3\int_0^a \int_0^{\frac{\pi}{2}} W\xi^2 \sin\Phi d\xi d\Phi + \frac{1}{2\pi} z(R = 0, Z = 0) \\ &\quad = a^2 \int_0^{\frac{\pi}{2}} [aW + S_{\xi r}(r - a r_{,\xi}) + S_{\xi z}(z - a z_{,\xi})] \sin\Phi d\Phi. \end{aligned} \tag{34}$$

To conclude this section, we restate the implications of the linearly elastic Boussinesq problem solution (1), (2) for this problem. The solution (1), (2) can be represented as

$$\begin{aligned} r(\xi, \Phi) &= \xi \sin\Phi + \frac{1}{4\pi\xi}[(1 - 2\nu)\csc\Phi(1 - \cos\Phi) - \sin\Phi\cos\Phi], \\ z(\xi, \Phi) &= \xi\cos\Phi + \frac{1}{4\pi\xi}[2(\nu - 1) - \cos^2\Phi]. \end{aligned}$$

Invoking H1 and letting $a \longrightarrow \infty$, we have the following asymptotic results for the terms contained in the above conservation laws:

$$W \longrightarrow O(a^{-4}), \tag{35}$$

$$\mathbf{S} \longrightarrow O(a^{-2}), \tag{36}$$

$$(r - \xi r_{,\xi})|_{\xi = a} \longrightarrow O(a^{-1}), \tag{37}$$

$$(z - \xi z_{,\xi})|_{\xi=a} \longrightarrow O(a^{-1}), \tag{38}$$

$$\left(\frac{\sin\Phi}{\xi} r_{,\Phi} - \cos\Phi r_{,\xi}\right)|_{\xi=a} \longrightarrow O(a^{-2}), \tag{39}$$

$$\left(\frac{\sin\Phi}{\xi} z_{,\Phi} - \cos\Phi z_{,\xi}\right)|_{\xi=a} \longrightarrow -1 + O(a^{-2}). \tag{40}$$

On converting the conservation law results (32) and (34) to spherical representations in the reference configuration, letting $a \longrightarrow \infty$, and employing the above, we obtain

$$0 < 2\pi \int_0^\infty W|_{\Phi=\frac{\pi}{2}} \xi d\xi + 1 = z_{,\xi}\,(0^+, 0) \tag{41}$$

and

$$6\pi \int_0^\infty \int_0^{\frac{\pi}{2}} W\xi^2 \sin\Phi d\xi d\Phi = -z(0^+, 0), \tag{42}$$

respectively. The results (41) and (42), which relate the deflection of the half-space and its partial derivative at the point of application of the load to the stored-energy function, will be useful (as in [1, 3]) in the asymptotic analysis to follow.

5. Asymptotics. We now adopt a third hypothesis (H3) successfully used previously for asymptotic analyses of problems involving singularities in finite elasticity (see [1, 3, 5, 6]):

H3: The unknowns r, z, and v have the following asymptotic forms as $\xi \longrightarrow 0$:

$$r(\xi, \Phi) = \xi^\alpha F(\Phi) + o(\xi^\alpha) \quad \text{with } F(0) = 0, \quad F(\Phi) > 0 \quad \text{for } 0 < \Phi \leq \tfrac{\pi}{2}, \tag{43}$$

$$z(\xi, \Phi) = z(0^+, 0) + \xi^\beta G(\Phi) + o(\xi^\beta), \quad G(0) \neq 0, \tag{44}$$

$$v(\xi, \Phi) = \xi^\delta I(\Phi) + o(\xi^\delta), \quad I(\Phi) \not\equiv 0. \tag{45}$$

Partial derivatives also have analogous asymptotic forms.

As discussed in [3], the above restrictions on F ensure that particles on the z-axis remain there and that particles do not pass through the z-axis, while the restriction on $G(0)$ follows from (41). Additionally, the exponents in (43)–(45) must be such that

$$\alpha > 0, \tag{46}$$

$$\beta > 0, \tag{47}$$

$$\beta \leq 1, \tag{48}$$

$$\delta = 0. \tag{49}$$

These restrictions result from requiring no cavity to form beneath the point load, a finite deflection under the point load, (41), and (30), respectively.

Following the development of [1, 3], we define for brevity

$$\sigma(x) = \begin{cases} 1, & x \geq 0, \\ 0, & x < 0, \end{cases} \tag{50}$$

$$\omega = \min\{\alpha, \beta\}, \tag{51}$$

and thus from (21)–(23) and the above, we have that, as $\xi \longrightarrow 0+$,

$$A \longrightarrow \xi^{2(\omega-1)} \bar{A}(F, F', G, G'), \tag{52}$$

$$B \longrightarrow \xi^{\alpha+\beta-2}\bar{B}(F, F', G, G'), \tag{53}$$

$$C \longrightarrow \xi^{2(\alpha-1)}F^2 \csc^2\Phi, \tag{54}$$

where $\bar{A}$ and $\bar{B}$ are defined for convenience as

$$\bar{A} \equiv \sigma(\beta-\alpha)[\alpha^2F^2 + F'^2] + \sigma(\alpha-\beta)[\beta^2G^2 + G'^2], \tag{55}$$

$$\bar{B} \equiv \beta F'G - \alpha FG' \not\equiv 0 \tag{56}$$

and the final restriction on $\bar{B}$ is necessary to preclude the unacceptable conclusion that $F \equiv 0$. Similarly to [1], we consider the asymptotic forms for the partial derivatives of a strain energy function $W(A, B, C)$, given generally as

$$W_A \longrightarrow \xi^{n(\alpha,\beta)}N(F, F', G, G', \sin\Phi), \tag{57}$$

$$W_B \longrightarrow \xi^{m(\alpha,\beta)}M(F, F', G, G', \sin\Phi), \tag{58}$$

$$W_C \longrightarrow \xi^{p(\alpha,\beta)}P(F, F', G, G', \sin\Phi), \tag{59}$$

where m, n, and p are functions of α and β and M, N, and P are functions of the arguments shown.

5.1. Asymptotic tests: Derivation. We consider now general compressible hyperelastic materials and derive in what follows some simple asymptotic tests on the form of its stored energy to determine whether such a material can support a tensile point load. In section 5.2, we apply our tests to a variety of well-known material models. We remind the reader that the equilibrium equations and boundary condition for this problem are given by (25)–(27), respectively. We now consider the asymptotic implications of these equations.

On substituting (43)–(45) and (57)–(59) into (25), differentiating, and collecting like powers of ξ, we obtain the following asymptotic form of the equilibrium equation (25):

$$\begin{aligned} &2\xi^{n+\alpha}[\alpha(n+\alpha+1)\sin\Phi FN + (\sin\Phi F'N)'] - 2\xi^{p+\alpha}\csc\Phi FP \\ &\quad + \xi^{m+\beta}[(\beta\sin\Phi GM)' - (m+\beta+1)\sin\Phi G'M] = 0. \end{aligned} \tag{60}$$

Next, substituting from (43)–(45), the asymptotic forms of (26) are

$$I' = (2\beta\xi^{n+\beta+1}GN + \xi^{m+\alpha+1}F'M)\sin\Phi \tag{61}$$

and

$$0 = 2\xi^{n+\beta}G'N - \alpha\xi^{m+\alpha}FM, \tag{62}$$

respectively. Finally, the boundary condition (27) becomes

$$2\xi^{n+\alpha}F'N + \beta\xi^{m+\beta}GM = 0 \quad \text{at } \Phi = \tfrac{\pi}{2}, \quad \xi > 0. \tag{63}$$

In addition, the Jacobian

$$J = \det\mathbf{F} = I_3^{\frac{1}{2}}$$

for the deformation of concern here is given by the above, (20), (22), and (23) as

$$J = \frac{r}{\xi^2\sin\Phi}(r_{,\Phi}\, z_{,\xi} - r_{,\xi}\, z_{,\Phi}\,). \tag{64}$$

On substituting from (43) and (44), we have that asymptotically

$$J \longrightarrow \xi^{2\alpha+\beta-3} F(\beta F' G - \alpha F G') \csc\Phi \quad \text{as } \xi \longrightarrow 0. \tag{65}$$

Thus, to maintain the fundamental physical and mathematical restrictions that $\det \mathbf{F}$ is positive and bounded, we must have

$$2\alpha + \beta - 3 = 0 \Rightarrow \alpha = \frac{3-\beta}{2} \tag{66}$$

and, noting $(43)_{2,3}$ and (56),

$$\bar{B} \equiv \beta F' G - \alpha F G' > 0, \qquad F\bar{B} \csc\Phi \text{ bounded}. \tag{67}$$

Recall from (47), (48) that $0 < \beta \leq 1$, and thus (66) requires $1 \leq \alpha < \frac{3}{2}$, and so $\beta \leq \alpha$ and $\beta = 1 \Leftrightarrow \alpha = 1$. Thus from (51)–(54), A, B, and C are independent of ξ when $\alpha = \beta = 1$, and hence the invariants (18)–(20) would not depend asymptotically on ξ, which is clearly unphysical. Thus we restrict attention to values of β such that

$$0 < \beta < 1, \tag{68}$$

which together with (66) gives $1 < \alpha < \frac{3}{2}$, and so by (51), $\omega = \beta$. Thus the asymptotic forms (52)–(54) of A, B, and C reduce to

$$A \longrightarrow \xi^{2(\beta-1)} \bar{A}(F, F', G, G'), \tag{69}$$

$$B \longrightarrow \xi^{\frac{\beta-1}{2}} \bar{B}(F, F', G, G'), \tag{70}$$

$$C \longrightarrow \xi^{(1-\beta)} \bar{C}, \tag{71}$$

where, using $\omega = \beta$, (50), (55), (56), and (67),

$$\bar{A} \equiv \beta^2 G^2 + G'^2, \tag{72}$$

$$\bar{B} \equiv \beta F' G - \alpha F G' > 0, \tag{73}$$

$$\bar{C} \equiv F^2 \csc^2\Phi, \tag{74}$$

and α is given in terms of β by (66). The asymptotic forms for the partial derivatives of a strain energy function $W(A, B, C)$ can then be given generally as

$$W_A \longrightarrow \xi^{n(\beta)} N(F, F', G, G', \sin\Phi), \tag{75}$$

$$W_B \longrightarrow \xi^{m(\beta)} M(F, F', G, G', \sin\Phi), \tag{76}$$

$$W_C \longrightarrow \xi^{p(\beta)} P(F, F', G, G', \sin\Phi). \tag{77}$$

Consider (62) first. As $\xi \longrightarrow 0+$, if $m + \alpha < n + \beta$, then the coefficient of $\xi^{m+\alpha}$ dominates, implying

$$0 = F(\Phi) M.$$

However, recalling $(43)_3$ and (75), $F(\Phi)$ is not identically zero, nor should W_B be identically zero for a reasonable strain energy. Thus

$$m + \alpha \geq n + \beta. \tag{78}$$

In view of (78) then, the $\xi^{n+\beta+1}$ term(s) dominate the right-hand side of (61). Thus we must determine now whether the exponent $n+\beta+1$ can be >0, <0, or $=0$. If $n+\beta+1>0$, then the term on the left in (61) dominates as $\xi \longrightarrow 0+$, resulting asymptotically in $I'=0$ so that

$$I = I_0, \tag{79}$$

where I_0 is an unknown constant. However, $(12)_2$ and (30) asymptotically imply via (45) and (49) that

$$I(0^+) = 0 \quad \text{and} \quad I\left(\frac{\pi}{2}\right) = \frac{1}{2\pi}, \tag{80}$$

and thus (79) is not possible. If $n+\beta+1<0$, then terms on the right of (61) dominate as $\xi \longrightarrow 0+$, and hence I has no influence asymptotically on the solution of the problem since (61) is the only governing equation containing I. However, by $(8)_2$, the point load is applied through S_{Zz}, and

$$S_{Zz} = \cos\Phi S_{\xi z} - \sin\Phi S_{\Phi z} = \xi^{-2}\cot\Phi v_{,\Phi} + \xi^{-1} v_{,\xi} \tag{81}$$

upon using (11). Therefore, by (45) and (49), the point load manifests itself through I, making the above alternative for the case $n+\beta+1<0$ unacceptable. Thus, in a manner analogous to [1], we arrive at the important result that

$$n+\beta+1=0. \tag{82}$$

Recalling that $n=n(\beta)$ was introduced in (74), the function $n(\beta)$ is known for a given strain energy function W, and thus (82) is an equation to determine β. This brings us to our first result, which is analogous to the situation for incompressible hyperelastic solids treated in [1]: *For a given strain energy function W, if β calculated through* (82) (*where n is determined from* (74)) *does not fall in the range* $0<\beta<1$ *given by* (68), *then the material modeled by W cannot sustain a point load.*

We now continue our analysis under the assumption that β satisfying (82) is consistent with (68); i.e., $n(\beta)+\beta+1=0$ and $0<\beta<1$. Recall that by (78) determined above, $m+\alpha \geq n+\beta$. We consider first the possibility that

$$m+\alpha = n+\beta \tag{83}$$

so that the two terms on the right-hand sides of (61) and (62) balance. We will show below that (83) can occur *only* for special material models. The final case of when strict inequality holds in (78) will be treated last in the section.

We proceed under the requirement of (82) and suppose that (83) holds as well. Then (61) and (62) reduce asymptotically to

$$I' = (2\beta GN + F'M)\sin\Phi \tag{84}$$

and

$$0 = 2G'N - \alpha FM, \tag{85}$$

respectively, where by (66), $\alpha = \frac{3-\beta}{2}$. Similarly, the powers of ξ appearing in (60) are given by

$$n+\alpha = \frac{1}{2} - \frac{3}{2}\beta, \quad m+\beta = \frac{3}{2}\beta - \frac{5}{2}, \quad \text{and} \quad p+\alpha = p + \frac{3}{2} - \frac{1}{2}\beta, \tag{86}$$

and so (60) takes the form

$$
\begin{aligned}
(87)\qquad & 2\xi^{\frac{1}{2}-\frac{3}{2}\beta}\left[\frac{3}{4}(3-\beta)(1-\beta)\sin\Phi FN+(\sin\Phi F'N)'\right]-2\xi^{p+\frac{3}{2}-\frac{1}{2}\beta}\csc\Phi FP\\
&+\xi^{\frac{3}{2}\beta-\frac{5}{2}}\left[(\beta\sin\Phi GM)'-\frac{3}{2}(\beta-1)\sin\Phi G'M\right]=0.
\end{aligned}
$$

In view of (68), the $\xi^{\frac{3}{2}\beta-\frac{5}{2}}$ term dominates the $\xi^{\frac{1}{2}-\frac{3}{2}\beta}$ term in (87). Now if the $\xi^{p+\frac{3}{2}-\frac{1}{2}\beta}$ term dominates the $\xi^{\frac{3}{2}\beta-\frac{5}{2}}$ term, then

$$\csc\Phi FP=0. \tag{88}$$

However, as before, $F(\Phi)$ is not identically zero, nor should W_C be identically zero for a reasonable strain energy. On the other hand, if the $\xi^{\frac{3}{2}\beta-\frac{5}{2}}$ term dominates the $\xi^{p+\frac{3}{2}-\frac{1}{2}\beta}$ term in (87), the final term in brackets must be zero, which results in the equation

$$\alpha(\sin\Phi M)G'+\beta G(\sin\Phi M)'=0. \tag{89}$$

Solving (89) yields

$$G=\kappa(\sin\Phi M)^{-\frac{\beta}{\alpha}}=\kappa(\sin\Phi M)^{-\frac{2\beta}{3-\beta}}, \tag{90}$$

where κ is a constant of integration. However, the boundary condition (63) will contradict this possibility. Using (68) and $(86)_{1,2}$, we have $m+\beta<n+\alpha$, and thus the second term on the left in (63) dominates, implying that either

$$G\left(\frac{\pi}{2}\right)=0\quad\text{or}\quad M|_{\Phi=\frac{\pi}{2}}=0. \tag{91}$$

The latter of (91) would imply from (90) that $G(\frac{\pi}{2})\to\pm\infty$, which violates our second hypothesis H2 and thus is unacceptable. The former of (91) is equally unrealistic, as this would imply from (44) that $z(\xi,\frac{\pi}{2})=z(0^+,0)$, and thus in some neighborhood of the origin, the deflection of the top surface of the half-space would be exactly the same as the deflection at the point of application of the point load. Thus, for (83) to hold, the $\xi^{p+\frac{3}{2}-\frac{1}{2}\beta}$ and $\xi^{\frac{3}{2}\beta-\frac{5}{2}}$ terms in (87) must balance, requiring

$$p=2\beta-4. \tag{92}$$

In this case, (87) becomes

$$-2\csc\Phi FP+\alpha(\sin\Phi M)G'+\beta G(\sin\Phi M)'=0. \tag{93}$$

In view of (92), however, $p(\beta)$, introduced in (76), is a *specific* function of β for a given strain energy, and (82), which determines β, has already been derived. Thus if p does not satisfy (92), then (83) is not possible. From (92) and (82), (83) is thus *possible only* in the special case where p and n are related by

$$p=-2(n+3). \tag{94}$$

Thus, in the special case of the material model where (94) occurs and thus p does satisfy (92), the above asymptotic analysis gives a system of coupled first-order ordinary differential equations for F, G, and I introduced in (43)–(45) with hypothesis H3.

This system consists of (93), (84), and (85), which followed from asymptotic analysis of the governing equations (25) and (26), where (66)–(68), (82), and (83) also hold. This system is subject to the boundary conditions in (43), (44), (80), and $(91)_2$. Accordingly, here and subsequently, we have that near the point load, the deformed radius $r(\xi,\Phi)$, deflection $z(\xi,\Phi)$, and $v(\xi,\Phi)$ are such that

$$r \longrightarrow O(\xi^{\frac{3}{2}-\frac{1}{2}\beta}), \qquad z \longrightarrow z(0^+,0)+O(\xi^{\beta}), \qquad v \longrightarrow O(\xi^0) \quad \text{as } \xi \longrightarrow 0^+. \tag{95}$$

This concludes our study of the case (83) since if (92) does not hold, then this, together with our previous analysis, precludes the possibility of (83). Thus it remains to consider the case of (78) when strict inequality holds, assuming that (92) is violated.

Given the above discussion, we now assume (92) does not hold, and hence, neither does (83). Thus, from (78),

$$m+\alpha > n+\beta. \tag{96}$$

Before proceeding, we remind the reader that the exact governing equations (25) and (26) and boundary condition (27) have resulted in (60)–(63). Consider first (62), for which the inequality (96) implies that the first term on the right dominates, resulting in

$$2G'N = 0. \tag{97}$$

We note that $N \not\equiv 0$, as this would imply from (61) (see (99) below) that I is constant, which violates (80). Thus (97) results in

$$G(\Phi) = G_0 \neq 0, \tag{98}$$

where G_0 is an unknown constant with the latter restriction following from (44).

We next consider (61). The inequality (96) further indicates that the first term on the right in (61) also dominates. Since β must be chosen so that (82) holds, (61) gives

$$I' = 2\beta GN\sin\Phi, \tag{99}$$

which along with (98) implies that

$$I = 2\beta G_0\int \bar{N}\sin\Phi d\Phi + I_0, \tag{100}$$

where I_0 is an unknown constant and the expression $\bar{N}$ indicates that the arguments of N are to be evaluated at G, given by (98), and F, which is yet to be determined. Assuming that the integrand in (100) can be integrated, we express I in the form

$$I = 2\beta G_0\Psi(\Phi) + I_0. \tag{101}$$

Imposing the boundary condition $(80)_1$, we have

$$I_0 = -2\beta G_0\Psi(0^+), \tag{102}$$

while $(80)_2$ determines

$$G_0 = \frac{1}{4\pi\beta\left(\Psi\left(\frac{\pi}{2}\right)-\Psi(0^+)\right)}. \tag{103}$$

We remark that by virtue of (44), $G_0 > 0$. Thus I is given in terms of $\Psi(\Phi) = \int \bar{N}(\Phi) \sin\Phi d\Phi$ as

$$I = \frac{(\Psi(\Phi) - \Psi(0^+))}{2\pi\left(\Psi\left(\frac{\pi}{2}\right) - \Psi(0^+)\right)}. \tag{104}$$

Thus G and I of hypothesis H3 are completely determined in terms of Ψ.

What now remains is to determine $F(\Phi)$ and to consider (60) and the boundary condition (63). We first note that the second term of (60) cannot dominate the other two, as this would imply that $P = 0$ and thus W would not depend on C, which is unacceptable. Similarly, the third term of (60) cannot dominate the other two, as this would imply that $M = \frac{K}{\sin\Phi}$, $K \neq 0$ constant, while (63) would in turn require $M|_{\Phi=\frac{\pi}{2}} = 0$, which is contradictory. Thus five remaining possibilities exist for (60), corresponding to all three terms balancing, pairs of terms balancing, and the first term dominating. However, our representation of the strain-energy density function in terms of A, B, C or I_1, I_2, I_3 is also advantageous to rule out two of these five possibilities. In view of (18)–(20), we have

$$\begin{aligned} W_A &= W_1 + CW_2, \\ W_B &= 2B(W_2 + CW_3), \\ W_C &= W_1 + AW_2 + B^2W_3, \end{aligned}$$

where the usual notation $W_i = \frac{\partial W}{\partial I_i}$ is employed. Since $W_1 \longrightarrow \xi^\gamma$, $W_2 \longrightarrow \xi^\eta$, $W_3 \longrightarrow \xi^\delta$ for some γ, η, δ, the above along with (66), (68), (69)–(71), and (75)–(77) can be used to determine that the $\xi^{n+\alpha}$ term in (60) cannot dominate alone, nor can the $\xi^{n+\alpha}$ and $\xi^{m+\beta}$ terms balance and dominate the remaining term.[1] Thus the only remaining cases are the following: the $\xi^{n+\alpha}$ and $\xi^{p+\alpha}$ terms balance and dominate the remaining term (case I), all three terms of (60) balance (case II), and the $\xi^{p+\alpha}$ and $\xi^{m+\beta}$ terms balance and dominate the remaining term (case III).

Consider first case I. Then the asymptotic form of (60) yields the following ordinary differential equation for F:

$$\frac{3}{4}(1-\beta)(3-\beta)\sin\Phi F\bar{N} + (\sin\Phi F'\bar{N})' - \csc\Phi F\bar{P} = 0. \tag{105}$$

In addition, the boundary condition (63) for this case becomes

$$\bar{N}|_{\Phi=\frac{\pi}{2}} = 0. \tag{106}$$

Thus, for a given material model, (105) and (106) provide a first-order ordinary differential equation and boundary condition to determine the final unknown F in the asymptotic solution of this problem. In case II, the asymptotic form of (60) implies that

$$\frac{3}{2}(1-\beta)(3-\beta)\sin\Phi F\bar{N} + 2(\sin\Phi F'\bar{N})' - 2\csc\Phi F\bar{P} + \beta G_0(\sin\Phi\bar{M})' = 0 \tag{107}$$

[1] We note that if $W = W(I_3)$ only, then $W_A \equiv 0$. Reanalysis of $(26)_1$ immediately results in $I(\Phi) =$ constant, which violates (80), and thus compressible materials of this form cannot support a tensile point load. As a related matter, this paper will not consider special compressible materials of the form $W = W(I_1)$ only.

while the boundary condition (63) gives

$$2F'\left(\frac{\pi}{2}\right)\bar{N}|_{\Phi=\frac{\pi}{2}} + \beta G_0 \bar{M}|_{\Phi=\frac{\pi}{2}} = 0. \tag{108}$$

Finally, case III gives

$$-2\csc\Phi F\bar{P} + \beta G_0(\sin\Phi\bar{M})' = 0 \tag{109}$$

along with the boundary condition

$$\bar{M}|_{\Phi=\frac{\pi}{2}} = 0. \tag{110}$$

We remark that the superposed bars on M, N, P denote that these are to be evaluated at G given by (103). Note that $\bar{M}$, $\bar{N}$, $\bar{P}$ also depend on the unknown $F(\Phi)$ and its derivative. In addition, in each of the above three cases, F must satisfy $(43)_{2,3}$ and (67), and thus we collect below the additional conditions, which require

$$F(0) = 0, \;\; F(\Phi) > 0, \;\; \text{and} \;\; F'(\Phi) > 0 \;\; \text{for } 0 < \Phi \leq \tfrac{\pi}{2}, \;\; FF'\csc\Phi \text{ bounded}. \tag{111}$$

On using (66), (82), and (96), we note that

$$m + \alpha > n + \beta \Leftrightarrow m > \frac{1}{2}\beta - \frac{5}{2} \Leftrightarrow m > -\frac{1}{2}n - 3, \tag{112}$$

and thus the three cases above correspond to

$$\begin{cases} \text{I:} & m > \dfrac{1}{2} - \dfrac{5}{2}\beta & \text{and} \quad p = -\beta - 1, \\ \text{II:} & m = \dfrac{1}{2} - \dfrac{5}{2}\beta & \text{and} \quad p = -\beta - 1, \\ \text{III:} & m \in \left(\dfrac{1}{2}\beta - \dfrac{5}{2}, \dfrac{1}{2} - \dfrac{5}{2}\beta\right) & \text{and} \quad p = m + \dfrac{3}{2}(\beta - 1). \end{cases} \tag{113}$$

Alternatively, in view of (82), we may write the above equivalently as

$$\begin{cases} \text{I:} & m > \dfrac{5}{2}n + 3 & \text{and} \quad p = n, \\ \text{II:} & m = \dfrac{5}{2}n + 3 & \text{and} \quad p = n, \\ \text{III:} & m \in \left(-\dfrac{1}{2}n - 3, \dfrac{5}{2}n + 3\right) & \text{and} \quad p = m - \dfrac{3}{2}n - 3. \end{cases} \tag{114}$$

We recall that it was shown earlier (see the discussion and analysis containing (83)–(94)) that a material for which $m < \frac{1}{2}\beta - \frac{5}{2}$ or, equivalently, $m < -\frac{1}{2}n - 3$ cannot support a tensile point load. When $m = \frac{1}{2}\beta - \frac{5}{2} = -\frac{1}{2}n - 3$, the material must be such that $p = 2\beta - 4 = -2(n + 3)$ to sustain the point load. We present below tables summarizing the main tests to determine whether a material can sustain a tensile point load (Table 1) and the remaining differential equation(s) and boundary condition(s) to solve for the complete asymptotic solution when the point load can be supported (Tables 2 and 3). We remind the reader that n, m, and p are determined from a given material model W via (74)–(76), respectively, and that β is determined from (82), i.e., $n + \beta + 1 = 0$, which always applies.

TABLE 1
Asymptotic tests for the ability of a compressible material to support a tensile point load.

$n(\beta)+\beta+1=0$	Material may sustain a point load	Material cannot sustain a point load
β	$\in(0,1)$	$\notin(0,1)$
m	$\geq \frac{1}{2}\beta-\frac{5}{2}=-\frac{1}{2}n-3$	$<\frac{1}{2}\beta-\frac{5}{2}=-\frac{1}{2}n-3$
$m=\frac{1}{2}\beta-\frac{5}{2}=-\frac{1}{2}n-3$	$p=2\beta-4$ $=-2(n+3)$ (Case IV)	$p\neq 2\beta-4$ $=-2(n+3)$
$m\in(\frac{1}{2}\beta-\frac{5}{2},\frac{1}{2}-\frac{5}{2}\beta)$ $=(-\frac{1}{2}n-3,\frac{5}{2}n+3)$	$p=m+\frac{3}{2}(\beta-1)$ $=m-\frac{3}{2}n-3$ (Case III)	$p\neq m+\frac{3}{2}(\beta-1)$ $=m-\frac{3}{2}n-3$
$m=\frac{1}{2}-\frac{5}{2}\beta=\frac{5}{2}n+3$	$p=-\beta-1=n$ (Case II)	$p\neq-\beta-1=n$
$m>\frac{1}{2}-\frac{5}{2}\beta=\frac{5}{2}n+3$	$p=-\beta-1=n$ (Case I)	$p\neq-\beta-1=n$

TABLE 2
Asymptotic equations when a material may support a point load.

Case	m	p	Unknown(s)	ODE(s)	BC(s)
I	$>\frac{1}{2}-\frac{5}{2}\beta$	$=-\beta-1$	$F(\Phi)$	(105)	(106)
II	$=\frac{1}{2}-\frac{5}{2}\beta$	$=-\beta-1$	$F(\Phi)$	(107)	(108)
III	$\in(\frac{1}{2}\beta-\frac{5}{2},\frac{1}{2}-\frac{5}{2}\beta)$	$=m+\frac{3}{2}(\beta-1)$	$F(\Phi)$	(109)	(110)
IV	$=\frac{1}{2}\beta-\frac{5}{2}$	$=2\beta-4$	$F(\Phi)$, $G(\Phi)$, $I(\Phi)$	(93), (85), (84)	$(43)_2$, $(44)_2$, (80), $(91)_2$

TABLE 3
Asymptotic equations when a material may support a point load.

Case	m	p	Unknown(s)	ODE(s)	BC(s)
I	$>\frac{5}{2}n+3$	$=n$	$F(\Phi)$	(105)	(106)
II	$=\frac{5}{2}n+3$	$=n$	$F(\Phi)$	(107)	(108)
III	$\in(-\frac{1}{2}n-3,\frac{5}{2}n+3)$	$=m-\frac{3}{2}n-3$	$F(\Phi)$	(109)	(110)
IV	$=-\frac{1}{2}n-3$	$=-2(n+3)$	$F(\Phi)$, $G(\Phi)$, $I(\Phi)$	(93), (85), (84)	$(43)_2$, $(44)_2$, (80), $(91)_2$

Table 1 should be interpreted as presenting sequential tests on the material to determine whether it may sustain a tensile point load. Thus, in testing a particular material, one must first ensure that $\beta\in(0,1)$, then determine if $m\geq\frac{1}{2}\beta-\frac{5}{2}=-\frac{1}{2}n-3$, and based on this result, continue with the appropriate test(s) for cases I–IV. If a material is not excluded somewhere along this process, then the material can support the point load provided that the remaining ordinary differential equation(s), boundary condition(s), and restrictions following Tables 2 and 3 can be satisfied. Tables 2 and 3 summarize the remaining differential equation(s) and boundary condition(s) when the material modeled by W may sustain a point load. Categories in Table 2 present the results when comparing m, p with β, while Table 3 presents the results when comparing m, p, and n.

For cases I–III, G and I are given by (103)–(104). In addition, we have the additional conditions in (111) for cases I–III, while for case IV, $(43)_3$ and (67) must hold.

5.2. Asymptotic tests: Applications. The asymptotic tests derived above can easily be applied to particular constitutive models in order to test a material's ability to support a finite deflection under a tensile point load. Below we present the results obtained from invoking our tests. The remainder of this subsection is divided into two sections according to hyperelastic materials that (i) cannot sustain the point load and (ii) may sustain the point load. The materials tested encompass many well-known compressible material models proposed in the literature.

5.2.1. Materials that cannot support a tensile point load: Examples.
(a) Special Hadamard materials of the form

$$W = c_1(I_1 - 3) + H(I_3), \quad c_1 > 0 \text{ constant}, \tag{115}$$

with $H(1) = 0$ and $c_1 + H'(1) = 0$ for vanishing stored-energy and stress, respectively, in the undeformed state: On substituting from (18) and (20) we can see directly that $W_A = c_1$, and thus by (75), $n = 0$ for materials of the form (115). However, from Table 1, (82) ($n + \beta + 1 = 0$) implies that $\beta = -1 \notin (0, 1)$, and thus we can immediately conclude that compressible materials of the form (115) cannot support a finite deflection under a tensile point load. A variety of strain-energy functions proposed in the literature, including models of Blatz, Ogden, and Christensen (see [10] and references cited therein), have the form (115).

(b) The generalized Blatz–Ko material [11]:

$$\begin{aligned} W = {} & \frac{\mu}{2} f \left(I_1 - 1 - \frac{1}{\nu} + \frac{1-2\nu}{\nu} I_3^{\frac{-\nu}{1-2\nu}} \right) \\ & + \frac{\mu}{2}(1-f) \left(\frac{I_2}{I_3} - 1 - \frac{1}{\nu} + \frac{1-2\nu}{\nu} I_3^{\frac{\nu}{1-2\nu}} \right) \\ & \text{with } \mu > 0, \quad 0 < \nu < \frac{1}{2}, \quad 0 \le f \le 1. \end{aligned} \tag{116}$$

On substituting from (18)–(20), differentiating, and employing (70), we have

$$W_A = \frac{\mu}{2}[f + (1-f)B^{-2}] \longrightarrow \frac{\mu}{2}[f + (1-f)\xi^{(1-\beta)}\bar{B}^{-2}]. \tag{117}$$

Considering (75), $f = 1$ implies $n = 0$, which was ruled out in (a) above, while $f = 0$ implies that $n = 1 - \beta$, which clearly violates (82). Finally, since by (68) we have $0 < \beta < 1$, the remaining case $0 < f < 1$ implies again that $n = 0$. Thus the generalized Blatz–Ko material cannot support a finite deflection under a tensile point load for *any* range of its parameters. We note that for the special Blatz–Ko material ($f = 0$, $\nu = \frac{1}{4}$) given by

$$W = \frac{\mu}{2} \left(\frac{I_2}{I_3} + 2I_3^{\frac{1}{2}} - 5 \right), \quad \mu > 0, \tag{118}$$

this result was obtained by a different argument in [3]. Here it follows immediately as a consequence of the above asymptotic tests. As another example, we note that the polynomial material proposed by Levinson and Burgess (see [10, reference [24]]) also contains parameters μ, ν, f and has W_A given exactly as in (117). Thus for the polynomial material, we obtain the same conclusion as that for the generalized Blatz–Ko material, regardless of the choice of material parameters.

(c) Generalized Hadamard materials:

$$W = H_1(I_3)(I_1 - 3) + H_2(I_3)(I_2 - 3) + H_3(I_3), \tag{119}$$

with $H_3(1) = 0$, $H_1(1)+2H_2(1)+H_3'(1) = 0$, and $H_1(1)+H_2(1) > 0$. Using (18)–(20) and (69)–(71), it follows that

$$W_A = H_1(B^2C) + H_2(B^2C)C \longrightarrow H_1(\bar{B}^2\bar{C}) + H_2(\bar{B}^2\bar{C})\xi^{(1-\beta)}\bar{C}. \tag{120}$$

From (68) and (75) it is follows immediately that $n = 0$ if $H_1(I_3) \not\equiv 0$, while $n = 1-\beta$ if $H_1(I_3) \equiv 0$, $H_2(I_3) \not\equiv 0$, and thus as before, we see that generalized Hadamard materials *cannot* support a tensile point load.

(d) A model proposed by Gao (see [8]):[2]

$$W = a\left[\left(\frac{I_1}{I_3^{\frac{1}{3}}}\right)^k - 3^k\right] + b(I_3-1)^j I_3^{-q}, \quad a,k>0, \quad b,j,q \geq 0. \tag{121}$$

The material (121) is used by Gao and Liu [8] in an asymptotic analysis of a rubber cone under a concentrated tensile force. On substituting from (18) and (20) into (121) and differentiating, we obtain

$$W_A = \frac{ak}{(B^2C)^{\frac{1}{3}}}\left[\frac{A}{(B^2C)^{\frac{1}{3}}} + \left(\frac{C}{B}\right)^{\frac{2}{3}}\right]^{k-1}, \tag{122}$$

while employing (69)–(71) and factoring results in

$$W_A \longrightarrow \xi^{2(\beta-1)(k-1)}\frac{ak}{(\bar{B}^2\bar{C})^{\frac{1}{3}}}\left[\frac{\bar{A}}{(\bar{B}^2\bar{C})^{\frac{1}{3}}}\right]^{k-1}[1+\eta]^{k-1}, \tag{123}$$

where

$$\eta = \xi^{3(1-\beta)}\left(\frac{\bar{C}}{\bar{B}}\right)^{\frac{2}{3}}\frac{(\bar{B}^2\bar{C})^{\frac{1}{3}}}{\bar{A}}. \tag{124}$$

Thus, in view of (68) and expanding (123) in powers of η, we obtain that, as $\xi \longrightarrow 0^+$, the asymptotic form (75) for W_A yields

$$n = 2(\beta-1)(k-1), \qquad N = \frac{ak\bar{A}^{k-1}}{(\bar{B}^2\bar{C})^{\frac{k}{3}}}, \tag{125}$$

where $\bar{A}$, $\bar{B}$, $\bar{C}$ are given by (72)–(74), respectively. Solving (82) determines

$$\beta = \frac{2k-3}{2k-1}, \quad \text{and so} \quad 0<\beta<1 \Leftrightarrow k > \frac{3}{2}. \tag{126}$$

Thus proceeding with Table 1, we must now determine m (see (76)) and test for $m \geq \frac{1}{2}\beta - \frac{5}{2}$ or, equivalently, $m \geq -\frac{1}{2}n - 3$. Calculating W_B, factoring, expanding, and taking the dominant term imply that for the material (121) we have (76) with

$$m = \frac{(\beta-1)(4k-1)}{2}, \qquad M = -\frac{2\bar{A}N}{3\bar{B}}, \tag{127}$$

[2]We note that the variable K is used in place of I_3 for the material model in [8]; however, K is readily seen to be equivalent to the I_3 in this paper by application of the Cayley–Hamilton theorem.

where β, k, N are as above. It is simple to check that $m = \frac{1-4k}{2k-1} = -\frac{n}{2} - 3$, which then determines that the remaining conditions are those of case IV. Thus we must next determine p (see (77)) and test for $p = 2\beta - 4$ or, equivalently, $p = -2(n+3)$. In a similar manner, we consider W_C and determine that the material (121) implies that (77) is such that

$$p = (\beta - 1)(2k + 1), \qquad P = -\frac{\bar{A}N}{3\bar{C}}, \tag{128}$$

from which it then follows that $p = -2\frac{(2k+1)}{2k-1} = -2(n+3)$, and so case IV is satisfied. Thus, turning to Tables 2 and 3, we must be able to satisfy the system of ordinary differential equations and boundary conditions imposed for this material, as well as the additional conditions $(43)_3$ and (67). However, one can deduce that this is not possible by examining the boundary condition $(91)_2$ (traction-free condition on the surface of the half-space), which from $(125)_2$ and $(127)_2$ becomes

$$M|_{\Phi=\frac{\pi}{2}} = \left.\frac{-2ak\bar{A}^k}{3\bar{B}(\bar{B}^2\bar{C})^{\frac{k}{3}}}\right|_{\Phi=\frac{\pi}{2}} = 0. \tag{129}$$

In view of $(43)_3$, $(44)_2$, (67), (72)–(74), and $(121)_2$, the traction-free boundary condition (129) is satisfied only if $\bar{B} \longrightarrow \pm\infty$ as $\Phi \longrightarrow \frac{\pi}{2}$; however, (67) and continuity of F would then imply that $F(\frac{\pi}{2}) = 0$, which violates $(43)_3$. Thus, under the hypotheses of this paper, a half-space consisting of the material (121) cannot support a finite deflection under a tensile point load and maintain a traction-free boundary.

(e) A model for biological tissue (see [10, 16] and references cited therein):

$$W = \frac{\gamma}{2}[f(I_3)e^{k(I_1-I_2)}(I_1 - 3) + g(I_3)e^{-k(I_1-I_2)}(I_2 - 3) + h(I_3)],$$

where $\gamma > 0, k \neq 0$ are constants, $h(1) = 0$, $f(1)+2g(1)+h'(1) = 0$, and $f(1)+g(1) > 0$. As was the case in [1] for the incompressible biological material model of Fung, W and its partial derivatives go to infinity faster than ξ raised to any power, and thus W is not integrable, which from (42) violates our second hypothesis H2 (displacements must be bounded everywhere). Thus the biological tissue model given above cannot support a tensile point load.

We remark that, as we shall continue to see in the remainder of this subsection, a great utility of the present treatment is the ability to test large classes of materials and quite simply determine whether they may be capable of sustaining the point load. As demonstrated above, many of the well-known models for compressible hyperelastic materials are *not* able to support a finite deflection under a tensile point load. This seems to be consistent with the results of [1] in the sense that, in [1], it was noted that incompressible hyperelastic materials were able to sustain the point load only when they were sufficiently stiff, and thus it is not surprising that many compressible materials fail to do so. To consider the variety of materials treated in this paper (or, similarly, in [1]) within the contexts of the analyses of [3] or [8] would require complete rederivations of the asymptotic forms of the governing equations and subsequent reanalysis for each material model.

5.2.2. Materials that may support a tensile point load: Examples.

(a) An Antman [12] material:

$$W = k_1\left(\frac{I_2}{I_3}\right)^{\mu_1} + k_2\left(\frac{I_1}{I_3}\right)^{\mu_2} + k_3 I_3^{-\mu_3} + k_4 I_1^{\frac{\nu}{2}} + k_5(I_1^2 - 2I_2) + k_6 I_2 + k_7 I_3, \tag{130}$$

where the k_i, μ_j, and ν are all constants satisfying the usual conditions for normalization and physically reasonable response (which we omit for our purposes here). The material (130) is used in [12] to study a class of boundary-value problems for nonlinearly elastic deformations, and it is also attractive for the problem considered in this paper. At this time, we shall not make an exhaustive study of the material (130) for our tensile point load problem, as the potential cases are too numerous for our purposes here, and instead determine the varied possibilities for the exponents n, m, p of (75)–(77). On substituting from (18)–(20), we represent (130) in the form

$$\begin{aligned} W = {} & k_1\left(\frac{A}{B^2}+\frac{1}{C}\right)^{\mu_1} + k_2\left(\frac{A}{B^2C}+\frac{1}{B^2}\right)^{\mu_2} + k_3(B^2C)^{-\mu_3} + k_4(A+C)^{\frac{\nu}{2}} \\ & + k_5(A^2-B^2+C^2) + k_6(AC+B^2) + k_7B^2C. \end{aligned} \tag{131}$$

On computing W_A, substituting from (69)–(71), factoring, and expanding, we obtain the asymptotic form

$$\begin{aligned} W_A \longrightarrow {} & \frac{k_1\mu_1}{\bar{B}^2}\left(\frac{\bar{A}}{\bar{B}^2}+\frac{1}{\bar{C}}\right)^{\mu_1-1}\xi^{(\beta-1)(\mu_1-2)} + \frac{k_2\mu_2}{\bar{B}^2\bar{C}}\left(\frac{\bar{A}}{\bar{B}^2\bar{C}}\right)^{\mu_2-1}\xi^{2(\beta-1)(\mu_2-1)} \\ & + \frac{1}{2}k_4\nu\bar{A}^{\frac{\nu-2}{2}}\xi^{(\beta-1)(\nu-2)} + 2k_5\bar{A}\xi^{2(\beta-1)} + k_6\bar{C}\xi^{1-\beta}. \end{aligned} \tag{132}$$

In view of (68), the final term in (132) can never be dominant and so the asymptotic form of (132) depends on which of the four remaining terms balance and/or dominate the others. There are 15 such possibilities, corresponding to the various cases within each of the four categories below:

$$\begin{cases} (n_1): & n = 2(\beta-1) \Rightarrow \beta = \dfrac{1}{3}, & \nu \le 4, & \mu_1 \le 4, & \text{and} & \mu_2 \le 2, \\ (n_2): & n = (\beta-1)(\nu-2) \Rightarrow \beta = \dfrac{\nu-3}{\nu-1}, & \nu > 4, & \mu_1 \le \nu, & \text{and} & \mu_2 \le \dfrac{\nu}{2}, \\ (n_3): & n = 2(\beta-1)(\mu_2-1) \Rightarrow \beta = \dfrac{2\mu_2-3}{2\mu_2-1}, & \nu < 2\mu_2, & \mu_1 \le 2\mu_2, & \text{and} & \mu_2 > 2, \\ (n_4): & n = (\beta-1)(\mu_1-2) \Rightarrow \beta = \dfrac{\mu_1-3}{\mu_1-1}, & \nu < \mu_1, & \mu_1 > 4, & \text{and} & \mu_2 < \dfrac{\mu_1}{2}. \end{cases} \tag{133}$$

In each equation of (133), we have $\beta \in (0,1)$, while N will vary depending on the case (n_1)–(n_4) and the particular combinations of restrictions placed on ν, μ_1, and μ_2. Next, to determine m, we calculate the asymptotic form of W_B, which yields

$$\begin{aligned} W_B \longrightarrow {} & -\frac{2k_1\mu_1\bar{A}}{\bar{B}^3}\left(\frac{\bar{A}}{\bar{B}^2}+\frac{1}{\bar{C}}\right)^{\mu_1-1}\xi^{\frac{1}{2}(\beta-1)(2\mu_1-1)} \\ & - \frac{2k_2\mu_2}{\bar{B}^{2\mu_2+1}}\left(\frac{\bar{A}}{\bar{C}}\right)^{\mu_2}\xi^{\frac{1}{2}(\beta-1)(4\mu_2-1)} \\ & + 2(k_6-2k_5)\bar{B}\xi^{\frac{1}{2}(\beta-1)}. \end{aligned} \tag{134}$$

The asymptotic form of (134) depends on which of its three terms balance and/or dominate the others, creating seven such possibilities corresponding to the follow-

ing cases:

$$
\begin{cases}
(m_1): & m = \dfrac{1}{2}(\beta-1), & \mu_1 \le 1, & \text{and} & \mu_2 \le \dfrac{1}{2}, \\
(m_2): & m = \dfrac{1}{2}(\beta-1)(4\mu_2-1), & \mu_1 \le 2\mu_2, & \text{and} & \mu_2 > \dfrac{1}{2}, \\
(m_3): & m = \dfrac{1}{2}(\beta-1)(2\mu_1-1), & \mu_1 > 1, & \text{and} & \mu_2 < \dfrac{\mu_1}{2}.
\end{cases} \tag{135}
$$

Again, M will vary with each of the cases (m_1)–(m_3) and the particular combinations of restrictions placed on μ_1 and μ_2. Finally, to determine p, the asymptotic form of W_C results in

$$
\begin{aligned}
W_C \longrightarrow\ & \frac{-k_1\mu_1}{\bar{C}^2}\left(\frac{\bar{A}}{\bar{B}^2}+\frac{1}{\bar{C}}\right)^{\mu_1-1}\xi^{(\beta-1)(\mu_1+1)} - \frac{k_2\mu_2}{\bar{C}}\left(\frac{\bar{A}}{\bar{B}^2\bar{C}}\right)^{\mu_2}\xi^{(\beta-1)(2\mu_2+1)} \\
& + \frac{1}{2}k_4\nu\bar{A}^{\frac{\nu-2}{2}}\xi^{(\beta-1)(\nu-2)} + k_6\bar{A}\xi^{2(\beta-1)},
\end{aligned} \tag{136}
$$

and thus we have four possibilities for p corresponding to the 15 combinations of the parameters ν, μ_1, and μ_2 (and thus P) given below:

(137)

$$
\begin{cases}
(p_1): & p = 2(\beta-1), & \nu \le 4, & \mu_1 \le 1, & \text{and} & \mu_2 \le \dfrac{1}{2}, \\
(p_2): & p = (\beta-1)(\nu-2), & \nu > 4, & \mu_1 \le \nu-3, & \text{and} & \mu_2 \le \dfrac{\nu-3}{2}, \\
(p_3): & p = (\beta-1)(2\mu_2+1), & \nu < 2\mu_2+3, & \mu_1 \le 2\mu_2, & \text{and} & \mu_2 > \dfrac{1}{2}, \\
(p_4): & p = (\beta-1)(\mu_1+1), & \nu < \mu_1+3, & \mu_1 > 1, & \text{and} & \mu_2 < \dfrac{\mu_1}{2}.
\end{cases}
$$

Thus the material (130) offers an array of possibilities for study of the asymptotic equations associated with this problem, which, as mentioned above, we shall explore elsewhere.

(b) A recent model due to Gao [13]:[3]

$$
W = a\left(I_1^k + \left(\frac{I_2}{I_3}\right)^k\right), \quad a, k > 0 \text{ constants.} \tag{138}
$$

Gao [13] studies the asymptotic large deformation elastostatic field near a crack tip for a compressible hyperelastic material described by (138). For the material (138), the asymptotic form (75) for W_A is such that

$$
n = 2(\beta-1)(k-1), \qquad N = ak\bar{A}^{k-1}, \tag{139}
$$

where $\bar{A}$ is given by (72). Solving (82) determines

$$
\beta = \frac{2k-3}{2k-1}, \quad \text{and so} \quad 0 < \beta < 1 \Leftrightarrow k > \frac{3}{2}. \tag{140}
$$

[3]We remark that this material does not satisfy the usual normalization condition of zero strain-energy in the undeformed state ($W(3,3,1) = 0$). A simple remedy would be, e.g., to include a term of the form $-2a3^k$ in (138) or replace I_1, I_2 with $I_1 - 3$, $I_2 - 3$, respectively.

Thus, for the point load to be supported, we must have $k > \frac{3}{2}$ in (138). In continuing with the asymptotic tests, one finds that when $\frac{3}{2} < k < 3$, the corresponding m and p satisfy case III; when $k = 3$, m and p satisfy case II; and when $k > 3$, m and p satisfy case I. However, upon further examination, one finds that the boundary conditions (106) and (110) consistent with cases I and III, respectively, cannot be satisfied for the associated $\bar{N}$ and $\bar{M}$ in these cases. Thus, under the hypotheses of this paper, a half-space consisting of the material (138) with $\frac{3}{2} < k < 3$ or $k > 3$ cannot support a finite deflection under a tensile point load and maintain a traction-free boundary. The boundary condition (108) for the case $k = 3$, however, is not incompatible with the associated $\bar{N}$ and $\bar{M}$. In this case, it can be seen that

$$k = 3 \Rightarrow \beta = \frac{3}{5}, \qquad m = -1, \qquad n = p = -\frac{8}{5}, \qquad \alpha = \frac{6}{5} \tag{141}$$

and that

$$\bar{N} = 3a\bar{A}^2, \quad \bar{M} = -6a\frac{\bar{A}}{\bar{B}^3}\left(\frac{\bar{A}}{\bar{B}^2} + \frac{1}{\bar{C}}\right)^2, \quad \bar{P} = 3a\left[\bar{A}^2 - \frac{1}{\bar{C}^2}\left(\frac{\bar{A}}{\bar{B}^2} + \frac{1}{\bar{C}}\right)^2\right], \tag{142}$$

where

$$\bar{A} = (\beta G_0)^2 = (12\pi a)^{-\frac{2}{5}}, \tag{143}$$

$$\bar{B} = \beta G_0 F'(\Phi) = (12\pi a)^{-\frac{1}{5}} F'(\Phi), \tag{144}$$

$$\bar{C} = \frac{F^2(\Phi)}{\sin^2 \Phi} \tag{145}$$

and

$$G(\Phi) \equiv G_0 = \frac{5}{3}(12\pi a)^{-\frac{1}{5}}, \tag{146}$$

$$I(\Phi) = \frac{1}{2\pi}(1 - \cos\Phi). \tag{147}$$

The remaining equations and conditions that must be satisfied are the ordinary differential equation (107), which in this case results in the nonlinear second-order differential equation for $F(\Phi)$ given by

$$\begin{aligned}(12\pi a)^{-\frac{4}{5}} &\left[2(\sin\Phi F')' + \frac{36}{25}\sin\Phi F - \frac{2F}{\sin\Phi}\right] \\ &+ 2\left\{\left(\frac{\sin\Phi}{F}\right)^3\left[\frac{1}{(F')^2} + \frac{\sin^2\Phi}{F^2}\right]^2 - \left(\frac{\sin\Phi}{(F')^3}\left[\frac{1}{(F')^2} + \frac{\sin^2\Phi}{F^2}\right]^2\right)'\right\} = 0;\end{aligned} \tag{148}$$

the traction-free boundary condition (108), which implies that

$$(12\pi a)^{-\frac{2}{5}}\left[F'\left(\frac{\pi}{2}\right)\right]^2 = \frac{1}{\left[F'\left(\frac{\pi}{2}\right)\right]^2} + \frac{1}{\left[F\left(\frac{\pi}{2}\right)\right]^2}; \tag{149}$$

and the conditions contained in (111), namely,

$$F(0) = 0, \quad F(\Phi) > 0, \quad \text{and} \quad F'(\Phi) > 0 \quad \text{for } 0 < \Phi \le \tfrac{\pi}{2}, \quad FF'\csc\Phi \text{ bounded}. \tag{150}$$

(c) A Jiang–Ogden [14] material:

$$W = f(I_1)h_1(I_3) + h_2(I_3), \tag{151}$$

where $f(3)h_1(1) + h_2(1) = 0$, $f'(3)h_1(1) + f(3)h_1'(1) + h_2'(1) = 0$, and $f'(3)h_1(1) > 0$. Jiang and Ogden [14] consider materials of the form (151) in their study of azimuthal shear of circular cylindrical tubes, where the function $f(I_1)$ has the form $f(I_1) = c_1(I_1 - 1)^j$ with $j \geq \frac{1}{2}$. Here we shall similarly treat materials of the form (151) with $f(I_1) = c_1(I_1 - 1)^k$, $k > 0$ constant. Thus we consider

$$W = c_1(I_1 - 1)^k h_1(I_3) + h_2(I_3), \quad k > 0, \tag{152}$$

for which the asymptotic tests developed earlier determine the following:

$$n = 2(\beta - 1)(k - 1), \qquad N = c_1 k \bar{A}^{k-1} h_1(\bar{B}^2\bar{C}), \tag{153}$$

where $\bar{A}$, $\bar{B}$, $\bar{C}$ are given by (72)–(74), respectively. Solving (82) determines

$$\beta = \frac{2k-3}{2k-1}, \quad \text{and so} \quad 0 < \beta < 1 \Leftrightarrow k > \frac{3}{2}, \tag{154}$$

and via (66), $\alpha = \frac{3-\beta}{2}$. We note that [14] treats four special cases of (152) with $k = \frac{1}{2}$, 1, $\frac{3}{2}$, $\frac{3}{4}$, and thus by $(154)_2$, none of these special cases can support a tensile point load. Continuing with $k > \frac{3}{2}$, it can be seen that

$$m = \frac{(\beta-1)(4k-1)}{2} = \frac{1}{2}\beta - \frac{5}{2}, \qquad M = 2c_1\bar{A}^k\bar{B}\bar{C}h_1'(\bar{B}^2\bar{C}) \tag{155}$$

and

$$p = (\beta - 1)(2k + 1) = 2\beta - 4, \qquad P = c_1\bar{A}^k\bar{B}^2h_1'(\bar{B}^2\bar{C}), \tag{156}$$

and thus case IV is satisfied. By Tables 2 and 3, we must then satisfy the system of ordinary differential equations and boundary conditions imposed for this material, as well as the additional conditions $(43)_3$ and (67). We note here that the material (121) is a special case of (151), and we recall from section 5.2.1(d) that the traction-free boundary condition $(91)_2$ ($M|_{\Phi=\frac{\pi}{2}} = 0$) could not be satisfied for (121). Noting $(155)_2$ and the fact that $\bar{A}$, $\bar{B}$, and $\bar{C}$ are all strictly positive at $\Phi = \frac{\pi}{2}$, $(91)_2$ can then be seen as a condition on the form of the function $h_1(I_3)$ such that a traction-free boundary is possible. As an example, note that $I_3 = B^2C = \bar{B}^2\bar{C}$ and that

$$I_3|_{\Phi=\frac{\pi}{2}} = c \tag{157}$$

for some $c > 0$ constant. Suppose now that

$$h_1(I_3) = \frac{1}{c}I_3\left[\ln\left(\frac{I_3}{c}\right) - 1\right], \tag{158}$$

and so

$$h_1'(I_3) = \frac{1}{c}\ln\left(\frac{I_3}{c}\right), \tag{159}$$

which, in turn, implies via (157) that

$$h_1'(\bar{B}^2\bar{C})|_{\Phi=\frac{\pi}{2}} = 0, \tag{160}$$

and thus the traction-free boundary condition $(91)_2$ is identically satisfied for the material (152) with $h_1(I_3)$ given by (158), where c is as in (157). With this choice for $h_1(I_3)$ and N, M, P, β, and α given above, the remaining equations to be satisfied are (93), (85), and (84), along with boundary conditions $(43)_2$, $(44)_2$, and (80), with the additional conditions $(43)_3$ and (67).

(d) A Beatty–Jiang-type material [15]:

(161)
$$\begin{aligned} W = {} & H_1(I_3)(I_1-3) + H_2(I_3)(I_2-3) + H(I_3) \\ & + c_3(I_1-3)^2 + c_4(I_2-3)^2 + c_5(I_1-3)(I_2-3) + c_6(I_1-3)^3 \\ & + c_7(I_2-3)^3 + c_8(I_1-3)(I_2-3)^2 + c_9(I_1-3)^2(I_2-3) + c_{10}(I_1-3)(I_2-3)^3, \end{aligned}$$

where the c_i are all constants. Beatty and Jiang [15] consider the material (161) with $c_{10} = 0$ in their study of azimuthal shear of nonlinearly elastic compressible solids. When $c_{10} = 0$, the material (161) satisfies the conditions on n, β, m, and p for case III; however, again we find that the associated traction-free boundary condition (110) cannot be satisfied. Thus we augment the material considered in [15] by including the c_{10} term so that (110) might be satisfied. Following the considerations of [15], we require $H_3(1) = 0$, $H_1(1) + 2H_2(1) + H_3'(1) = 0$, $H_1(1) + H_2(1) > 0$, $c_3 + c_4 + c_5 \geq 0$, and $c_6 + c_7 + c_8 + c_9 + c_{10} \geq 0$ for physically reasonable material response. Proceeding with our analysis, we find that

$$n = 4(\beta - 1), \qquad N = 3c_6\bar{A}^2, \tag{162}$$

while solving (82) determines

$$\beta = \frac{3}{5} \Rightarrow n = -\frac{8}{5} \quad \text{and} \quad \alpha = \frac{6}{5}. \tag{163}$$

In addition, one can determine that

$$m = \frac{9}{2}(\beta - 1) = -\frac{9}{5}, \qquad M = 2\bar{A}\bar{B}[c_9\bar{A} + 3c_{10}(\bar{A}\bar{C} + \bar{B}^2)^2], \tag{164}$$

where we recall that $\bar{A}$, $\bar{B}$, $\bar{C}$ are given by (72)–(74), respectively, and

$$\begin{aligned} p &= 6(\beta - 1) = -\frac{12}{5} = m + \frac{3}{2}(\beta - 1), \\ P &= \bar{A}^2[c_9\bar{A} + 3c_{10}(\bar{A}\bar{C} + \bar{B}^2)^2] = \frac{\bar{A}}{2\bar{B}}M. \end{aligned} \tag{165}$$

Thus case III is satisfied, and it then follows that

$$G(\Phi) \equiv G_0 = \frac{5}{3}(12\pi c_6)^{-\frac{1}{3}}, \tag{166}$$

$$I(\Phi) = \frac{1}{2\pi}(1 - \cos\Phi). \tag{167}$$

Further, from Tables 2 and 3, $F(\Phi)$ must then satisfy the conditions in (111) as well as the ordinary differential equation (109) and traction-free boundary condition (110), which for the material (161) become

$$\begin{aligned} & -\frac{F}{\sin\Phi}\left[c_9 + 3c_{10}\beta^2 G_0^2\left(\frac{F^2}{\sin^2\Phi} + F'^2\right)^2\right] \\ & \quad + \left\{\sin\Phi F'\left[c_9 + 3c_{10}\beta^2 G_0^2\left(\frac{F^2}{\sin^2\Phi} + F'^2\right)^2\right]\right\}' = 0 \end{aligned} \tag{168}$$

and

$$\left[F^2\left(\frac{\pi}{2}\right)+F'^2\left(\frac{\pi}{2}\right)\right]^2=-\frac{c_9}{3c_{10}\beta^2G_0^2}, \tag{169}$$

respectively, where by $(163)_1$ and (166), $\beta^2G_0^2=(12\pi c_6)^{-\frac{2}{3}}$. To conclude this paper, we consider the special case of (168), (169) when

$$\frac{F^2}{\sin^2\Phi}+F'^2=c^2,\quad c\neq 0 \text{ constant}, \tag{170}$$

and

$$c_9=-3c_{10}\beta^2G_0^2c^4. \tag{171}$$

Since (168), (169) follow automatically if (170), (171) hold, we will take c_9 in (161) to be given by (171) and consider the first-order nonlinear ordinary differential equation (170) subject to the conditions in (111), which we repeat here as

$$F(0)=0,\quad F(\Phi)>0,\quad \text{and}\quad F'(\Phi)>0\quad \text{for } 0<\Phi\leq\tfrac{\pi}{2},\quad FF'\csc\Phi \text{ bounded}. \tag{172}$$

While (170) is a significant simplification over (168) and appears deceptively simple, we have not found (170) to be amenable to analysis for determining a closed-form solution. Instead, we present in Figures 1 and 2 plots of the direction fields along with the solutions of (170) satisfying $(172)_1$, as well as the remaining conditions in (172), for representative values of the constant $c>1$. We remark that Figures 1 and 2 were generated for simplicity using Maple. The solutions plotted for various c agree with data obtained from computing a numerical solution implementing a fourth-order Runge–Kutta backwards shooting method.

Thus with our numerical solution for $F(\Phi)$, the unknown functions and parameters from (43)–(45) are determined according to (163) for α and β and (166) and (167) for G and I, respectively, where we recall $\delta=0$ from (49). In addition, with (170), (171), we have $\bar{M}=\bar{P}=0$, and thus the asymptotic forms of the nonzero stress components are as follows (where we note that $S_{\Phi z}=0$, $S_{\Theta\theta}=0$):

$$S_{\xi r}\longrightarrow \xi^{-\frac{7}{5}}\frac{12}{5}\bar{N}F(\Phi), \tag{173}$$

$$S_{\xi z}\longrightarrow \xi^{-2}\frac{6}{5}\bar{N}G_0, \tag{174}$$

$$S_{\Phi r}\longrightarrow \xi^{\frac{1}{5}}2\bar{N}F'(\Phi), \tag{175}$$

where by (162), (73), and (166), $\bar{N}=3c_6\bar{A}^2=3c_6(\beta G_0)^4$ is constant. The asymptotic forms of the nonzero stress components, on using the more traditional cylindrical polar coordinates and bases for both the undeformed and deformed configurations, are then computed for this problem as

$$S_{Rr}\longrightarrow \sin\Phi S_{\xi r}, \tag{176}$$

$$S_{Rz}\longrightarrow \sin\Phi S_{\xi z}, \tag{177}$$

$$S_{Zr}\longrightarrow \cos\Phi S_{\xi r}, \tag{178}$$

$$S_{Zz}\longrightarrow \cos\Phi S_{\xi z}, \tag{179}$$

where $S_{\xi r}$, $S_{\xi z}$ are given in (173)–(174) and $\xi\sin\Phi=R$, $\xi\cos\Phi=Z$, $\xi^2=R^2+Z^2$.

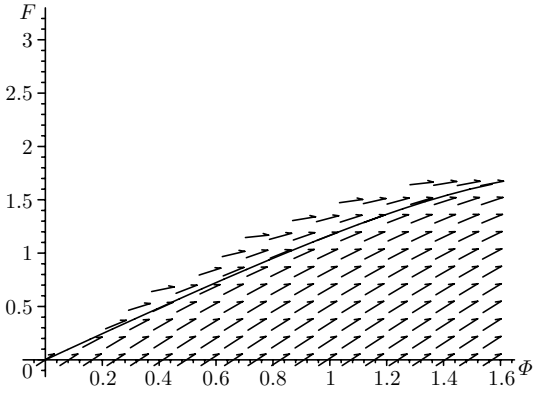


FIG. 1. *Plot of the direction field and solution of* (170) *when* $c^2 = 3$.

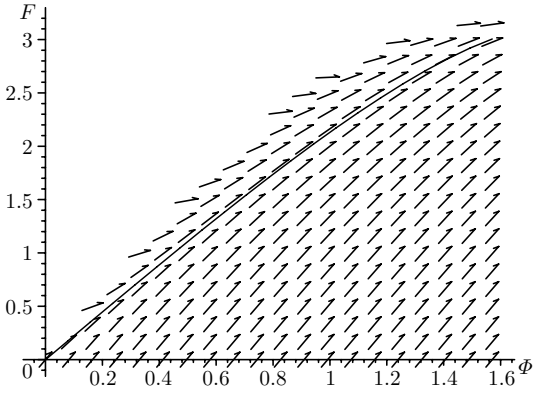


FIG. 2. *Plot of the direction field and solution of* (170) *when* $c^2 = 10$.

REFERENCES

[1] D. A. POLIGNONE WARNE AND P. G. WARNE, *An asymptotic finite deformation analysis for an isotropic incompressible hyperelastic half-space subjected to a tensile point load*, SIAM J. Appl. Math., 62 (2001), pp. 107–128.

[2] S. P. TIMOSHENKO AND J. N. GOODIER, *Theory of Elasticity*, 3rd ed., McGraw–Hill, New York, 1970.

[3] J. G. SIMMONDS AND P. G. WARNE, *Notes on the nonlinearly elastic Boussinesq problem*, J. Elasticity, 34 (1994), pp. 69–82.

[4] P. CHADWICK, *Applications of an energy-momentum tensor in non-linear elastostatics*, J. Elasticity, 5 (1975), pp. 249–258.

[5] J. K. KNOWLES AND E. STERNBERG, *An asymptotic finite-deformation analysis of the elastostatic field near the tip of a crack*, J. Elasticity, 3 (1973), pp. 67–107.

[6] J. K. KNOWLES AND E. STERNBERG, *Finite deformation analysis of the elastostatic field near the tip of a crack: Reconsideration and higher-order results*, J. Elasticity, 4 (1974), pp. 201–233.

[7] J. K. KNOWLES, *The finite anti-plane shear field near the tip of a crack for a class of incompressible elastic solids*, Internat. J. Fracture, 13 (1977), pp. 611–639.

[8] Y. C. GAO AND B. LIU, *A rubber cone under the tension of a concentrated force*, Internat. J. Solids Structures, 32 (1995), pp. 1485–1493.

[9] R. W. OGDEN, *Non-Linear Elastic Deformations*, Ellis Horwood, Chichester, UK, 1984.

[10] D. A. POLIGNONE AND C. O. HORGAN, *Axisymmetric finite anti-plane shear of compressible nonlinearly elastic circular tubes*, Quart. Appl. Math., 50 (1992), pp. 323–341.

[11] P. J. Blatz and W. L. Ko, *Application of finite elasticity to the deformation of rubbery materials*, Trans. Soc. Rheol., 6 (1962), pp. 223–251.
[12] S. S. Antman, *Regular and singular problems for large elastic deformations of tubes, wedges, and cylinders*, Arch. Rational Mech. Anal., 83 (1983), pp. 1–52.
[13] Y. C. Gao, *Large deformation field near a crack tip in rubber-like material*, Theoret. Appl. Fract. Mech., 26 (1997), pp. 155–162.
[14] X. Jiang and R. W. Ogden, *On azimuthal shear of a circular cylindrical tube of compressible elastic material*, Quart. J. Mech. Appl. Math., 51 (1998), pp. 143–158.
[15] M. F. Beatty and Q. Jiang, *On compressible materials capable of sustaining axisymmetric shear deformations, part 2: Rotational shear of isotropic hyperelastic materials*, Quart. J. Mech. Appl. Math., 50 (1997), pp. 212–237.
[16] M. F. Beatty and Q. Jiang, *On compressible materials capable of sustaining axisymmetric shear deformations, part 3: Helical shear of isotropic hyperelastic materials*, Quart. Appl. Math., 57 (1999), pp. 681–697.

SIAM J. APPL. MATH.
Vol. 63, No. 1, pp. 195–221

HYDRODYNAMIC CLEANSING OF PULMONARY ALVEOLI*

DAPHNE ZELIG† AND SHIMON HABER‡

Abstract. The inside wall of the pulmonary alveolus is lined with a thin viscous fluid layer and a monolayer of surfactants. Inhaled foreign particles that reach the lung alveoli are normally neutralized by macrophages and remain inside the lung. Nevertheless, Podgorski and Gradon [*Ann. Occup. Hyg.*, 37 (1993), pp. 347–365] suggested that a hydrodynamic cleansing mechanism may exist in which particles are swept out by the net fluid flow from the alveolar viscous layer to the adjacent airways. Hawgood [*The Lung: Scientific Foundations*, 2nd ed., R. G. Crystal and J. B. West, eds., Lippincott-Raven, Philadelphia, 1997, pp. 557–571] has also reported that surfactants exit the alveoli during every breathing period. Based upon the foregoing observations, we examine a possible mechanism of hydrodynamic cleansing and predict its effectiveness. Our central assumption is that the amount of surfactant remains periodic during breathing and that a certain regulatory mechanism exists that causes excess surfactant (reported by Hawgood) to leave the alveoli. Owing to the latter, surfactant concentration gradients are induced inside the alveoli, which in turn generate fluid motion (a Marangoni effect) and concomitant fluid discharge. Our analysis predicts that a typical value of the outflow velocity is 10^{-9}[m/sec]; i.e., it takes a fluid particle almost two days to travel a distance equal to an alveolar radius. It is also shown that the outflow velocity depends almost linearly on the discharge rate of the surfactants. Hence, a small artificial addition of surfactants into the lung may enhance alveolar cleansing, provided that a biological mechanism exists that maintains normal surfactant concentration over the lining fluid layer.

Key words. lung alveoli, hydrodynamic cleansing, surfactants

AMS subject classifications. 76Z05, 92C35

PII. S0036139901386090

1. Introduction.

Zeltner et al. [29] observed that a nonuniform pattern of particle deposition exists within the rodent lung. Specifically, the density of particles deposited on the alveolar entrance rim is five times higher than that on septal alveolar surfaces. Are hydrodynamic forces driving the particles from their initial deposition locations toward the entrance rim? The fluid dynamical problem addressed in this paper is motivated by the search for such a possible cleansing mechanism inside the lung alveoli.

Environmental and occupational hazards resulting from aerosol inhalation have been the subject of intensive research (see Harvey and Crystal [14]). An understanding of aerosol kinetics may also prove to be a meaningful step towards improving diagnostic and therapeutic methods [1], [5]. In humans, the respiratory airway system consists of the nasal cavity, the throat, the voice box, the trachea, the two primary bronchi that bifurcate from the trachea, the bronchi, and bronchiole that divide and subdivide, becoming steadily smaller until there are about 20–23 generations of branching. From the sixteenth generation, the airways become increasingly alveolated. The bronchioles terminate with berry-shaped group of sacs and acinar ducts (the acinus). During breathing, the alveoli and the alveolar ducts expand and contract in a way roughly consistent with geometric similarity. Thus, all dimensions scale approximately as the 1/3 power of the lung volume (Gil and Weibel [8], Gil et al. [7]; Weibel [27]; Ardila,

*Received by the editors March 7, 2001; accepted for publication (in revised form) April 2, 2002; published electronically August 28, 2002.
http://www.siam.org/journals/siap/63-1/38609.html
†Department of Mathematics, Technion, Israel Institute of Technology, Haifa 32000, Israel (motiz@012.net.il).
‡Department of Mechanical Engineering, Technion, Israel Institute of Technology, Haifa 32000, Israel (mersh01@tx.technion.ac.).

Horie, and Hildebrandt [2]). Tsuda, Henry, and Butler [25], Tsuda, Otani, and Butler [26], and recently Haber et al. [12] considered the effect of alveolar expansion and contraction on the fluid flow inside the alveoli. In [25] and [26], the authors assumed that the pulmonary acinus could be viewed as a self-similar expanding axisymmetric thoroughfare surrounded by a toroidal sac, a configuration that simplified the numerical calculations. In [12], the alveolus was geometrically approximated by a self-similar expanding spherical cap attached at its rim to the alveolar duct (see also Gil et al. [7]), a geometry that is likely to represent a more faithful portrayal of the acinus.

Little attention has been paid in the past to the effect of alveolar expansion and contraction, since in the case of gas exchange the Peclet number controlling the transport of the gas molecules is much smaller than unity. Thus, convection due to the acinar flow is negligibly small when compared with the diffusive transport. (It takes only a few milliseconds for a gas molecule to reach the alveolar wall from its entrance ring.) However, in the case of aerosol transport, the Peclet number is much larger, and particle convection and diffusion may play a comparable role. Under normal conditions, particles 0.5 to 4 μm in diameter may often reach the acinus and pose the greatest hazard to human health (see, e.g., Dockery et al. [6]).

Particles that enter the respiratory system and are deposited over the airway walls are mechanically removed by the rhythmical motion of cilia (Sleigh, Blake, and Liron [20]). Particles are forced upwards along the bronchiolar tree and are finally removed from the respiratory system by forced convection of air (coughing). Nonetheless, a similar cleansing mechanism does not exist within the acinus. Generally, particles that reach the alveoli are neutralized by macrophages [4] and remain deposited inside the acinus. Indeed, several experimental studies (e.g., Zeltner et al. [29], Heyder et al. [16], Schultz et al. [24]) have investigated such aerosol mixing and deposition. However, Gradon and Podgorski [11] proposed that a purely hydrodynamic effect may assist in cleansing the alveoli. They suggested that gradients in surfactant concentrations induce the thin fluid lining that covers the inner alveolar wall to flow slowly outside the alveolus rim. Thus, particles deposited on the alveolar wall are carried with the fluid toward the entrance rim. They predicted a characteristic clearance time of about one hour.

Scarpelli [23] described the main stages of the surfactant's transition between the air-fluid interface and the fluid body as follows: During expiration, the alveolus contracts and the distance between the surfactant molecules decreases; in other words, their concentration increases and consequently the surface tension diminishes. When the alveolar radius reaches a threshold value, some of the molecules of the surfactant leave the interface and penetrate the fluid. During inspiration, the alveolus expands, the concentration of the surfactant decreases, and concomitantly, the surface tension increases. In addition, surfactants return to the interface from the bulk of the fluid by diffusion. More detailed models for surfactant transition can be found in [9], [10].

In [15], the metabolism of surfactants is explained, and the secretion rate is evaluated. There is clear evidence for the existence of a regulatory mechanism for surfactant production and clearance rates that keeps it from excessive accumulation or dilution [15]. Surfactants are created in Type II cells, which form part of the alveolus wall. After diffusing to the interface, most of them (about 80%) return to these cells and are then recycled for additional use. About 10–20% are consumed by macrophages, which lie at the alveolar wall, and the remaining few percent exit the alveolus.

In this article the alveolar hydrodynamic clearance mechanism is analyzed. We adopt the spherical model that has been extensively used in the past (e.g., Podgorski

and Gradon [22] and Haber et al. [12]) to describe the alveolus configuration. We also use measured experimental data for alveoli expansion/contraction rates and the known measured correlation between surfactant concentration and surface tension. We focus on the dynamical behavior of the surfactants, the main mechanism that controls the lining fluid flow, and assume that no surfactants are accumulated or depleted inside the alveolus during a breathing cycle. The boundary condition at the alveolar rim is based upon the known experimental value of the small amount of surfactant exiting the alveolus per breathing cycle. An open and valid question is what the specific mechanism that causes surfactant to exit the alveolus might be. One might assume, for instance, that airflow in the adjacent airway contributes to the sweeping effect, and reformulate the boundary conditions accordingly. Another possibility is that there is a biological mechanism that discharges excess surfactant from inside the alveolus. We try to avoid such ad hoc assumptions and focus on a cleansing mechanism that is based upon known and validated experimental data. The solution methodology is based on the assumption that, had no fluid been driven through the alveolus opening, surfactant concentration would have been uniform and the lining fluid would have expanded and contracted in a radially symmetric manner to conserve fluid mass. Thus, scaling of the cleansing mechanism is based upon the amount of surfactant leaving the alveolus, a markedly different approach from that used by Podgorski and Gradon [22], who relate the continuity of the fluid and the surfactant layers at the alveolar rim. A source term is also added to the surfactant mass conservation equation to account for surfactants entering or leaving the interface from the bulk fluid, and this facilitates the condition that no mass accumulation or depletion of surfactants per cycle occurs. As a result, the whole set of equations and the solution differ markedly from that obtained by Podgorski and Gradon [22].

In section 2, we define the geometrical, kinematical, and physiological parameters that scale the variables of the problem. In section 3, we obtain the resulting governing equations, boundary conditions, and parameters that control the problem. In section 4, we present the asymptotic expansion of the flow variables in terms of two smallness parameters and obtain the equations and boundary conditions that govern the zero and first order approximations. In section 5.1, we present the analytic solution of the zero order approximation, and in section 5.2, a finite element analysis is utilized to obtain a solution for the first order approximation. In section 6, we discuss our results, and we present our concluding remarks in section 7.

2. The alveolus model: Configuration and typical parameters. Assume that the alveolus can be approximated by a hollow spherical cap of radius $R(t)$ attached at its rim to the alveolar duct (see Figure 1). Typical alveolar mean radius ranges between 40 and 200μm. The alveolus is rhythmically expanding and contracting with a breathing rate of 12–14 breaths per minute for adults and about 33 breaths per minute for infants. The expansion amplitude is about 0.1 alveolar radius. The dependence of R on time is based on experimental data described in Podgorski and Gradon [22] and approximated here by a natural cubic spline interpolation to achieve continuity of its time derivatives (see Figure 2(a)).

A spherical coordinate system (r, θ, ϕ) is located at the center of the spherical cap, where r stands for the radial coordinate and θ and ϕ denote the latitude and azimuthal angles, respectively.

The alveolus rim location is defined by the half-cone angle θ_b. (Henceforth, we assume that $\theta_b = 60^0$ and that the subscript "b" denotes evaluation at the rim.) The inside wall of the alveolus is lined with a thin fluid layer of thickness $h(\theta, \phi, t)$. The

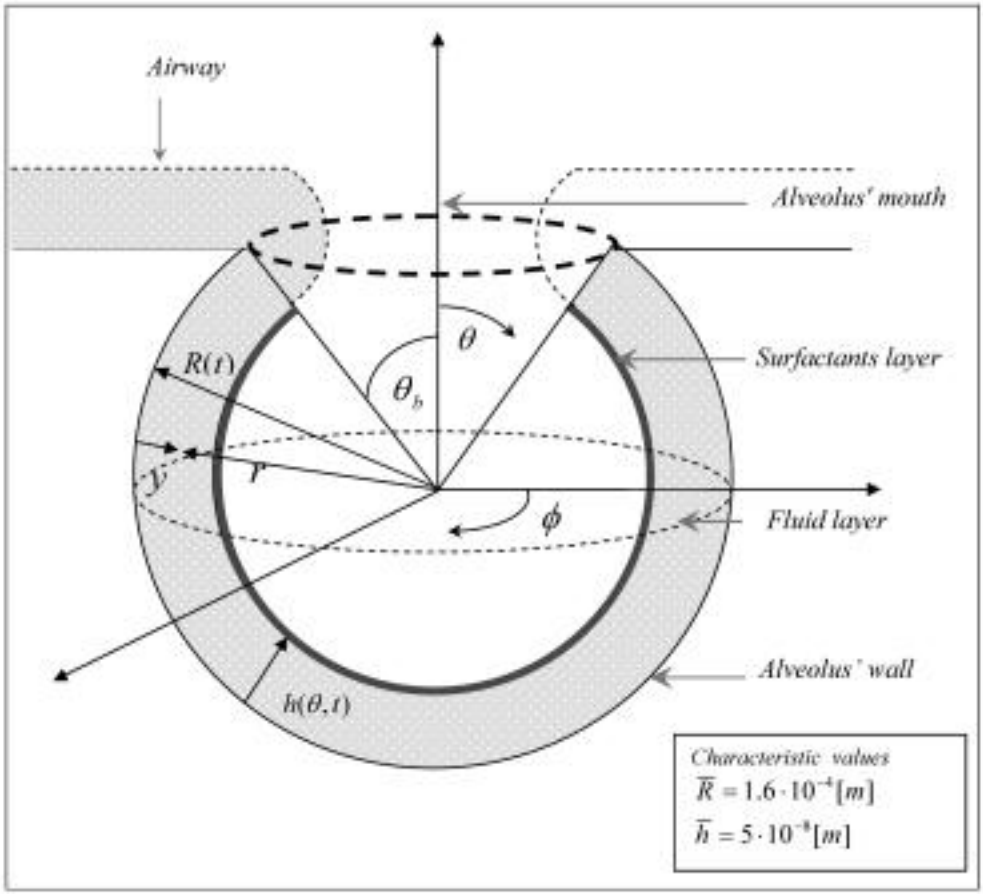


FIG. 1. *A schematic description of the alveolus.*

fluid layer is lined with a single layer of surfactant that lies at the fluid-air interface.

During expansion, additional surfactants are produced at the alveolus wall and diffuse through the fluid bulk into the fluid-air interface. Most of these retract to the fluid layer when the alveolus contracts. A residual part is cleared through the alveolar rim at $\theta = \theta_b$. Thus, a useful partition of the total rate of surfactant production $F(t)$ is

$$F(t) = \frac{\overline{m}(\lambda_b F_b(t) + \lambda_{ec} F_{ec}(t))}{T},$$

where T stands for the breathing period and $\overline{m}$ is the time-averaged amount of surfactants found in the alveolus. (Henceforth, the overhead-bar sign denotes either an average or a typical value.) The first term $\overline{m}\lambda_b F_b(t)/T$ is the rate of production of surfactants that are cleared from the alveolus rim at $\theta = \theta_b$. The prefactor $\overline{m}\lambda_b/T$ is used to scale the production rate so that the time dependent function $F_b(t)$ is of order unity.

The second term $\lambda_{ec}\overline{m}F_{ec}(t)/T$ is a periodic function with zero mean that stands for the rate of transit of surfactant between the fluid bulk and the air-fluid interface during the expansion and contraction process. The prefactor $\lambda_{ec}\overline{m}/T$ scales its

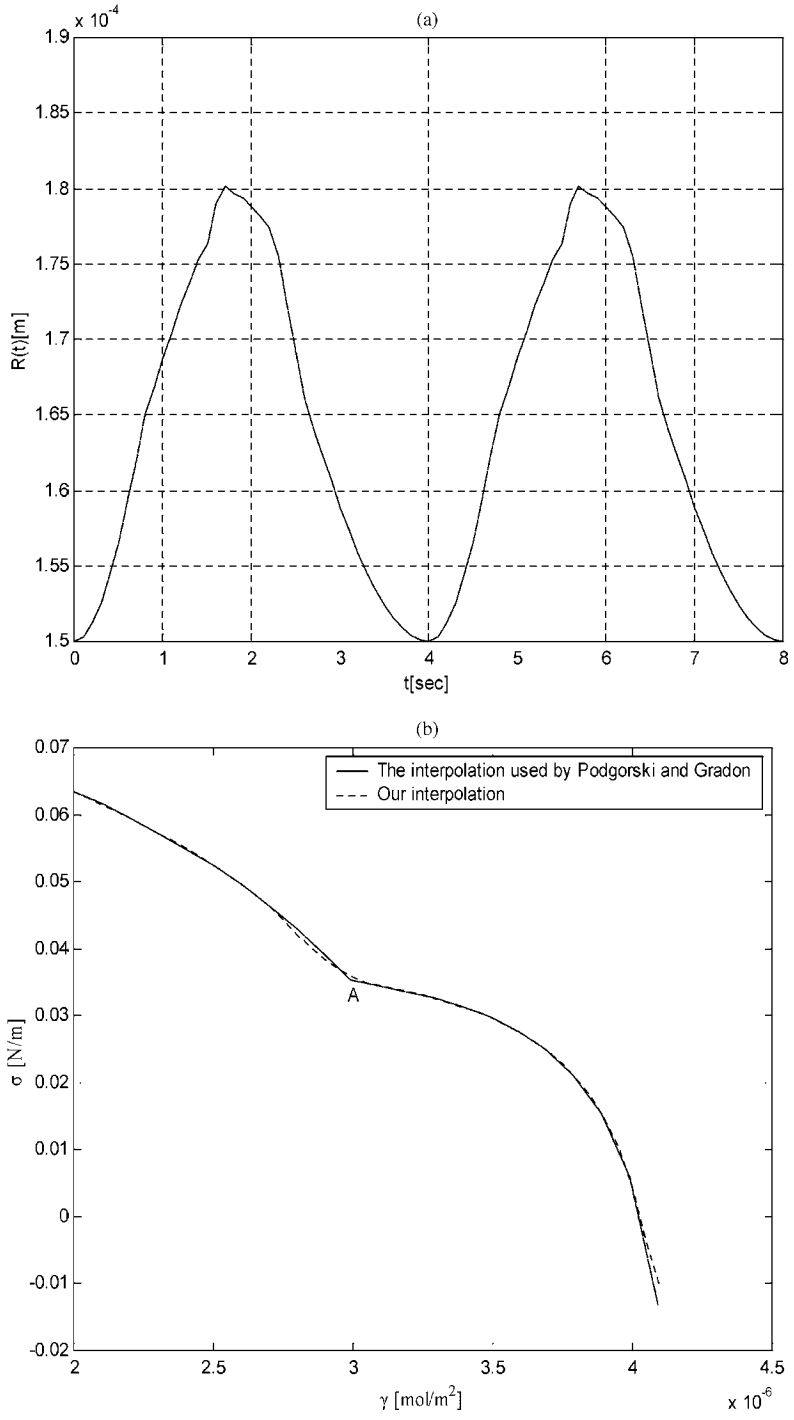


FIG. 2. (a) *Alveolar radius dependency upon time during breathing as suggested in* [22]. *The breathing cycle is* 4 *sec.* (b) *Surface tension dependency upon concentration of the DPPC surfactant. Illustrated are the Podgorski and Gradon* [22] *correlation and the cubic spline interpolation we used in our analysis. Notice that the latter possesses a continuous derivative also at point* A.

amplitude.

Hawgood [15] provided experimental data pertaining to the total amount of surfactant cleared from the rim during a breathing cycle. However, there is practically no data on how the production rate $F(t)$ varies with time and location. With no such prior knowledge, we believe that a leading order approximation can be obtained if we assume that $F(t)$ is expanded in a time Fourier series with period T and consider only the first two leading terms with $F_b(t) \equiv 1$ and $F_{ec}(t) = \sin(\frac{2\pi t}{T})$. This is equivalent to assuming that surfactants are uniformly produced at the alveolus wall and that the rate of excess surfactants leaving the rim is fixed and scales with $\overline{g} = \overline{m}\lambda_b/T$.

According to [15], the amount of surfactant secretion per hour is about 10–40% of the total amount of the surfactant present at the alveolus. If we pick $T = 4$ sec, the amount of surfactant produced per breathing cycle is about $1.1 \cdot 10^{-4}\overline{m}$ to $4.4 \cdot 10^{-4}\overline{m}$. Hawgood [15] also reported that 1–10% of the secreted amount is cleared from the alveolus. Thus, the range of λ_b is $1 \cdot 10^{-6}$–$4 \cdot 10^{-5}$, and that of λ_{ec} is $9 \cdot 10^{-5}$–$3.9 \cdot 10^{-4}$. Henceforth, we set $\lambda_b = 1 \cdot 10^{-5}$ and $\lambda_{ec} = 1.9 \cdot 10^{-4}$ as appropriate scaling values.

The surfactant surface concentration γ scales with $\overline{\gamma} = \overline{m}/2\pi(\overline{R} - \overline{h})^2 d$, where $d = 1 + \cos(\theta_b)$ and thus $\overline{g} = 2\pi\overline{R}^2 d\lambda_b\overline{\gamma}/T$. The velocity of the lining fluid and the diffusion at the interface layer govern the surfactant flux through the rim. Thus, the amount of surfactant leaving the alveolus per unit time is $g = 2\pi(R-h)\sin(\theta_b)(\gamma u_\theta - D\frac{\partial}{\partial\theta}\gamma)_{r=R-h,\theta=\theta_b}$, where $D = 10^{-10}\mathrm{m}^2/\mathrm{sec}$ is the surfactant surface diffusion coefficient (provided in [22]) and u_θ is the tangential surface velocity. Hence, the surface velocity scales with $\overline{u}_\theta = \lambda_b\frac{\overline{R}}{T} \approx 10^{-9}$ m/sec since the flux due to diffusion is of a lesser effect.

Table 1 furnishes a summary of all the additional physical parameters that are employed in the analysis with mean numerical values taken from [10], [15], and [22].

3. Flow equations, boundary conditions, dimensionless parameters and controlling variables. The differential equations that govern the flow of the lining fluid layer are the following:

(a) The continuity equation for an incompressible fluid,

$$\nabla \cdot \mathbf{u} = 0. \tag{1}$$

(b) The quasi-steady linear momentum equation (neglecting body forces and the disjoint pressure),

$$\mu\nabla^2\mathbf{u} = \nabla p. \tag{2}$$

Here, the local acceleration and convection terms have been neglected since the Reynolds numbers, $R_{eT} = \overline{h}^2/Tv = 5 \times 10^{-10}$, $R_{e\theta} = \overline{u}_\theta\overline{h}/v = 4 \times 10^{-11}$, and $R_{er} = |(\overline{u}_r - \dot{\overline{R}})\overline{h}/v| = 10^{-6}$, are much smaller than unity. The disjoint pressure effect may be neglected since the time scale of an instability (Oron, Davis, and Bankoff [18]) that may cause rupture of the thin lining layer is of the order $96\pi^3\overline{h}^5\rho v\overline{\sigma}/A^2 = 100$ sec (for a Hamaker constant A of the order of 10^{-20} J and surface tension as low as $\overline{\sigma} = 1$ dyne/cm), a much slower process than the breathing cycle of 4 seconds. In addition, Wit, Gallez, and Christov [28] concluded that the cutoff wave number is independent of the Marangoni effect.

(c) The mass conservation equation for the surfactant layer is (see Aris [3, p. 86] for the Reynolds transport theorem in a two-dimensional curved space)

TABLE 1
Geometrical and phenomenological properties.

Description	**Typical Value [units]**
Lining fluid thickness	$\overline{h} = 5 \cdot 10^{-8}$ [m]
Alveolar radius	$\overline{R} = 1.6 \cdot 10^{-4}$ [m]
Breathing period	$\overline{t} = 4$ [sec]
Lining fluid outflow velocity	$\overline{u}_\theta = \dfrac{\lambda_b \overline{R}}{\overline{t}} \approx 10^{-9}$ [m/sec]
Surface tension	$\overline{\sigma} = 2.5 \cdot 10^{-2}$ [N/m]
Surfactant concentration at the fluid-air-interface	$\overline{\gamma} = 3.3 \cdot 10^{-6}$ [mol/m^2]
Capillary pressure	$\overline{P} = \dfrac{\overline{\sigma}}{\overline{R}} \approx 156$ [N/m^2]
Amount of surfactants in the lining fluid interface	$\overline{m} = 2\pi \overline{R}^2 d\overline{\gamma} \approx 5 \cdot 10^{-13} d$ [mol]
Ratio of lining fluid thickness to alveolar radius	$\varepsilon = \dfrac{\overline{h}}{\overline{R}} \approx 3 \cdot 10^{-4}$
Ratio of amount of surfactant leaving the alveolus during a breathing period to $\overline{m}$	$\lambda_b = 10^{-5}$
Ratio of amount of surfactant staying in the lining fluid to $\overline{m}$	$\lambda_{ec} = 1.9 \cdot 10^{-4}$
Diffusion coefficient of the surfactants at the fluid interface	$D = 10^{-10}$ [m^2/sec]
Modified capillary number	$\overline{C}_a = \dfrac{\mu \overline{R}^2}{\overline{h}\overline{\sigma} T} = 0.0614$
Alveolus fluid viscosity	$\mu = 12 \cdot 10^{-3}$ [Pa $\cdot$ sec]

$$\mathbf{n} \cdot \frac{\partial(\gamma \mathbf{n})}{\partial t} + \mathbf{u}_s \cdot \nabla_s(\gamma \mathbf{n}) \cdot \mathbf{n} - \gamma \mathbf{n} \cdot (\nabla \mathbf{u}) \cdot \mathbf{n} = D \nabla_s^2 \gamma + \frac{F(t)}{2\pi (R-h)^2 d}, \tag{3}$$

where $\mathbf{n}$ is a unit vector perpendicular to the interface, $\nabla_s = (\mathbf{I} - \mathbf{nn}) \cdot \nabla$ is the surface gradient, $\mathbf{u}_s = (\mathbf{I} - \mathbf{nn}) \cdot \mathbf{u}$ is the surface velocity, and $\mathbf{I}$ is the idem dyadic. The second term on the right-hand side of (3) is a source term that accounts for the amount of surfactant entering the interface from the fluid bulk.

(d) The equation that governs the fluid layer interface location $h(\theta, t)$ is

$$\frac{\partial E}{\partial t} + \mathbf{u} \cdot \nabla E = 0, \tag{4}$$

where $E = r - R(t) + h(\theta, t) = 0$, $\mathbf{n} = \nabla E / |\nabla E|$, and the time derivative is taken for r and θ held fixed.

Assuming that the problem is axisymmetric, (1)–(4) constitute an appropriate set of equations for the five unknown fields p, γ and h, u_r, and u_θ. The latter fields are subject to the following boundary conditions:

$$u_r = \dot{R} - \frac{\lambda_b \overline{R} \hat{U}(t)}{T}, \quad u_\theta = 0, \qquad r = R(t), \tag{5a,b}$$

$$\text{(6a,b)} \qquad \frac{\partial u_r}{\partial \theta} = 0, \quad u_\theta = 0, \qquad\qquad \theta = \pi,$$

$$\text{(7)} \qquad \mathbf{n} \cdot \boldsymbol{\tau} \cdot \mathbf{t} = \mathbf{t} \cdot \nabla \sigma, \qquad\qquad r = R - h,$$

$$\text{(8)} \qquad p - p_a - \mathbf{n} \cdot \boldsymbol{\tau} \cdot \mathbf{n} - \sigma \nabla \cdot \mathbf{n} = 0, \qquad\qquad r = R - h,$$

$$\text{(9)} \qquad \frac{\partial \gamma}{\partial \theta} = 0, \qquad\qquad \theta = \pi, \quad r = R - h,$$

$$\text{(10)} \qquad 2\pi r \sin(\theta_b) \left[\gamma u_\theta - \left(\frac{D}{r} \right) \frac{\partial \gamma}{\partial \theta} \right] = -\frac{\overline{m} \lambda_b}{T}, \qquad \theta = \theta_b, \quad r = R - h,$$

where σ stands for the surface tension at the surfactant layer, $\boldsymbol{\tau} = \mu[\nabla \mathbf{u} + (\nabla \mathbf{u})^T]$ is the viscous part of the stress tensor, and $\mathbf{t}$ stands for a unit vector tangential to the interface.

Equation (5a) accounts for the unknown velocity of lining fluid $U = \lambda_b \overline{R} \hat{U}(t)/T$ that is generated at the alveolus boundary and compensates for fluid leaving the alveolus every period. We made here the reasonable assumptions that the production rate scales with the amount of surfactants leaving the alveolus and that the fluid is generated uniformly at the alveolus wall. Equation (5b) is a manifestation of the no-slip condition imposed on the flow, (6a,b) and (9) result from the geometrical symmetry of the alveolus, and (7) represents the jump condition in the tangential component of the stress tensor due to surface tension gradients. Equation (8) considers the jump condition in the normal component of the stress tensor stemming from interface curvature, and (10) demonstrates that a given amount of surfactant leaves the alveolus during every breathing period. (That excess amount is produced at the alveolus wall and diffuses through the lining fluid toward the interface.)

To achieve closure of the problem, it seems that we need an additional boundary condition at $\theta = \theta_b$. However, for very thin fluid layers, lubrication theory applies, and such a condition is redundant. The initial conditions are

$$\text{(11)} \qquad h = \overline{h}, \qquad \gamma = \overline{\gamma},$$

where $\overline{h}$ and $\overline{\gamma}$ are constants and stand for the respective fluid layer thickness and surfactant concentration evaluated at time $t = \overline{t}$ at which the alveolar radius R assumes the value $\overline{R}$.

It shall be demonstrated that a periodic solution is readily obtained for any physical values of $\overline{h}$ and $\overline{\gamma}$. A specific set of initial conditions is required to initiate the numerical scheme but is of no consequence in the final periodic solution. For the sake of convenience, we shall assume that $\overline{t} = 0$.

Based upon experimental observations (Philips and Chapman [21]), a constitutive equation $\sigma = \sigma(\gamma)$ was suggested by Gradon and Podgorski [11], which correlates surface tension to the concentration of DPPC (diacylphosphatidylcholine). The correlation function (Figure 2(b)) includes two smooth regions and a dividing point (A) at which the function is not differentiable. The latter fact results in an aphysical, discontinuous velocity solution near the dividing point. To circumvent this difficulty, we employ a smooth, natural, cubic spline interpolation function that matches well with the Podgorski and Gradon [22] data outside A, predicts a slightly higher value near A, and is differentiable everywhere (see Figure 2(b)).

To render the differential equations and boundary conditions dimensionless, we define the following dimensionless variables (denoted henceforth with a caret symbol):

$$\hat{y} = \frac{(R-r)}{\overline{h}}, \qquad \hat{t} = \frac{t}{T}, \qquad \hat{\nabla}_s = \overline{R}\nabla_s, \qquad \hat{R} = \frac{R(\hat{t})}{\overline{R}},$$

$$\hat{u}_y = \frac{(\dot{R}-u_r)T}{\overline{h}}, \qquad \hat{\sigma} = \frac{\sigma}{\overline{\sigma}}, \qquad \hat{u}_\theta = \frac{u_\theta T}{\overline{R}}, \tag{12}$$

$$\hat{p} = \frac{(p_a-p)\overline{R}}{\overline{\sigma}}, \qquad \hat{\gamma} = \frac{\gamma}{\overline{\gamma}}, \qquad \hat{h} = \frac{h}{\overline{h}}.$$

Substituting (12) into (1)–(11) yields an equivalent set of dimensionless equations and boundary conditions, where the dimensionless unknowns depend on the independent variables $\hat{y}$, θ, $\hat{t}$ and parameters λ_b, λ_{ec}, ε, P_e, C_a, θ_b. Here $\varepsilon = \overline{h}/\overline{R} = 3 \times 10^{-4}$ is the lining fluid depth ratio, $P_e = \overline{R}^2/DT = 64$ stands for the Peclet number, and $C_a = \mu\overline{u}_\theta/\overline{\sigma} = 4.8 \times 10^{-11}$ is the capillary number. Equations (1)–(4) are highly nonlinear and couple the velocity field with surfactant concentration and the location of the interface. In the next chapter, we employ an asymptotic expansion in the two smallness parameters λ_b, ε, which makes it possible to solve the problem semianalytically.

4. The asymptotic formulation. A possible clue for a coherent asymptotic representation of the unknown functions is that the cleansing mechanism results from the generation of an excess amount of surfactant determined by λ_b, a parameter that plays a paramount role in the solution. The value of λ_b is of the order of 10^{-5}; thus gradients in surface tension driving the flow are expected to be very small, albeit not zero, resulting in a nonzero small tangential velocity. Had λ_b vanished, the lining fluid would have remained inside the alveolus at all times, covered the alveolus wall uniformly, and grown thicker during exhalation and thinner during inhalation to conserve mass. In this case, the unknown functions h, γ, σ, p, and $\mathbf{u}$ would have been radially symmetric, i.e., depended upon t but not upon θ. Consequently, the following regular asymptotic expansions in λ_b and ε are suggested:

$$\hat{u}_y = \overset{0}{u}_y\,(\hat{y},\hat{t};\varepsilon) + \lambda_b[\hat{U}_y(\hat{y},\theta,\hat{t}) + \varepsilon\hat{U}_y^{(1)}(\hat{y},\theta,\hat{t}) + \cdots] + O(\lambda_b^2), \tag{13a}$$

$$\hat{u}_\theta = \lambda_b[\hat{U}_\theta(\hat{y},\theta,\hat{t}) + \varepsilon\hat{U}_\theta^{(1)}(\hat{y},\theta,\hat{t}) + \cdots] + O(\lambda_b^2), \tag{13b}$$

$$\hat{h} = \overset{0}{h}\,(\hat{t};\varepsilon) + \lambda_b[\hat{H}(\theta,\hat{t}) + \varepsilon\hat{H}^{(1)}(\theta,\hat{t}) + \cdots] + O(\lambda_b^2), \tag{13c}$$

$$\hat{\gamma} = \overset{0}{\gamma}\,(\hat{t};\varepsilon) + \lambda_b[\hat{\Gamma}(\theta,\hat{t}) + \varepsilon\hat{\Gamma}^{(1)}(\theta,\hat{t}) + \cdots] + O(\lambda_b^2), \tag{13d}$$

$$\hat{\sigma} = \overset{0}{\sigma}\,(\hat{t};\varepsilon) + \lambda_b[\hat{\Sigma}(\theta,\hat{t}) + \varepsilon\hat{\Sigma}^{(1)}(\theta,\hat{t}) + \cdots] + O(\lambda_b^2), \tag{13e}$$

$$\hat{p} = \overset{0}{p}\,(\hat{t};\varepsilon) + \lambda_b[\hat{P}(\theta,\hat{t}) + \varepsilon\hat{P}^{(1)}(\theta,\hat{t}) + \cdots] + O(\lambda_b^2). \tag{13f}$$

Here, the naught symbol denotes the radially symmetric solution, and uppercase symbols are used to denote asymptotic, first order fields in λ_b. Notice that the leading term $\overset{0}{u}_\theta$ vanishes identically in expansion (13b); i.e., a tangential velocity component stems solely from excess production of surfactant (see also section 3).

Henceforth, we shall focus our attention on the first two terms in the foregoing expansions and neglect the contribution of the third, order $O(\lambda_b\varepsilon)$, much smaller term. Substituting (13) into the dimensionless form of (1)–(11) and collecting the zero and first order terms in λ_b results in two respective sets of dimensionless differential equations and boundary conditions.

4.1. The zero order approximation. For the zero order, radially symmetric fields, the equations are as follows:
the continuity equation,

$$\frac{\partial}{\partial\hat{y}}\left[(\hat{R}-\varepsilon\hat{y})^2\left(\frac{d\hat{R}}{d\hat{t}}-\varepsilon\overset{0}{u}_y\right)\right]=0, \tag{14a}$$

the radial momentum equation,

$$\frac{1}{(\hat{R}-\varepsilon\hat{y})^2}\left\{\frac{\partial}{\partial\hat{y}}\left[(\hat{R}-\varepsilon\hat{y})^2\frac{\partial\overset{0}{u}_y}{\partial\hat{y}}\right]+2\varepsilon\left(\frac{d\hat{R}}{d\hat{t}}-\varepsilon\overset{0}{u}_y\right)\right\}=-\left(\frac{\overline{T}\overline{\sigma}}{\mu\overline{R}}\right)\frac{\partial\overset{0}{p}}{\partial\hat{y}}, \tag{14b}$$

the mass conservation equation of surfactants,

$$\frac{\partial\overset{0}{\gamma}}{\partial\hat{t}}-\overset{0}{\gamma}\left.\frac{\partial\overset{0}{u}_y}{\partial\hat{y}}\right|_{\hat{y}=\overset{0}{h}}=\frac{1}{(\hat{R}-\varepsilon\overset{0}{h})^2}\lambda_{ec}\sin(2\pi\hat{t}), \tag{14c}$$

and the kinematic condition for interface location,

$$\frac{\partial\overset{0}{h}}{\partial\hat{t}}=\left.\overset{0}{u}_y\right|_{\hat{y}=\overset{0}{h}}. \tag{14d}$$

The appropriate boundary conditions are

$$\overset{0}{u}_y=0,\qquad \hat{y}=0, \tag{15a}$$

$$\overset{0}{p}=\frac{2\overset{0}{\sigma}}{(\hat{R}-\varepsilon\overset{0}{h})},\qquad \hat{y}=\overset{0}{h}. \tag{15b}$$

The initial conditions are replaced by the requirement that the solution be periodic.

Notice that the foregoing equations are not expanded with respect to ε, since, as shall be demonstrated in the next section, an exact solution of (14) is feasible for any value of ε.

4.2. The first order approximation. Substituting (13) into (1)–(10) and collecting first order terms in λ_b yields the following set of equations and boundary conditions:
the continuity equation,

$$\hat{R}\frac{\partial}{\partial\hat{y}}\hat{U}_y+\frac{\partial}{\partial\theta}\hat{U}_\theta+\cot(\theta)\hat{U}_\theta=0, \tag{16a}$$

the momentum equation in the radial direction,

$$\frac{\partial \hat{P}}{\partial \hat{y}} = 0, \tag{16b}$$

the momentum equation in the tangential direction,

$$\frac{\partial^2 \hat{U}_\theta}{\partial \hat{y}^2} = 0, \tag{16c}$$

the mass conservation equation of surfactants,

$$\hat{R}^2 \frac{\partial \hat{\Gamma}}{\partial \hat{t}} + 2\hat{R}\hat{\Gamma}\frac{d\hat{R}}{d\hat{t}} - \frac{\partial \hat{U}_y}{\partial \hat{y}} - \frac{1}{P_e}\left[\frac{\partial^2 \hat{\Gamma}}{\partial \theta^2} + \cot(\theta)\frac{\partial \hat{\Gamma}}{\partial \theta}\right] = 1, \tag{16d}$$

and the kinematic condition for interface location,

$$\frac{\partial \hat{H}}{\partial \hat{t}} = [\hat{U}_y]_{\hat{y}=\overset{0}{h}}. \tag{16e}$$

Notice that time derivatives in (16d,e) are carried out for y and θ held fixed. The appropriate boundary conditions are

$$\hat{U}_y = \hat{U}, \qquad \hat{y} = 0, \tag{17a}$$

$$\hat{U}_\theta = 0, \qquad \hat{y} = 0, \tag{17b}$$

$$\frac{\partial \hat{\Gamma}}{\partial \theta} = 0, \qquad \theta = \pi, \tag{17c}$$

$$\frac{\partial \hat{U}_y}{\partial \theta} = 0, \qquad \theta = \pi, \tag{17d}$$

$$\hat{U}_\theta = 0, \qquad \theta = \pi, \tag{17e}$$

$$\overline{C}_a^{-1}\frac{\partial \hat{\Sigma}}{\partial \theta} - \hat{R}\frac{\partial \hat{U}_\theta}{\partial \hat{y}} = 0, \qquad \hat{y} = \frac{1}{\hat{R}^2}, \tag{17f}$$

$$\hat{P} = 0, \qquad \hat{y} = \frac{1}{\hat{R}^2}, \tag{17g}$$

$$\hat{R}\left[\overset{0}{\gamma}\hat{U}_\theta - \frac{1}{P_e\hat{R}}\frac{\partial \hat{\Gamma}}{\partial \theta}\right] = -\cot\left(\frac{\theta_b}{2}\right), \qquad \theta = \theta_b,\ r = R - h, \tag{17h}$$

where $\overline{C}_a = \frac{\mu \overline{R}/T}{\overline{\sigma}}\frac{\overline{R}}{\overline{h}} = 0.0614$ is the modified capillary number whose inverse scales the Marangoni effect. Notice that $\overline{C}_a$ is the governing capillary number that results from the balance between the shear forces and the surface tension gradients at the interface (17f). A velocity scale defined by $\overline{R}/T$ would be improper since it governs the zero order radially symmetric fields.

In the next section, solutions for the zero and first order approximation fields are addressed.

5. The solution of the zero and first order perturbations.

5.1. The zero order, radially symmetric solution. The exact solutions for the radially symmetric fields (14a–d) are[1]

$$\overset{0}{u}_y = \frac{1}{\varepsilon}\frac{d\hat{R}}{d\hat{t}}\left(1 - \frac{\hat{R}^2}{(\hat{R}-\varepsilon\hat{y})^2}\right) = -2\frac{\hat{y}}{\hat{R}}\frac{d\hat{R}}{d\hat{t}} + O(\varepsilon), \tag{18a}$$

$$\overset{0}{p} = \frac{2\hat{\sigma}(\overset{0}{\gamma})}{(\hat{R}-\varepsilon\overset{0}{h})}, \tag{18b}$$

$$\overset{0}{h} = \frac{1}{\varepsilon}[\hat{R} - (\hat{R}^3 - 1 + (1-\varepsilon)^3)^{1/3}] = \frac{1}{\hat{R}^2} + O(\varepsilon), \tag{18c}$$

$$\overset{0}{\gamma} = \frac{(1-\varepsilon)^2}{(\hat{R}-\varepsilon\overset{0}{h})^2}\left(1 - \frac{\lambda_{ec}}{2\pi}\cos\left[2\pi\left(\hat{t} - \frac{\bar{t}}{T}\right)\right]\right) = \frac{1}{\hat{R}^2} + O(\lambda_{ec}) + O(\varepsilon). \tag{18d}$$

Notice that the radially symmetric pressure is uniform across the lining layer and that for small values of ε the leading terms of the radially symmetric solutions are of order unity.

5.2. The first order perturbation in $\boldsymbol{\lambda_b}$. The solution of (16a–e)–(17a–h) is divided into two consecutive steps. First, an analytic expression is obtained for the velocity $\hat{U}_\theta$, which is substituted into (16d). A numerical scheme is then employed, in which a finite-element method is utilized along θ and a finite difference predictor-corrector method is employed along t to solve the transformed equation (16d).

Integrating (17f) and (16c) and employing boundary condition (17b) yields

$$\hat{U}_\theta = \hat{W}(\theta,\hat{t})\hat{y}. \tag{19a}$$

From (16d), the unknown function $\hat{W}(\theta,\hat{t})$ can easily be determined in terms of $\hat{\Sigma}$ or $\hat{\Gamma}$:

$$\hat{W}(\theta,t) = \frac{\overline{C}_a^{-1}}{\hat{R}}\frac{\partial\hat{\Sigma}}{\partial\theta} = \frac{\overline{C}_a^{-1}}{\hat{R}}\frac{\partial\hat{\sigma}}{\partial\hat{\gamma}}\bigg|_{\hat{\gamma}=1/\hat{R}^2}\frac{\partial\hat{\Gamma}}{\partial\theta}. \tag{19b}$$

The latter equality stems from the known constitutive relation between surface tension and surfactant concentration.

Introducing (16a) into (16d) and employing (19a) and (19b) yields the second order partial differential equation in $\hat{\Gamma}$,

$$\frac{\partial(\hat{R}^2\hat{\Gamma})}{\partial\hat{t}} + \left[\frac{\overline{C}_a^{-1}}{\hat{R}^4}\frac{\partial\hat{\sigma}}{\partial\hat{\gamma}}\bigg|_{\hat{\gamma}=1/\hat{R}^2} - \frac{1}{P_e}\right]\left[\frac{\partial^2\hat{\Gamma}}{\partial\theta^2} + \cot(\theta)\frac{\partial\hat{\Gamma}}{\partial\theta}\right] = 1, \tag{20a}$$

subject to the boundary conditions

$$\frac{\partial\hat{\Gamma}}{\partial\theta} = 0, \qquad \theta = \pi, \tag{20b}$$

[1] An easy route to obtaining the exact solutions is to consider the problem from a global point of view in which the total fluid and surfactant mass during breathing is conserved.

$$\left(\frac{\overline{C}_a^{-1}}{\hat{R}^4}\frac{\partial\hat{\sigma}}{\partial\hat{\gamma}}\bigg|_{\hat{\gamma}=1/\hat{R}^2} - \frac{1}{P_e}\right)\frac{\partial\hat{\Gamma}}{\partial\theta} = -\cot\left(\frac{\theta_b}{2}\right), \qquad \theta=\theta_b, \tag{20c}$$

and the initial condition

$$\hat{\Gamma} = 0. \tag{20d}$$

Notice that since $\partial\hat{\sigma}/\partial\hat{\gamma}$ is invariably negative, (20a) possesses the form of a diffusion equation with an effective time dependent diffusion coefficient that is always positive.

To simplify the finite-element formulation of the problem, we rewrite (20a) and (20c):

$$A(\hat{t})\frac{\partial\hat{\Gamma}}{\partial\hat{t}} + B(\hat{t})\hat{\Gamma} + C(\hat{t})\left(\frac{\partial^2}{\partial\theta^2}\hat{\Gamma} + \cot(\theta)\frac{\partial\hat{\Gamma}}{\partial\theta}\right) = 1, \qquad \pi<\theta<\theta_b, \tag{21a}$$

$$\frac{\partial\hat{\Gamma}}{\partial\theta} = G(\hat{t}), \qquad \theta=\theta_b, \tag{21b}$$

where

$$\begin{aligned} A(\hat{t}) &= \hat{R}^2, \\ B(\hat{t}) &= 2\hat{R}\frac{d\hat{R}}{d\hat{t}}, \\ C(\hat{t}) &= \frac{\overline{C}_a^{-1}}{\hat{R}^4}\frac{\partial\hat{\sigma}}{\partial\hat{\gamma}}\bigg|_{\hat{\gamma}=1/\hat{R}^2} - \frac{1}{P_e}, \\ G(\hat{t}) &= -\left(\frac{\overline{C}_a^{-1}}{\hat{R}^4}\frac{\partial\hat{\sigma}}{\partial\hat{\gamma}}\bigg|_{\hat{\gamma}=1/\hat{R}^2} - \frac{1}{P_e}\right)^{-1}\cot\left(\frac{\theta_b}{2}\right). \end{aligned} \tag{21c}$$

The equation governing the deviation of the interface from its spherical shape $\hat{H}$ is obtained from (16a,e) and (19a,b),

$$\frac{\partial\hat{H}}{\partial\hat{t}} = \frac{\hat{U}}{\hat{R}} - \frac{\overline{C}_a^{-1}}{2\hat{R}^6}\frac{\partial\hat{\sigma}}{\partial\hat{\gamma}}\bigg|_{\hat{\gamma}=1/\hat{R}^2}\left(\frac{\partial^2\hat{\Gamma}}{\partial\theta^2} + \cot(\theta)\frac{\partial\hat{\Gamma}}{\partial\theta}\right). \tag{22}$$

Little is known about the spatial distribution and the time evolution of $\hat{U}$. A global mass-conservation requires that the amount of fluid generated at the alveolus wall equal the amount exiting the alveolus during a single breathing period. Consequently,

$$\int_0^T\left(2\pi R\sin(\theta_b)\int_0^h u_\theta dy\right)dt = \lambda_b\frac{\overline{R}}{T}\int 2\pi R^2[1+\cos(\theta_b)]\hat{U}dt. \tag{23}$$

Substituting (19a,b) into (23) yields

$$\int_0^1 \hat{R}^2\hat{U}d\hat{t} = \frac{1}{2}\varepsilon\overline{C}_a^{-1}\tan\left(\frac{\theta_b}{2}\right)\int_0^1\frac{1}{\hat{R}^4}\frac{\partial\hat{\sigma}}{\partial\hat{\gamma}}\bigg|_{\hat{\gamma}=1/\hat{R}^2}\left[\frac{\partial\hat{\Gamma}}{\partial\theta}\right]_{\theta=\theta_b}dt. \tag{24}$$

Hence $\hat{U}$ is of order ε, and the second term in (22) determines the time evolution and spatial distribution of $\hat{H}$. Fortunately, the equations for $\hat{\Gamma}$ and $\hat{H}$ are decoupled, and we focus on solving $\hat{\Gamma}$, which makes it possible to predict the tangential velocity.

A weak form of the equation for $\hat{\Gamma}$ is obtained by integrating (21a) over the solution domain

$$\int_{\theta=\theta_b}^{\pi} w\left[A(\hat{t})\frac{\partial\hat{\Gamma}}{\partial\hat{t}}+B(\hat{t})\hat{\Gamma}+C(\hat{t})\left(\frac{\partial^2\hat{\Gamma}}{\partial\theta^2}+\cot(\theta)\frac{\partial\hat{\Gamma}}{\partial\theta}\right)-1\right]d\theta=0, \tag{25}$$

where w is any differentiable weighting function. Integrating (25) by parts and utilizing boundary conditions (20b) and (21b) yields

$$\begin{aligned}&\int_{\theta=\theta_b}^{\pi}\left\{w\left[A(\hat{t})\frac{\partial\hat{\Gamma}}{\partial\hat{t}}+B(\hat{t})\hat{\Gamma}+C(\hat{t})\cot(\theta)\frac{\partial\hat{\Gamma}}{\partial\theta}\right]-C(t)\frac{\partial w}{\partial\theta}\frac{\partial\hat{\Gamma}}{\partial\theta}\right\}d\theta\\&=\int_{\theta=\theta_b}^{\pi}wd\theta+w(\theta_b)C(\hat{t})G(\hat{t}).\end{aligned} \tag{26}$$

An element mesh is formed over the solution domain, and w and $\hat{\Gamma}$ are expanded in the following Galerkin sums (see, for example, [17]) for arbitrary c_A's:

$$\begin{aligned}w&=\sum_{A\in\Omega}c_AN_A(\theta),\\\hat{\Gamma}&=\sum_{B\in\Omega}d_B(\hat{t})N_B(\theta),\end{aligned} \tag{27}$$

where Ω denotes the nodes index group and N_A and N_B are the shape functions, $N_1(\theta)$ being the shape function of an element located at the alveolus opening. The unknown time dependent functions $d_B(\hat{t})$ are to be determined as follows. Substituting (27) into (26) yields

$$\begin{aligned}&\sum_{B\in\Omega}\frac{d}{d\hat{t}}d_B\int_{\theta=\theta_b}^{\pi}N_AA(\hat{t})N_Bd\theta\\&\quad+\sum_{B\in\Omega}d_B\int_{\theta=\theta_b}^{\pi}\left\{\begin{array}{l}N_AB(\hat{t})N_B+N_AC(\hat{t})\cot(\theta)\dfrac{dN_B}{d\theta}\\[2ex]-C(\hat{t})\dfrac{dN_A}{d\theta}\dfrac{dN_B}{d\theta}\end{array}\right\}d\theta\\&\quad=\int_{\theta=\theta_b}^{\pi}N_Ad\theta+N_1(\theta_b)C(\hat{t})G(\hat{t}).\end{aligned} \tag{28}$$

Thus (28) possesses the form

$$M\frac{d}{d\hat{t}}d(\hat{t})+Kd(\hat{t})=V, \tag{29}$$

where d is a vector consisting of the unknown functions $d_A(A\in\Omega)$, M and K are

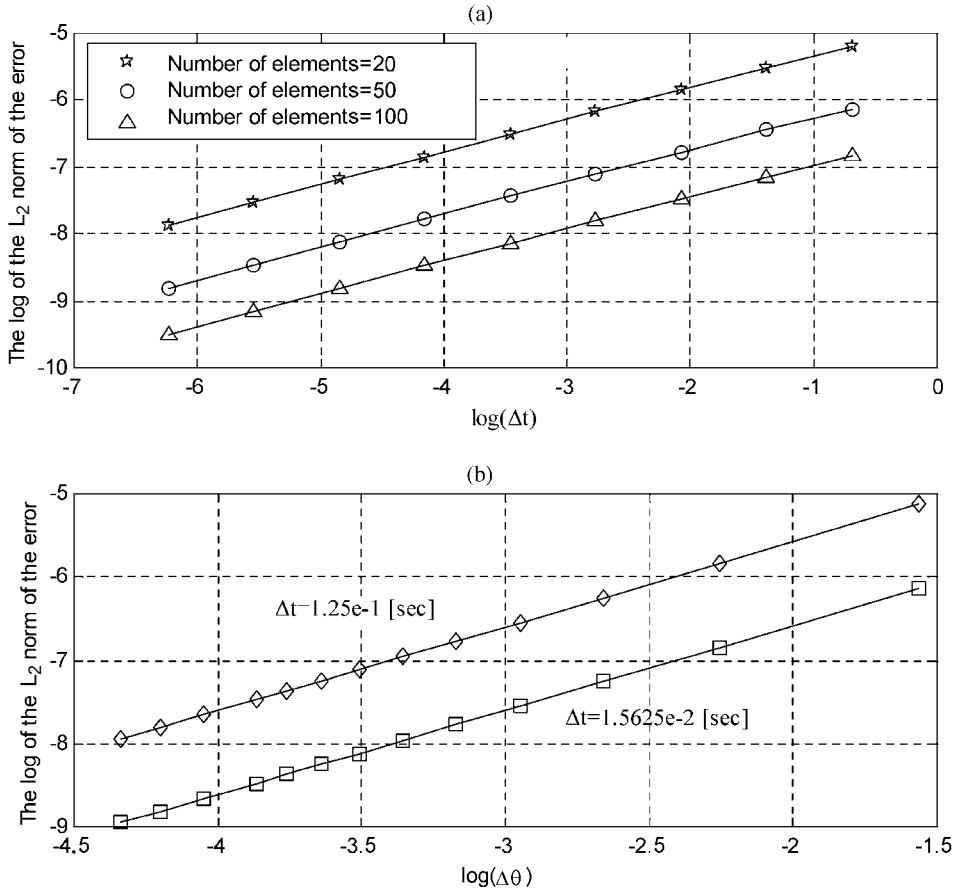


FIG. 3. *Error evaluation for various* (a) *mesh-sizes,* (b) *time-steps. The error is defined by the equation* $L\Gamma_{\text{calculated}} = \text{error}$*, where* L *is the differential operator defined in* (21a).

coefficient matrices, and V is a vector defined as follows:

$$
\begin{aligned}
M_{AB} &= \int_{\theta=\theta_b}^{\pi} N_A A(\hat{t}) N_B d\theta, \\
K_{AB} &= \int_{\theta=\theta_b}^{\pi} \left\{ N_A B(\hat{t}) N_B + N_A C(\hat{t}) \cot(\theta) \frac{dN_B}{d\theta} - C(t) \frac{dN_A}{d\theta} \frac{dN_B}{d\theta} \right\} d\theta, \\
V_A &= \int_{\theta=\theta_b}^{\pi} N_A d\theta + N_1(\theta_b) C(\hat{t}) G(\hat{t}).
\end{aligned} \tag{30}
$$

Choosing linear shape functions N_A, the matrices $M = [M_{AB}]$, $K = [K_{AB}]$ and the vector $V = [V_A]$ can be numerically calculated.

The time evolution equation (29) is numerically solved by a predictor-corrector code. Convergence and error properties of the numerical scheme, the time evolution of the surfactant distribution, the tangential velocities, and the effect of varying the phenomenological parameters are all addressed in the next section.

6. Results. We examined the convergence and accuracy of the numerical scheme; the results are illustrated in Figures 3–5. An L_2 norm was utilized to evaluate errors

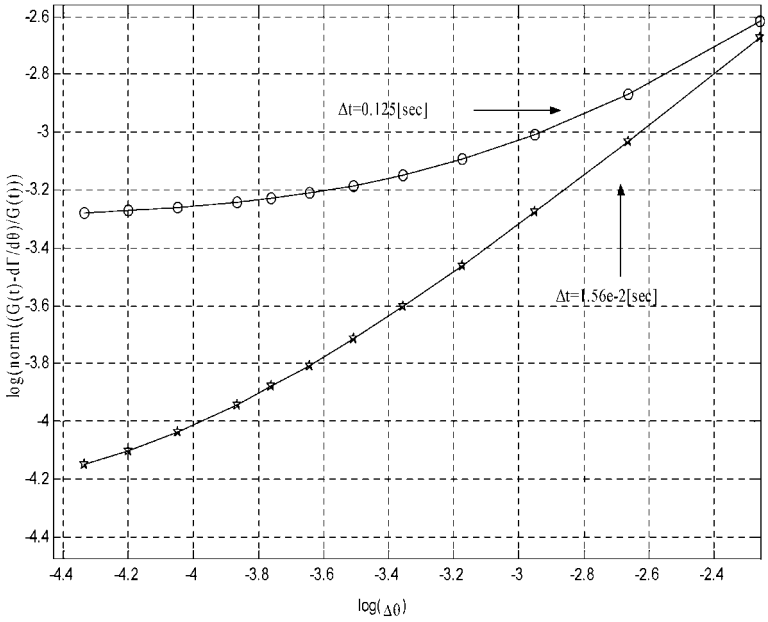


FIG. 4. *The relative error between the calculated derivative* $\partial\Gamma/\partial\theta$ *at* θ_b *and its known exact value from boundary conditions* (21b,c).

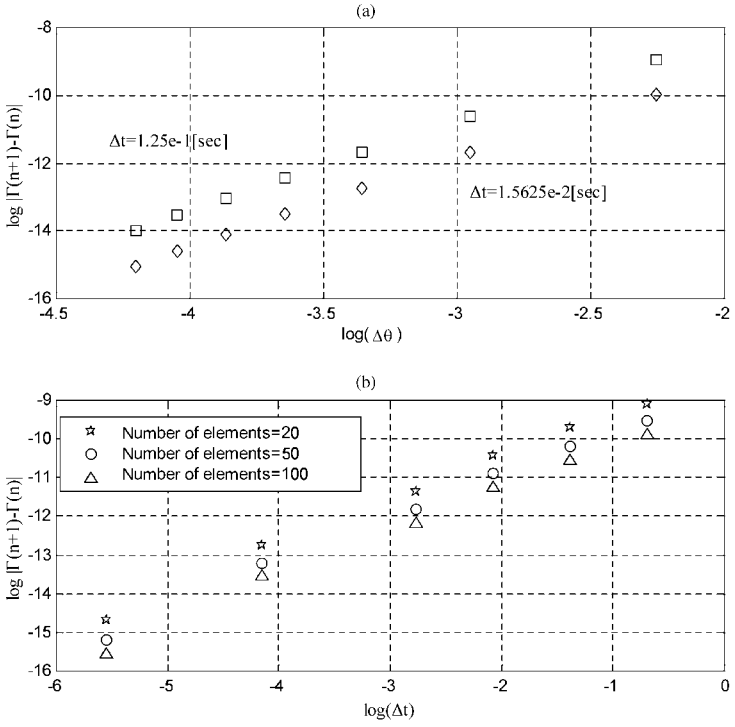


FIG. 5. *Solution convergence for various* (a) *mesh sizes,* (b) *time-steps. Here* $\Gamma(n)$ *is an* L_2 *norm of* Γ *in the solution domain, and* n *defines refinement order.*

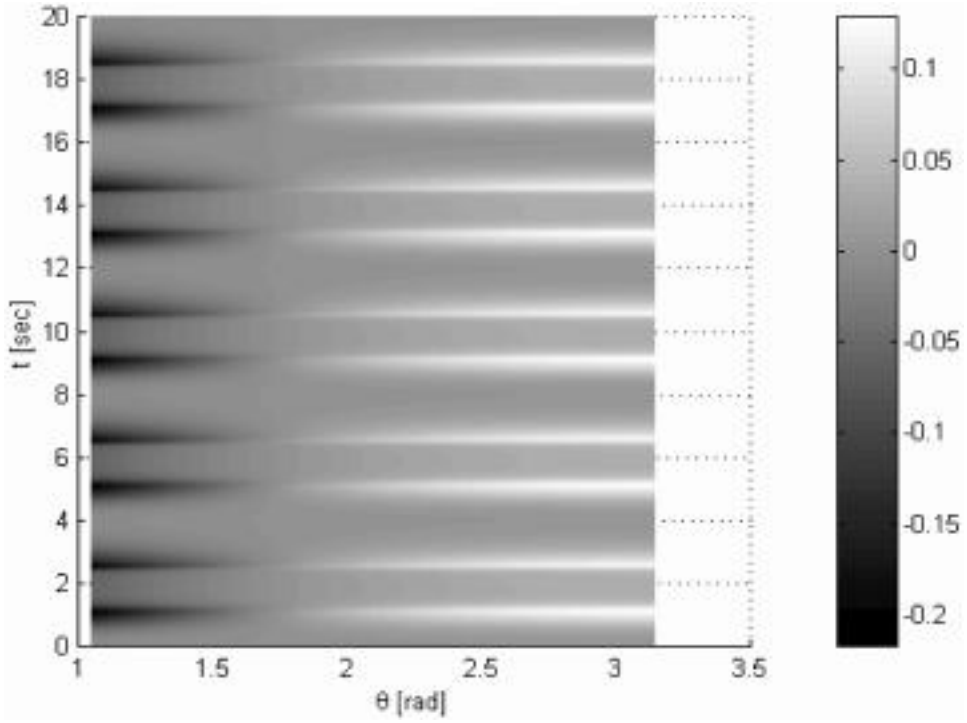


FIG. 6. *Surfactant concentration* Γ *as a function of position and time during five breathing periods.*

in the solution for $\hat{\Gamma}$. To that end, we used the parameter values defined in Table 1; the solution domain was defined by $\pi/3 \leq \theta \leq \pi$; and comparisons were made for the time interval $16 \leq t \leq 20$ sec. The chosen time span, the fifth breathing cycle, was picked to avoid transient effects that may exist at earlier times and are affected by the particular choice of initial conditions. The calculations were repeated for refined time-steps and elements in the θ-direction. We tested grids having 10 to 160 elements in the θ-direction, and time-step sizes ranging between 0.5 sec and $1.95 \cdot 10^{-3}$ sec. The results are summarized in Figure 3, which demonstrates that the estimated error decreases for both time and grid refinements. The best error estimate can be achieved at the boundaries, where a comparison can easily be made between known exact values of the derivatives of $\hat{\Gamma}$ and the respective numerical predictions (see Figure 4). The figure makes clear that the error decreases monotonically with reduced values of time-steps and increased number of elements.

To evaluate the convergence rate of the solution, an L_2 norm was also calculated for the difference between consecutive refined solutions (see Figure 5(a,b)). The figures illustrate vividly that convergence is achieved even for high values of time-steps (of order 0.1) and a small number of elements (of order 20).

Since the solution is approximated up to order ε, no greater precision than 10^{-4} is required. Consequently, from Figures 3–5, a time-step size of 0.015 sec was selected, and the θ-domain was divided into 100 elements, a parameter set that yields a converging solution with an estimated absolute error of order ε or less.

The time evolution of the surfactants and velocity fields is illustrated in Figures 6–10. Since $R(t)$ is a periodic function and, consequently, the time dependent coefficients

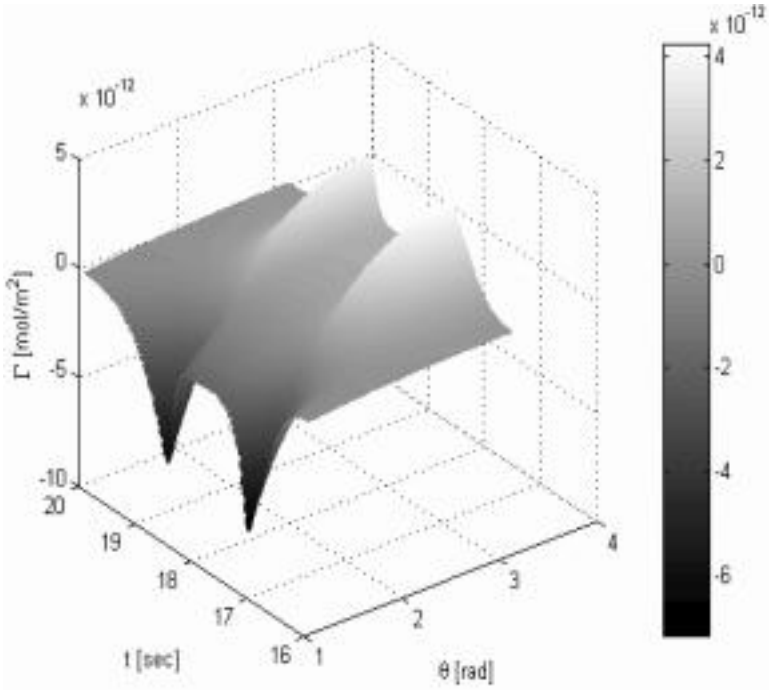


FIG. 7. *Surfactant concentration* Γ *as a function of position and time during the fifth breathing period.*

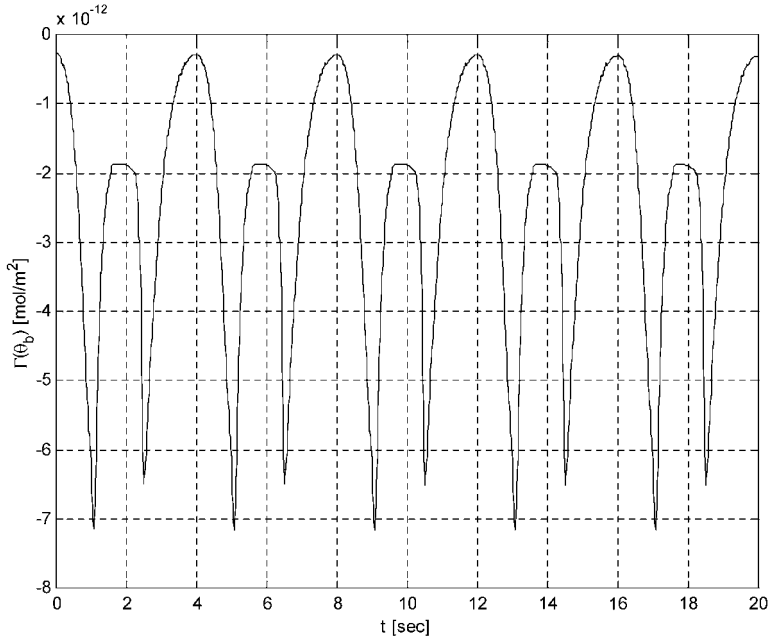


FIG. 8. *The evolution of surfactant concentration* Γ *at* θ_b *during five breathing periods. Notice the two-peak pattern occurring within every breathing period.*

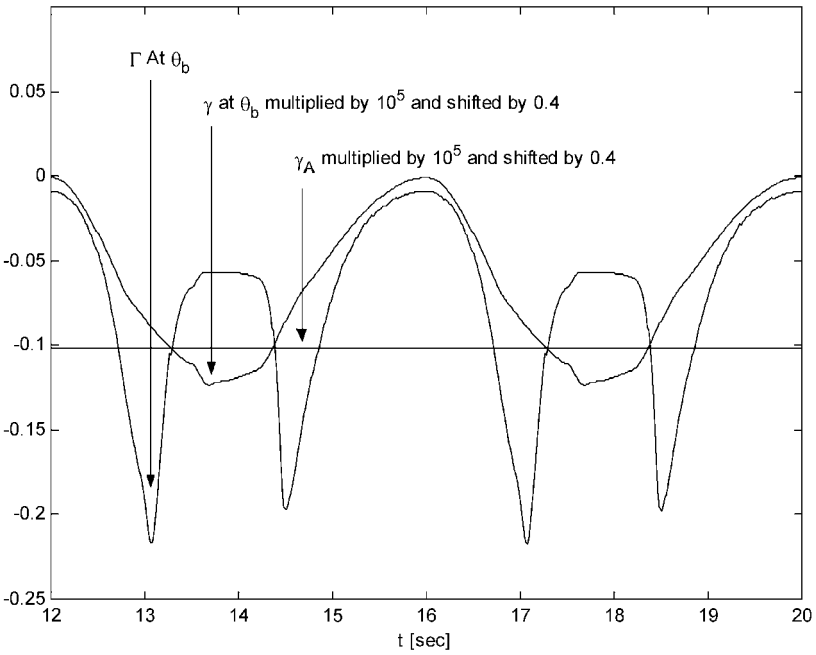


FIG. 9. *The temporal evolution of the radial surfactant concentration* $\overset{0}{\gamma}$ *and* Γ. *Notice that* Γ *reaches a maximum value when* $\gamma = \gamma_A$.

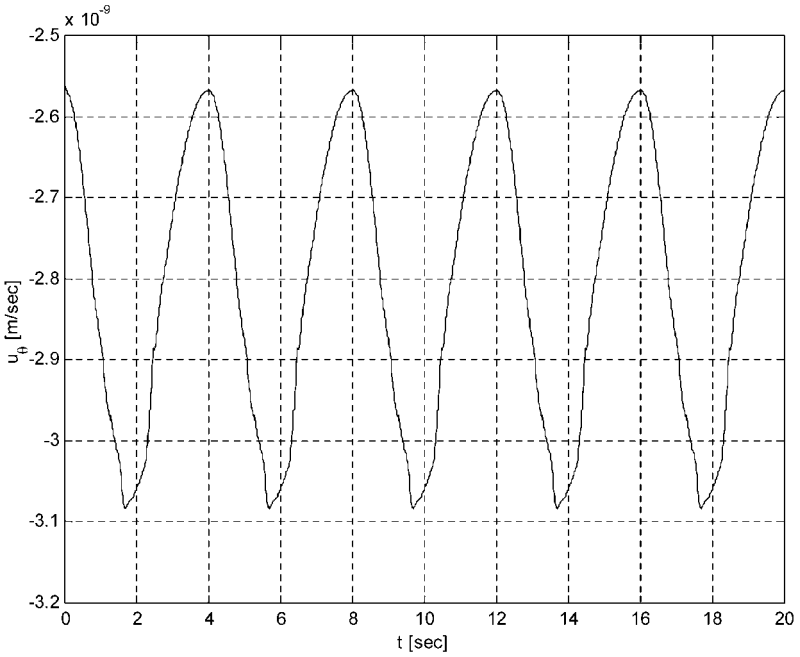


FIG. 10. *The temporal evolution of the tangential velocity* U_θ *at* $\theta_b = \pi/3$.

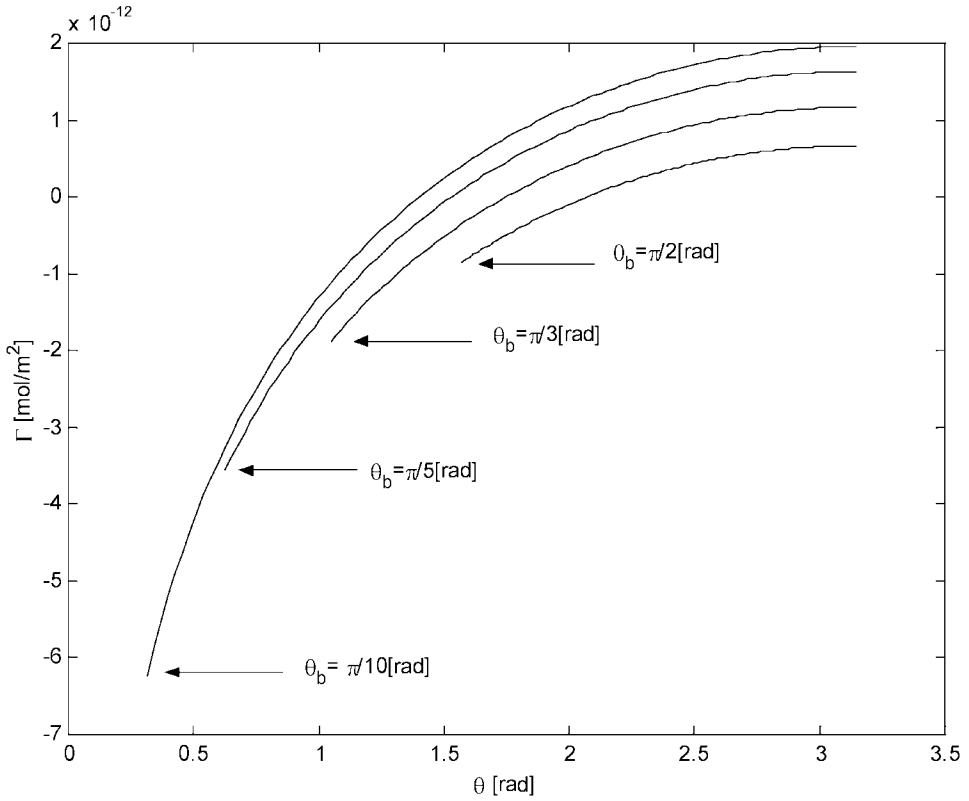


FIG. 11. *The spatial distribution of surfactant concentration* Γ *for various values of* θ_b.

of (20a–d) are periodic, we expect a periodic steady-state solution to the problem. Indeed, Figure 6 demonstrates that $\hat{\Gamma}$ reaches a steady state after a short transient period of less than a single breathing cycle.

During every breathing period, $\hat{\Gamma}$ (at θ_b) possesses a two-peak pattern (see Figures 7 and 8). We use, henceforth, the dimensional form of $\hat{\Gamma}$, namely, $\Gamma = \lambda_b \overline{\gamma} \hat{\Gamma}$, to describe the small perturbation in surfactant concentration. The peaks occur during inhalation and exhalation when the derivative of the surface tension with respect to surfactant concentration varies abruptly as γ crosses point A in Figure 2(b) (see Figure 9). The value of Γ remains negative throughout the breathing process. Thus, the total surfactant concentration $\gamma = \overline{\gamma} \overset{0}{\gamma} + \Gamma$ is lower than its radially symmetric concentration $\overline{\gamma} \overset{0}{\gamma}$. This is reflected in a higher than average surface tension at θ_b and a net fluid motion toward the alveolar edge. The latter conclusion is also illustrated in Figure 10, in which the time dependence of the tangential velocity at the interface is depicted. Notice that a negative value for u_θ means a flow direction toward the alveolar edge (Figure 1). It demonstrates that the velocity is a time-periodic function that possesses a negative mean; namely, there is a net flow exiting the alveolus.

Figures 11 and 12 illustrate a smooth spatial distribution of Γ and u_θ for various values of θ_b. Figure 11 validates the former conclusion that surfactant concentration is lowest (surface tension is highest) at θ_b, namely, fluid is drawn toward the alveolar edge. Figure 12 illustrates that the tangential velocity increases (in absolute value) as

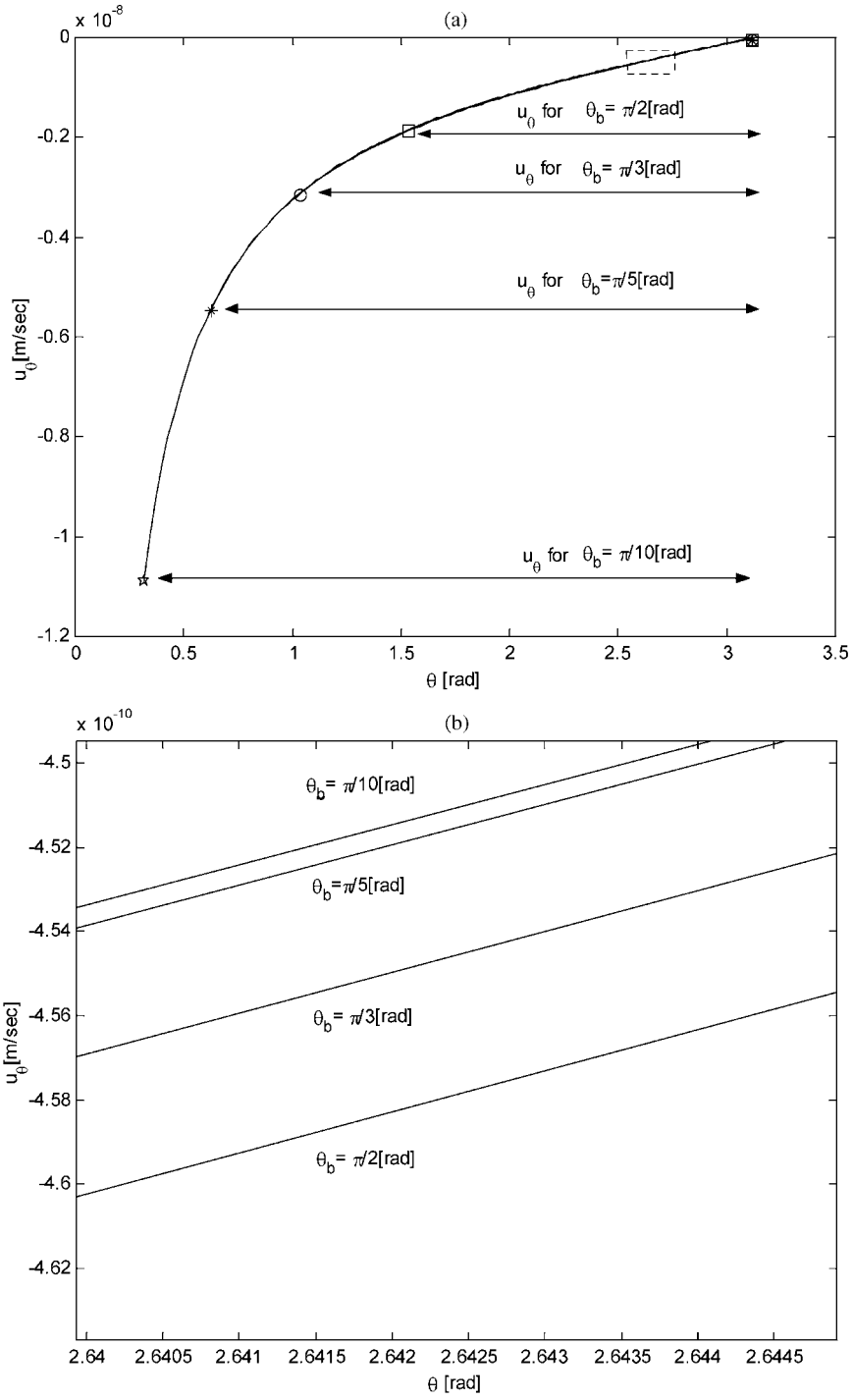


FIG. 12. (a) *The spatial distribution of the tangential velocity for various values of* θ_b. (b) *A blowup of the dotted small rectangle shown in* (a) *that manifests the small contribution of the angle* θ_b.

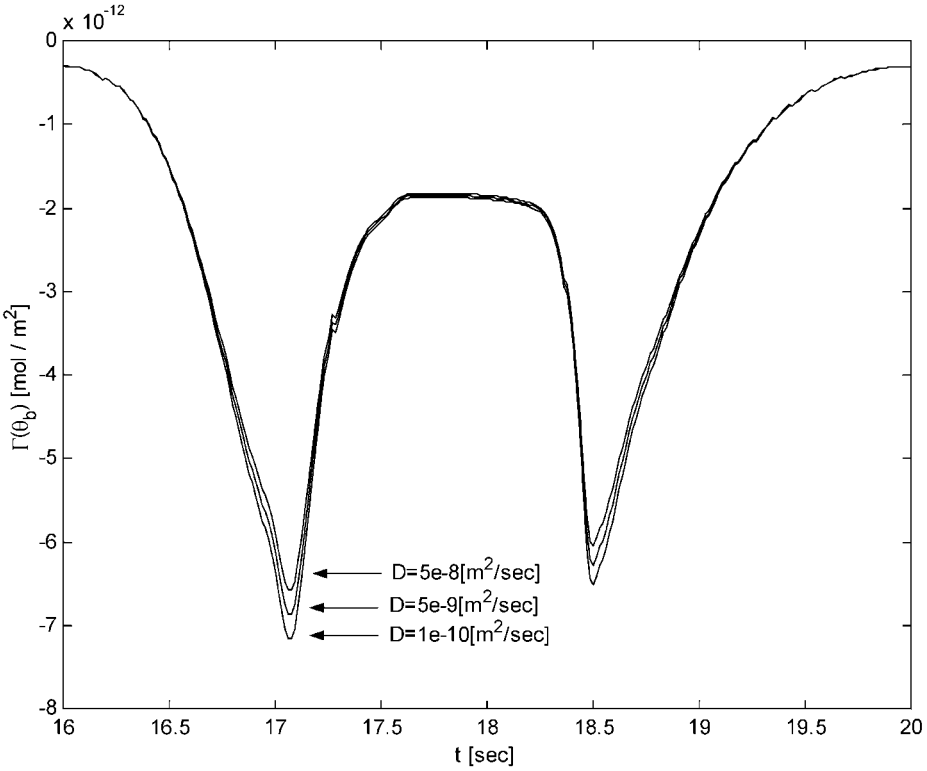


FIG. 13. *The temporal evolution of surfactant concentration* Γ *at* θ_b *for various values of the diffusion coefficient* D.

θ gets closer to θ_b. This result is consistent with the assumption that the fluid excess is generated uniformly at the alveolar surface.

The surfactant concentration and the tangential velocity dependence upon θ and θ_b are also illustrated in Figures 11 and 12, respectively. Figure 11 illustrates that $\gamma = \overline{\gamma} \overset{0}{\gamma} + \Gamma$ decreases and the surfactant concentration gradient increases as θ_b decreases. Hence, smaller values of θ_b yield a nonlinear increase in the magnitude of u_θ at the alveolar rim. This result is not surprising since, from (17h), if the Peclet number is large, u_θ varies like $\cot(\theta_b/2)$. Figure 12 illustrates how u_θ increases (in absolute value) as we approach the alveolar rim. It also illustrates that different values of θ_b result in almost identical values of u_θ, namely, all lines seem to collapse into a single graph within their mutual domain. However, a blowout of a small domain (shown by a small rectangle in the upper right corner of Figure 12) indicates that small deviations do exist between different values of θ_b (Figure 12(a)), with slightly smaller values of u_θ for smaller θ_b's.

The effect of the Peclet number upon surfactant distribution and the tangential velocity field is summarized in Figures 13–14. Figure 13 illustrates a double peaked pattern that results from the abrupt change in surface tension gradients at point A of Figure 2(b). Figure 14 illustrates dependence of u_θ upon time, with the highest (absolute) value occurring at the end of inhalation and the beginning of exhalation. Varying the diffusion coefficient has a minor effect on the results. This is not surprising since the Peclet number is quite high (Pe = 64) and the inverse of the capillary

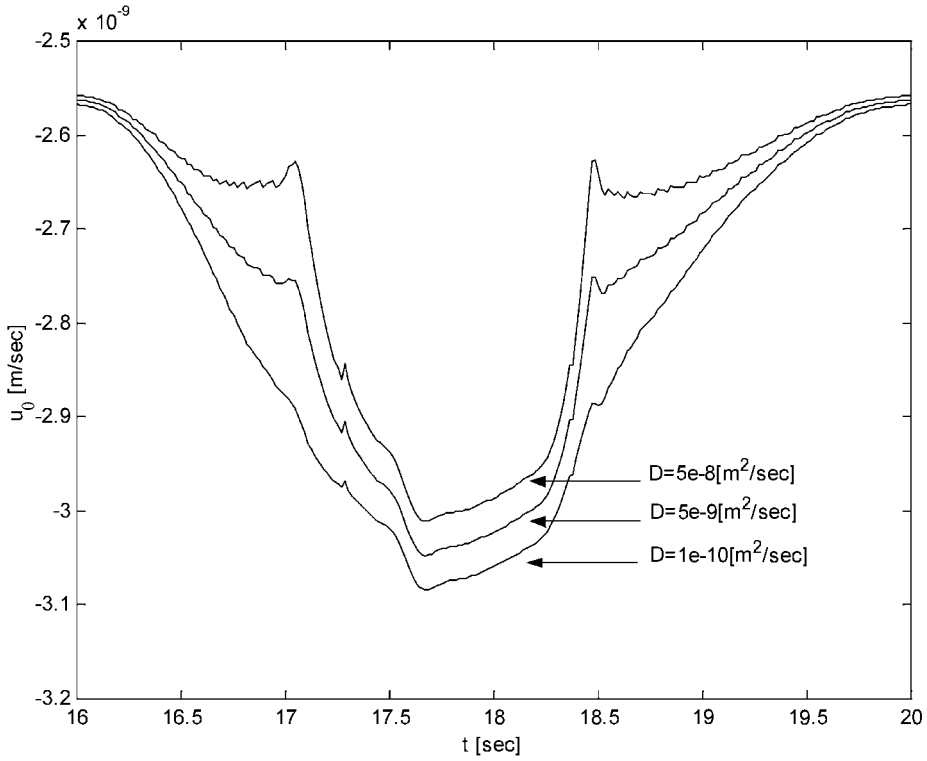


FIG. 14. *The temporal evolution of the tangential velocity U_θ at θ_b for various values of the diffusion coefficient D.*

number is about 16. A more significant effect would occur were D of the order of $D = 5 \times 10^{-8}$ m^2/sec, a much greater value than the estimated physical value of $D = 10^{-10}$ m^2/sec.

Finally, the surfactant production rate, λ_b, has a most significant effect on the tangential velocity. An increase in the production rate causes a concomitant increase in the tangential velocity.

7. Discussion and conclusions. The results in section 6 demonstrate that gradients in surfactant concentration at the lining layer interface induce tangential flow toward the alveolar edge (the Marangoni effect). Based upon experimental observations, we assumed that during every breathing cycle an excess amount of surfactant was secreted at the alveolus wall and removed to the adjacent airway. This excess amount is a given percentage of the existing average amount of surfactant that is embedded inside the lining layer. The removal of surfactants and the concomitant concentration gradients induce tangential flow inside the lining layer so that a small amount of the lining fluid exits the alveolus with a typical low rate on the order of 10^{-9} m/sec. The flow rate varies periodically with time and depends strongly upon how widely open the alveoli are. Pathologically wide cone angles θ_b result in a strong reduction in $\hat{U}_\theta$ and vice versa. However, since $u_\theta \sim \lambda_b \hat{U}_\theta$, the actual tangential velocity may either increase or decrease with θ_b. To make a rigorous conclusion, additional experimental evidence is required to correlate the flux of surfactant exiting the alveolus (proportional to λ_b) with the cone angle θ_b.

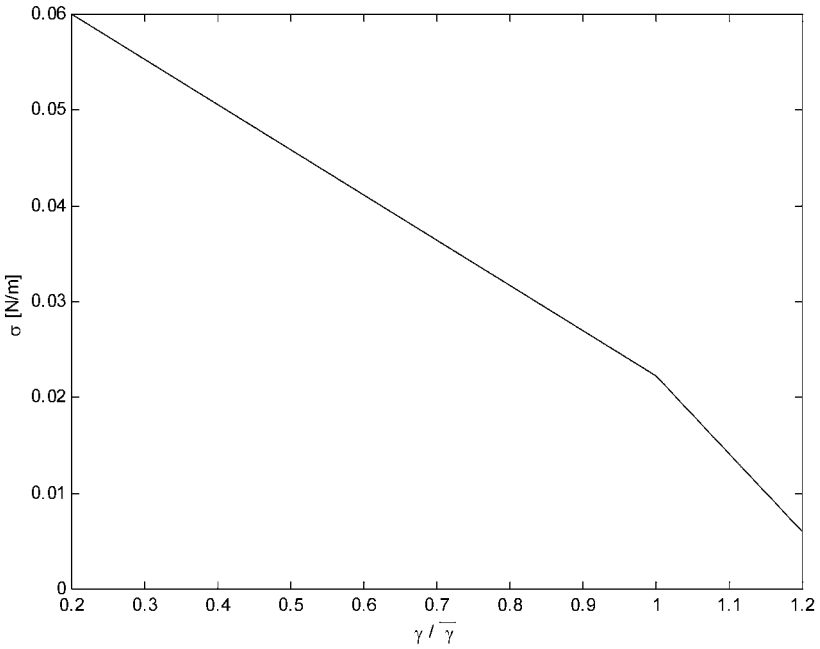


FIG. 15. *Surface tension dependency upon concentration of surfactant TA from Otis et al.* [19].

Particles that are deposited over the alveolus wall are subject to hydrodynamic drag and may be swept out of the alveolus due to the induced tangential velocity. The hydrodynamic cleansing rate is determined by particle velocity that, generally, need not be equal to the fluid velocity. However, the predicted fluid tangential velocity at the alveolar rim may provide a reasonable measure of the rate of hydrodynamic cleansing. With an average sweeping rate, it will take a particle about two days to move a distance equal to one alveolar radius, a very small rate indeed.

The effect of particle diffusion may add to the cleansing rate. However, this effect may be quite small. The diffusion coefficient of a particle 1μm in diameter in an *unbounded* lining flow field is $D_p = 3.8^{*}10^{-14}$ m^2/sec, based on the Stokes–Einstein equation. Thus, it seems that the time it takes a particle to travel a distance equal to one alveolar radius $R = 10^{-4}$m is of the order of $R^2/4D_p \sim 10^5$s, a value similar to the convection time. Notwithstanding this idea, Happel and Brenner [13] show that, due to the close proximity of the particle to the alveolar walls, the hydrodynamic drag coefficient can be several order of magnitudes higher than $6\pi\mu r_p$ (here r_p is the particle radius). Consequently, the value for the diffusion coefficient would be smaller and the resulting diffusion time longer.

We also tried to compare DPPC with an artificial surfactant TA (also known as Survanta; Ross Laboratories, Columbus, OH), widely used clinically to treat respiratory distress syndrome. From Otis et al. [19], a surfactant TA isotherm, relating the surface tension to surface concentration, is obtained (Figure 15) and approximated by two straight lines. Figures 16 and 17 illustrate the behavior of surfactant TA vis-à-vis DPPC, provided that their Peclet number is of similar order.[2] The time evolution of Γ differs markedly from that of DPPC; however, the calculated u_θ at θ_b is very similar.

[2]Note that synthetic surfactants do not undergo cellular secretion and adsorption. Thus, the results may depend on the time protocol by which TA is provided, but this is left for future work.

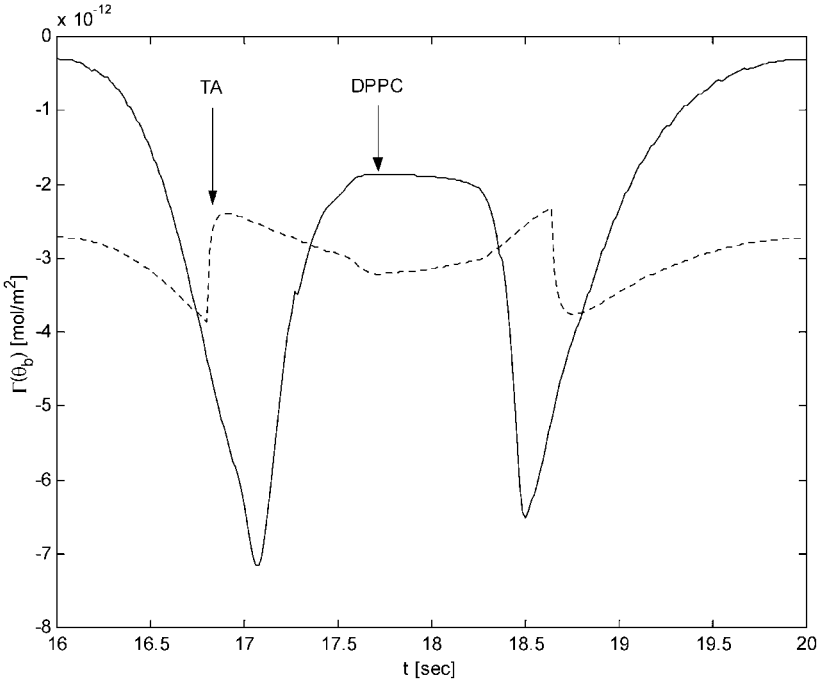


FIG. 16. *The temporal evolution of surfactant concentration* Γ *for DPPC and surfactant TA at* $\theta_b = \pi/3$.

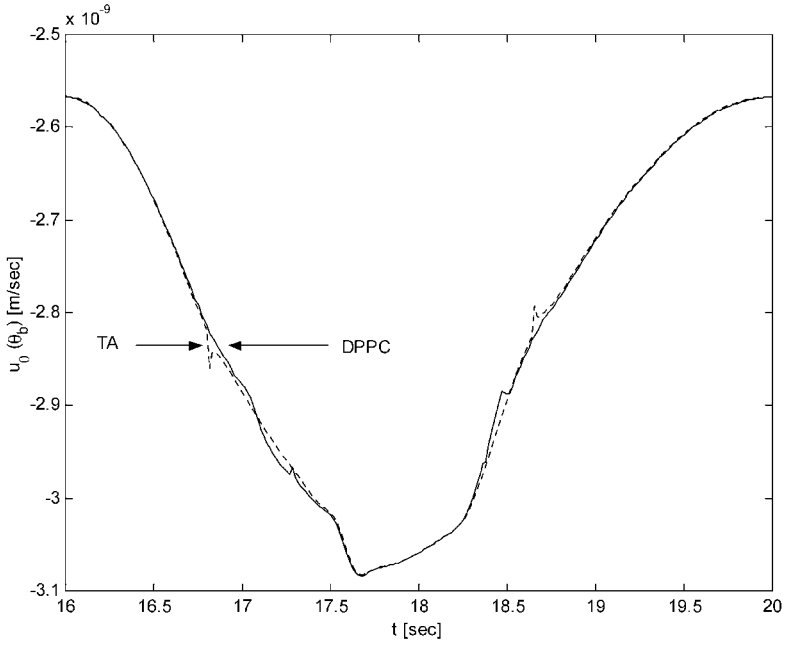


FIG. 17. *The temporal evolution of the tangential velocity for DPPC and surfactant TA at* $\theta_b = \pi/3$.

The small deviations stem from the discontinuity in $\partial\hat{\sigma}/\partial\hat{\gamma}$ assumed for surfactant TA. In fact, U_θ at θ_b should depend approximately linearly upon $\hat{R}$. Since, from (19a) and (19b), at $\hat{y} = 1/\hat{R}^2$ we obtain that

$$\hat{U}_\theta = \frac{\hat{W}}{\hat{R}^2} = \frac{\overline{C}_a^{-1}}{\hat{R}^3} \frac{\partial\hat{\sigma}}{\partial\hat{\gamma}}\bigg|_{\hat{\gamma}=1/\hat{R}^2} \frac{\partial\hat{\Gamma}}{\partial\theta},$$

consequently, at the alveolus edge $\theta = \theta_b$, for large Peclet numbers, boundary condition (21c) results in $\hat{U}_\theta|_{\theta=\theta_b} \sim -\hat{R}\cot(\theta_b/2)$. Thus, the major difference in the tangential velocity $u_\theta \sim \lambda_b \hat{U}_\theta$ between DPPC and surfactant TA stems from λ_b, provided that they possess similar diffusion coefficients.

In summary, a significant enhanced hydrodynamic cleansing can occur if the mechanism that keeps the surfactants from excessive accumulation or dilution functions over a wide range of surfactant concentrations. Notice that a very small deviation in surfactant concentration from the radially symmetric distribution is sufficient to induce flow in the lining layer. Thus, artificial stimulation of surfactant production at the alveolar wall tissue, or artificially administering a small excess amount of surfactant by inhalation, may result in an increased flow of surfactants exiting the alveoli and a concomitant sweeping flow of the lining layer. More research is required to investigate what the physiological mechanisms might be that cause surfactants to exit the alveolus and thereby determine/control the important parameter λ_b for various values of alveolus cone angle θ_b and surfactant composition. We hope that an artificial process can be devised and experimentally tested so that people exposed to a severe polluted environment could utilize the mechanism of enhanced hydrodynamic cleansing to reduce particle deposition of hazardous materials inside the lung alveoli.

Acknowledgments. The authors would like to thank Professor L. Gradon for his helpful explanations and Dr. A. Tsuda for illuminating discussions.

REFERENCES

[1] T. Ahmed, *Clinical testing of aerosol drugs*, in Respiratory Drug Delivery, P. R. Byron, ed., CRC Press, Boca Raton, FL, 1990, pp. 208–242.

[2] R. Ardila, T. Horie, and J. Hildebrandt, *Macroscopic isotropy of lung expansion*, Respir. Physiol., 20 (1974), pp. 105–115.

[3] R. Aris, *Vectors, Tensors, and the Basic Equations of Fluid Mechanics*, Prentice–Hall, Englewood Cliffs, NJ, 1962.

[4] P. Bezdicek and G. C. Crystal, *Pulmonary macrophages*, in The Lung: Scientific Foundations, 2nd ed., R. G. Crystal and J. B. West, eds., Lippincott-Raven, Philadelphia, 1997, pp. 859–874.

[5] J. Brain, J. D. Blanchard, and T. D. Sweeney, *Deposition and fate of inhaled pharmacological aerosol*, in Provocative Challenge Procedures: Background and Methodology, S. D. Spector, ed., Futura, Mount Kisco, NY, 1989, pp. 1–36.

[6] D. W. Dockery, J. D. Spengler, J. H. Ware, M. E. Fay, B. G. Ferris, and F. E. Speizer, *An association between air pollution and mortality in six U.S. cities*, N. Engl. J. Med., 329 (1993), pp. 1753–1759.

[7] J. Gil, H. Bachofen, P. Gehr, and E. R. Weibel, *Alveolar volume-surface area relation in air and saline filled lungs fixed by vascular perfusion*, J. Appl. Physiol., 45 (1979), pp. 990–1001.

[8] J. Gil and E. R. Weibel, *Morphological study of pressure-volume hysteresis in rat lungs fixed by vascular perfusion*, Respir. Physiol., 15 (1972), pp. 190–213.

[9] J. Goerke, *Pulmonary surfactant: Functions and molecular composition*, Biochim. Biophys. Acta, 1408 (1998), pp. 79–89.

[10] J. Goerke and J. A. Clements, *Alveolar surface tension and lung surfactant*, in Respiratory System, Vol. III, A. P. Fishman and S. R. Geiger, eds., Am. Physiol. Soc., Bethesda, MD, 1985, pp. 247–261.

[11] L. Gradon and A. Podgorski, *Hydrodynamical model of pulmonary clearance*, Chem. Eng. Sci., 44 (1989), pp. 741–749.

[12] S. Haber, J. P. Butler, H. Brenner, I. Emanuel, and A. Tsuda, *Shear flow over a self-similar expanding pulmonary alveolus during rhythmical breathing*, J. Fluid Mech., 405 (2000), pp. 243–268.

[13] J. Happel and H. Brenner, *Low Reynolds Number Hydrodynamics*, Prentice–Hall, Englewood Cliffs, NJ, 1965.

[14] B. G. Harvey and R. G. Crystal, *Pulmonary response to chronic inorganic dust exposure*, in The Lung: Scientific Foundations, 2nd ed., R. G. Crystal and J. B. West, eds., Lippincott-Raven, Philadelphia, 1997, pp. 2339–2352.

[15] S. Hawgood, *Surfactant: Composition, structure, and metabolism*, in The Lung: Scientific Foundations, 2nd ed., R. G. Crystal and J. B. West, eds., Lippincott-Raven, Philadelphia, 1997, pp. 557–571.

[16] J. Heyder, J. D. Blanchard, H. A. Feldman, and J. D. Brain, *Convective mixing in human respiratory tract: Estimates with aerosol boli*, J. Appl. Physiol., 64 (1988), pp. 1273–1278.

[17] T. J. R. Hughes, *The Finite Element Method, Linear, Static and Dynamic Finite Elements Analysis*, Prentice–Hall, Englewood Cliffs, NJ, 1987.

[18] A. Oron, S. H. Davis, and S. G. Bankoff, *Long-scale evolution of thin liquid films*, Rev. Modern Phys., 69 (1997), pp. 931–980.

[19] D. R. Otis, E. P. Ingenito, R. D. Kamm, and M. Johnson, *Dynamic surface tension of surfactant TA: Experiments and theory*, J. Appl. Physiol., 77 (1994), pp. 2681–2688.

[20] M. A. Sleigh, J. R. Blake, and N. Liron, *The propulsion of mucus by cilia*, Am. Rev. Respir. Dis., 137 (1988), pp. 726–741.

[21] M. C. Phillips and D. Chapman, *Monolayer characteristics of saturated* $1,2$*-diacyl phosphatidylcholines (lecitins) and phosphatidylethanolamines at the air-water interface*, Biochim. Biophys. Acta, 163 (1968), pp. 301–313.

[22] A. Podgorski and L. Gradon, *An improved mathematical model of hydrodynamical self-cleansing of pulmonary alveoli*, Ann. Occup. Hyg., 37 (1993), pp. 347–365.

[23] E. M. Scarpelli, *The Surfactant System of the Lung*, Lea & Febiger, Philadelphia, 1968.

[24] H. Schulz, P. Heilmann, A. Hillebrecht, J. Gebhart, M. Mayer, J. Piiper, and J. Heyder, *Convective and diffusive gas transport in canine intrapulmonary airways*, J. Appl. Physiol., 72 (1992), pp. 1557–1562.

[25] A. Tsuda, F. S. Henry, and J. P. Butler, *Chaotic mixing of alveolated fluid duct flow in rhythmically expanding pulmonary acinus*, J. Appl. Physiol., 79 (1995), pp. 1055–1063.

[26] A. Tsuda, Y. Otani, and J. P. Butler, *Acinar flow irreversibility caused by perturbations in reversible alveolar wall motion*, J. Appl. Physiol., 86 (1999), pp. 977–984.

[27] E. R. Weibel, *Functional morphology of lung parenchyma*, in Handbook of Physiology, The Respiratory System, Vol. 3, A. P. Fishman, ed., Am. Physiol. Soc., Bethesda, MD, 1986, pp. 89–111.

[28] A. de Wit, D. Gallez, and C. I. Christov, *Nonlinear evolution equations for thin films with insoluble surfactants*, Phys. Fluids, 6 (1994), pp. 3256–3266.

[29] T. B. Zeltner, T. D. Sweeney, W. A. Skornik, H. A. Feldman, and J. D. Brain, *Retention and clearance of* $0.9mm$ *particles inhaled by hamsters during rest or exercise*, J. Appl., Physiol., 70 (1991), pp. 1137–1156.

SIAM J. APPL. MATH.
Vol. 63, No. 1, pp. 222–240

ASYMPTOTIC SOLUTION TO AN INVERSE PROBLEM FOR A SHARED UNBUFFERED RESOURCE*

JOHN A. MORRISON† AND K. G. RAMAKRISHNAN‡

Abstract. We consider an unbuffered resource having capacity C, which is shared by several different services. Calls of each service arrive in a Poisson stream and request a fixed, integral amount of capacity, which depends on the service. An arriving call is blocked and lost if there is not enough free capacity. Otherwise, the capacity of the call is held for the duration of the call, and the holding period is generally distributed. The inverse problem of determining the traffic intensities in terms of the measured values of the carried loads for each service is investigated. It is assumed that C and the traffic intensities are commensurately large. The inverse problem is solved asymptotically in the critically loaded regime, and it involves the unique real solution of a nonlinear equation. An iterative solution of this equation is shown to lead to a contraction mapping and to monotonic and geometric convergence. A separate analysis is given for the overloaded regime, and it is shown that the result matches asymptotically with that for the critically loaded regime. Numerical results are presented for two examples.

Key words. asymptotics, carried loads, inverse problem, multiservice, traffic intensities, unbuffered resources

AMS subject classifications. 60K30, 90B12

PII. S0036139901388799

1. Introduction.

We consider an unbuffered resource having capacity C, which is shared by S different services. Calls of service s ($s = 1, 2, \ldots, S$) arrive in a Poisson stream with mean rate λ_s and request capacity d_s. An arriving call is blocked and lost if there is less than d_s free capacity. Otherwise, the call capacity d_s is held for the duration of the call, and the holding period is generally distributed with mean $1/\mu_s$ and independent of earlier arrival and holding times. The traffic intensity of calls of service s is $\nu_s = \lambda_s/\mu_s$, and the product form and the insensitivity property hold (see [4], [5], [7]); i.e., the joint stationary distribution of the number of active calls of each service depends on the distributions only through ν_s.

Let $\mathbb{L}_s(\mathbf{d}, \boldsymbol{\nu}, C)$ denote the loss probability for each service s, where $\mathbf{d} = (d_1, d_2, \ldots, d_S)$ and $\boldsymbol{\nu} = (\nu_1, \nu_2, \ldots, \nu_S)$. Then the carried loads are $Y_s = \nu_s[1 - \mathbb{L}_s(\mathbf{d}, \boldsymbol{\nu}, C)]$, $s = 1, 2, \ldots, S$. Of practical importance is the inverse problem of determining the traffic intensities $\boldsymbol{\nu}$ from the measured values of the carried loads Y_s ($s = 1, 2, \ldots, S$). Once the traffic intensities are known, the loss probabilities are readily determined.

A particular application of such resource sharing is in telecommunication networks, where the provider wants to ensure that service level agreements with customers are met [1], [2]. Specifically, the loss probabilities should not exceed prescribed values. At the same time, the provider wants to verify that customers do not exceed agreed-upon peak traffic intensities.

*Received by the editors April 30, 2001; accepted for publication (in revised form) April 3, 2002; published electronically September 5, 2002.
http://www.siam.org/journals/siap/63-1/38879.html
†Bell Laboratories, Lucent Technologies, 600 Mountain Avenue, Murray Hill, NJ 07974 (johnmorrison@lucent.com).
‡Winphoria Networks, 2 Highwood Drive, Tewksbury, MA 01876 (ram@winphoria.com). This work was performed while the author was at Bell Laboratories, Lucent Technologies, Murray Hill, NJ.

The model considered above applies to each link of the network. The results for links influence the design of the network and the routing of traffic through the network, so as to optimize revenue while meeting the service level agreements with customers [11], [12], [13]. The model also applies in policing, shaping, access, and call admission control of customer traffic at the edge of the network [17].

In this paper we assume that the capacity C and the traffic intensities $\boldsymbol{\nu}$ are commensurately large and that C is an integer. We also assume that d_s $(s = 1, 2, \ldots, S)$ are positive integers, not large relative to C, and, without loss of generality, that the greatest common divisor of $d_1, \ldots, d_S$ is 1. There are three regimes in which the behavior of the loss probabilities differs. The resource is overloaded, critically loaded, or underloaded depending on whether the total traffic intensity $\sum_{s=1}^{S} d_s \nu_s$ exceeds, is close to, or is less than its capacity C, respectively.

In section 2 we consider the critically loaded regime in which $C - \sum_{s=1}^{S} d_s \nu_s = \delta\sqrt{C}$, where $\delta = O(1)$ may have either sign. The lowest order asymptotic approximation (see [3], [10], [16]) to $\mathbb{L}_s(\mathbf{d}, \boldsymbol{\nu}, C)$ implies that $\nu_s \sim Y_s(1 + d_s\beta/\sqrt{C})$, $s = 1, 2, \ldots, S$, where β is independent of s. Both δ and β are determined asymptotically in terms of Y_r $(r = 1, 2, \ldots, S)$, where $0 < C - \sum_{r=1}^{S} d_r Y_r = O(\sqrt{C})$, from the unique real solution of a nonlinear equation. This leads to a critically loaded asymptotic approximation (CLAA) to the traffic intensities. We then use a refined approximation to $\mathbb{L}_s(\mathbf{d}, \boldsymbol{\nu}, C)$, which is obtained by specializing the uniform asymptotic approximation [10] to the critically loaded regime. This leads to a refined critically loaded asymptotic approximation (RCLAA) to the traffic intensities.

In section 3 we present an iterative refinement procedure for solving the nonlinear equation and show that it leads to a contraction mapping and to monotonic and geometric convergence. We also establish the connection between our iterative refinement procedure and the one proposed by Mitra [9] for the case of general loading.

In section 4 we consider the overloaded regime in which $0 > C - \sum_{s=1}^{S} d_s \nu_s = O(C)$. It is shown that asymptotically $\nu_s \sim Y_s[(1 + \gamma)/\gamma]^{d_s}$, where $0 < \gamma = C - \sum_{r=1}^{S} d_r Y_r = O(1)$. It is also shown that this result matches asymptotically with those for the critically loaded regime for $\gamma \gg 1$ and $0 < \gamma/\sqrt{C} \ll 1$.

In section 5 we present numerical results for the two examples considered in [14]. We first compare the CLAA and the RCLAA to the traffic intensities with the exact results. We emphasize that both the CLAA and RCLAA are valid only in the critically loaded regime, and that the former uses $O(1)$ and $O(1/\sqrt{C})$ terms, while the latter also includes the $O(1/C)$ term in the expansion. The CLAA gives moderately accurate values in the critically loaded regime, but less so in the overloaded regime. On the other hand, the RCLAA gives quite accurate values in the critically loaded regime, and moderately accurate results in the overloaded regime. Overall, the RCLAA provides a significant improvement of the CLAA, with only minimal additional numerical computations. Both results do become quite accurate in the underloaded regime, since the loss probabilities, although not well approximated there by the CLAA or RCLAA, are exponentially small. We also present the results of some numerical experiments using the iterative procedure proposed by Mitra [9]. The number of iterations required to obtain the required accuracy increases significantly with the load.

The above asymptotic approximations may be applied to each link of a multirate loss network [11] to determine the reduced load offered to the link.

2. Critically loaded regime. In the critically loaded regime,

$$C - \sum_{s=1}^{S} d_s \nu_s = \delta\sqrt{C}, \tag{2.1}$$

where $C \gg 1$, $\nu_s = O(C)$, $s = 1, 2, \ldots, S$, and $\delta = O(1)$ may have either sign. In Appendix A we derive a refined asymptotic approximation to the loss probability $\mathbb{L}_s(\mathbf{d}, \boldsymbol{\nu}, C)$ for service s. The approximation is obtained by specializing the uniform asymptotic approximation [10] to the critically loaded regime.

Let

$$\sigma^2 C = 2\sum_{s=1}^{S} d_s^2 \nu_s, \quad \sigma > 0, \qquad \eta C = \sum_{s=1}^{S} d_s^3 \nu_s, \tag{2.2}$$

so that $\sigma = O(1)$ and $\eta = O(1)$, since $\nu_s = O(C)$, $s = 1, 2, \ldots, S$, and

$$\beta = \frac{2e^{-(\delta/\sigma)^2}}{\sigma\sqrt{\pi}\,\mathrm{Erfc}\,(-\delta/\sigma)}, \tag{2.3}$$

where the complementary error function is given by

$$\mathrm{Erfc}\,(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-u^2}\, du\,. \tag{2.4}$$

Then, asymptotically,

$$\mathbb{L}_s(\mathbf{d}, \boldsymbol{\nu}, C) \sim \frac{d_s\beta}{\sqrt{C}} \left\{ 1 + \frac{\delta}{\sigma^2\sqrt{C}} \left[\frac{2\eta}{\sigma^2} \left(\frac{2\delta^2}{3\sigma^2} - 1 \right) + d_s - 1 \right] \right. \\ \left. - \frac{\beta}{2\sqrt{C}} \left[1 + \frac{2\eta}{3\sigma^2} \left(1 - \frac{2\delta^2}{\sigma^2} \right) \right] + \cdots \right\}. \tag{2.5}$$

Hence, asymptotically,

$$\begin{aligned} \nu_s &= \frac{Y_s}{[1 - \mathbb{L}_s(\mathbf{d}, \boldsymbol{\nu}, C)]} \\ &\sim Y_s \left\{ 1 + \frac{d_s\beta}{\sqrt{C}} + \frac{d_s\beta\delta}{\sigma^2 C} \left[\frac{2\eta}{\sigma^2} \left(\frac{2\delta^2}{3\sigma^2} - 1 \right) + d_s - 1 \right] \right. \\ &\quad \left. - \frac{d_s\beta^2}{2C} \left[1 + \frac{2\eta}{3\sigma^2} \left(1 - \frac{2\delta^2}{\sigma^2} \right) \right] + \frac{d_s^2\beta^2}{C} + \cdots \right\}. \end{aligned} \tag{2.6}$$

We expand in powers of $1/\sqrt{C}$ and let

$$\delta \sim \delta_0 + \frac{\delta_1}{\sqrt{C}} + \cdots, \qquad \sigma \sim \sigma_0 + \frac{\sigma_1}{\sqrt{C}} + \cdots \tag{2.7}$$

and

$$\beta \sim \beta_0 + \frac{\beta_1}{\sqrt{C}} + \cdots, \qquad \eta \sim \eta_0 + \cdots. \tag{2.8}$$

Then, from (2.1)–(2.3) and (2.6), to lowest order,

$$\delta_0\sqrt{C} = C - \sum_{s=1}^{S} d_s Y_s - \frac{\beta_0}{\sqrt{C}} \sum_{s=1}^{S} d_s^2 Y_s, \tag{2.9}$$

$$\sigma_0^2 C = 2\sum_{s=1}^{S} d_s^2 Y_s, \quad \sigma_0 > 0, \qquad \eta_0 C = \sum_{s=1}^{S} d_s^3 Y_s, \tag{2.10}$$

and

$$\beta_0 = \frac{2e^{-(\delta_0/\sigma_0)^2}}{\sigma_0\sqrt{\pi}\,\mathrm{Erfc}\,(-\delta_0/\sigma_0)}\,. \tag{2.11}$$

If we divide (2.9) by $\sigma_0\sqrt{C}$ and use (2.10) and (2.11), we obtain

$$\frac{\delta_0}{\sigma_0} + \frac{e^{-(\delta_0/\sigma_0)^2}}{\sqrt{\pi}\,\mathrm{Erfc}\,(-\delta_0/\sigma_0)} = \frac{\left(C - \sum_{s=1}^{S} d_s Y_s\right)}{\sqrt{2\sum_{s=1}^{S} d_s^2 Y_s}}\,. \tag{2.12}$$

Intuitively, $\sum_{s=1}^{S} d_s Y_s < C$, i.e., the total carried capacity is less than that of the resource. We let

$$a \triangleq \frac{(C - \sum_{s=1}^{S} d_s Y_s)}{\sqrt{2\sum_{s=1}^{S} d_s^2 Y_s}} > 0 \tag{2.13}$$

and

$$g(x) = \frac{e^{-x^2}}{\sqrt{\pi}\,\mathrm{Erfc}\,(x)} - x. \tag{2.14}$$

To lowest order, the inverse problem is asymptotically equivalent to solving the equation $g(\xi) = a$ for $\xi = -\delta_0/\sigma_0$. Then, from (2.11)–(2.13),

$$\beta_0\sigma_0 = 2(\xi + a), \tag{2.15}$$

and, from (2.6) and (2.10), we obtain the first order asymptotic approximation to the traffic intensities ν_s in terms of the carried loads Y_s,

$$\nu_s \sim Y_s\left(1 + \frac{d_s\beta_0}{\sqrt{C}}\right) = Y_s\left[1 + \frac{2d_s(\xi + a)}{\sqrt{2\sum_{r=1}^{S} d_r^2 Y_r}}\right]. \tag{2.16}$$

We note, from (2.14), that $g(0) = 1/\sqrt{\pi}$. We show in Appendix B that $g(-\infty) = \infty$, $g(\infty) = 0$, and $g'(x) < 0$ for $x < \infty$. Hence, there is a unique real solution to $g(\xi) = a > 0$. Since $g(-a) > a$, it follows that $\xi > -a$. In Figure 2.1, $\xi + a$ is depicted as a function of a. Because of the monotonicity, the equation for ξ may be solved numerically by any simple procedure, such as bisection or iterative refinement. In the next section we present an iterative refinement procedure for solving the equation, which leads to a contraction mapping and to monotonic and geometric convergence. This is important, since Mitra [9] has conjectured that his

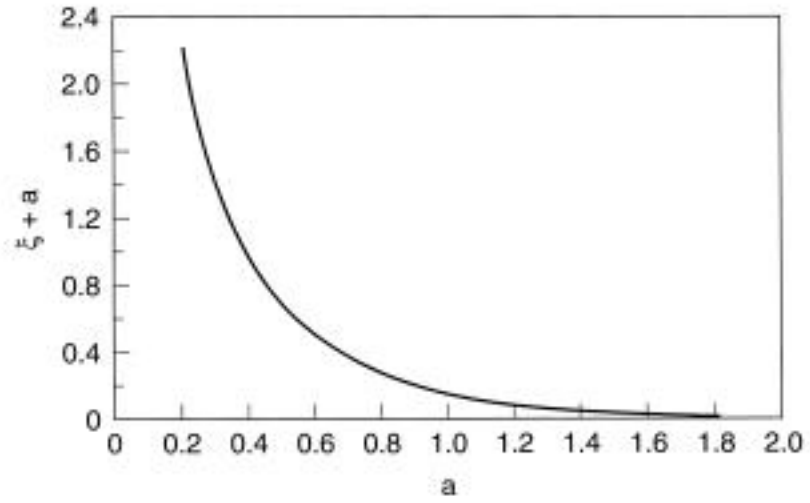


FIG. 2.1. *Plot of $\xi + a$ as a function of a, where $g(\xi) = a$.*

proposed iterative procedure, for the case of general loading, leads to a contraction mapping.

To the next order, from (2.6)–(2.8), we have

$$\nu_s \sim Y_s\left\{1 + d_s\left(\frac{\beta_0}{\sqrt{C}} + \frac{\beta_1}{C}\right) + \frac{d_s\beta_0\delta_0}{\sigma_0^2 C}\left[\frac{2\eta_0}{\sigma_0^2}\left(\frac{2\delta_0^2}{3\sigma_0^2} - 1\right) + d_s - 1\right]\right. \\ \left. - \frac{d_s\beta_0^2}{2C}\left[1 + \frac{2\eta_0}{3\sigma_0^2}\left(1 - \frac{2\delta_0^2}{\sigma_0^2}\right)\right] + \frac{d_s^2\beta_0^2}{C} + \cdots\right\}. \tag{2.17}$$

Hence, from (2.1), (2.9), and (2.10), we obtain

$$\delta_1 = -\frac{1}{2}\sigma_0^2\beta_1 + \frac{1}{2}\beta_0\delta_0 + \frac{1}{4}\beta_0^2\sigma_0^2 - \frac{1}{3}\beta_0\eta_0\left(\frac{5}{2}\beta_0 + \frac{\beta_0\delta_0^2}{\sigma_0^2} + \frac{2\delta_0^3}{\sigma_0^4}\right). \tag{2.18}$$

It follows that

$$\nu_s \sim Y_s\left[1 + \frac{d_s\beta_0}{\sqrt{C}} - \frac{2\delta_1 d_s}{\sigma_0^2 C} + \frac{\beta_0}{C}\left(\beta_0 + \frac{\delta_0}{\sigma_0^2}\right)d_s\left(d_s - \frac{2\eta_0}{\sigma_0^2}\right) + \cdots\right]. \tag{2.19}$$

Also, from (2.2), we obtain

$$\sigma_0\sigma_1 = \beta_0\eta_0. \tag{2.20}$$

Next,

$$\frac{\delta}{\sigma} \sim \frac{1}{\sigma_0}\left[\delta_0 + \frac{1}{\sqrt{C}}\left(\delta_1 - \frac{\delta_0\sigma_1}{\sigma_0}\right) + \cdots\right] \tag{2.21}$$

and

$$\beta\sigma \sim \beta_0\sigma_0 + \frac{1}{\sqrt{C}}(\sigma_0\beta_1 + \beta_0\sigma_1) + \cdots. \tag{2.22}$$

Hence, from (2.3), (2.4), and (2.11), after a straightforward calculation, it is found that

$$\sigma_0\beta_1 + \beta_0\sigma_1 = -\beta_0\left(\delta_1 - \frac{\delta_0\sigma_1}{\sigma_0}\right)\left(2\frac{\delta_0}{\sigma_0} + \beta_0\sigma_0\right). \tag{2.23}$$

If we use (2.18) and (2.20) to eliminate β_1 and σ_1 from (2.23), we obtain

$$\begin{aligned}\delta_1\left[1 - \frac{1}{2}\beta_0\sigma_0\left(2\frac{\delta_0}{\sigma_0} + \beta_0\sigma_0\right)\right] \\ = \frac{1}{2}\beta_0\delta_0 + \frac{1}{4}\beta_0^2\sigma_0^2 - \frac{1}{2}\beta_0^3\delta_0\eta_0 - \frac{\beta_0\eta_0}{3\sigma_0^2}\left(\beta_0\sigma_0^2 + 4\beta_0\delta_0^2 + 2\frac{\delta_0^3}{\sigma_0^2}\right).\end{aligned} \tag{2.24}$$

Since $\delta_0/\sigma_0 = -\xi$, it follows from (2.15) that

$$(1 - 2a^2 - 2a\xi)\delta_1 = (\xi + a)\left[a + \frac{4\eta_0}{3\sigma_0^2}(2a\xi^2 + 3a^2\xi - \xi - a)\right], \tag{2.25}$$

and from (2.19) that

$$\nu_s \sim Y_s\left\{1 + \frac{2d_s(\xi + a)}{\sigma_0\sqrt{C}} - \frac{2d_s}{\sigma_0^2 C}\left[\delta_1 + (\xi + a)(\xi + 2a)\left(\frac{2\eta_0}{\sigma_0^2} - d_s\right)\right]\right\}. \tag{2.26}$$

Hence, from (2.10), we obtain the refined asymptotic approximation

$$\begin{aligned}\nu_s \sim Y_s\Bigg\{1 + \frac{2d_s(\xi + a)}{\sqrt{2\sum_{r=1}^S d_r^2 Y_r}} + \frac{d_s(\xi + a)}{\sum_{r=1}^S d_r^2 Y_r}\left[d_s(\xi + 2a) - \frac{a}{(1 - 2a^2 - 2a\xi)}\right] \\ + \frac{d_s(\xi + a)\left(\sum_{r=1}^S d_r^3 Y_r\right)}{3\left(\sum_{r=1}^S d_r^2 Y_r\right)^2}\left[\frac{a(1 + 2a^2)}{(1 - 2a^2 - 2a\xi)} - (\xi + 5a)\right]\Bigg\}.\end{aligned} \tag{2.27}$$

We note, from (2.4) and (2.14), that

$$g'(\xi) = 2g(\xi)[g(\xi) + \xi] - 1. \tag{2.28}$$

Since $g(\xi) = a$, it follows that

$$1 - 2a^2 - 2a\xi = -g'(\xi) > 0. \tag{2.29}$$

Once ξ has been computed, the numerical calculation of the refined approximation (2.27) is straightforward.

3. Iterative refinement procedure. Let $f(x) = g(x) + x$, where $g(x)$ is given by (2.14). Then, the inverse problem is asymptotically equivalent to solving the equation

$$f(\xi) - \xi = a. \tag{3.1}$$

We consider the algorithm

$$\xi(m + 1) = f(\xi(m)) - a, \quad m = 0, 1, \ldots, \qquad \xi(0) < \infty. \tag{3.2}$$

In order to establish that this is a contraction mapping, we make use of the following, which is proved in Appendix C.

LEMMA 3.1. *Let*

$$f(x) = \frac{e^{-x^2}}{\sqrt{\pi}\,\mathrm{Erfc}\,(x)} = \frac{e^{-x^2}}{2\int_x^\infty e^{-u^2}du}. \tag{3.3}$$

Then $f'(x) > 0$ *and* $f''(x) > 0$ *for* $-\infty < x < \infty$.

We remark that this lemma is of importance in another problem. Reiman [15] observed that $h(x) = \sqrt{2}\,f(x/\sqrt{2})$ is the hazard rate of a standard (mean 0, variance 1) normal random variable. He made use of his assertion that h is strictly increasing and strictly convex, but provided no proof of the latter.

From (3.1) and (3.2), we have

$$\xi(m+1) - \xi = f(\xi(m)) - f(\xi). \tag{3.4}$$

Since $f'(x) > 0$, $\xi(m) \geq \xi \Rightarrow \xi(m+1) \geq \xi$ and $\xi(m) \leq \xi \Rightarrow \xi(m+1) \leq \xi$. By induction,

$$\xi(0) \geq \xi \Rightarrow \xi(m) \geq \xi, \quad m = 0, 1, \ldots, \tag{3.5}$$

and

$$\xi(0) \leq \xi \Rightarrow \xi(m) \leq \xi, \quad m = 0, 1, \ldots . \tag{3.6}$$

Now $f'(x) = g'(x) + 1 < 1$, since $g'(x) < 0$ for $x < \infty$. We suppose first that $\xi(0) \geq \xi$. Then,

$$0 \leq \xi(m+1) - \xi = \int_\xi^{\xi(m)} f'(x)dx \leq \xi(m) - \xi, \tag{3.7}$$

so that $\xi(m)$ is a monotonically nonincreasing function of m, and $\xi(m) \leq \xi(0)$, $m = 0, 1, \ldots$. Hence, since $f''(x) > 0$, we have

$$\xi(0) \geq \xi \Rightarrow |\xi(m+1) - \xi| \leq f'[\xi(0)]\ |\xi(m) - \xi|, \quad m = 0, 1, \ldots . \tag{3.8}$$

On the other hand, if $\xi(0) \leq \xi$, then

$$0 \leq \xi - \xi(m+1) = \int_{\xi(m)}^{\xi} f'(x)dx \leq \xi - \xi(m), \tag{3.9}$$

so that $\xi(m)$ is a monotonically nondecreasing function of m. Hence, from (3.6), since $f''(x) > 0$, we have

$$\xi(0) \leq \xi \Rightarrow |\xi(m+1) - \xi| \leq f'(\xi)\ |\xi(m) - \xi|, \quad m = 0, 1, \ldots . \tag{3.10}$$

We let

$$\alpha = \max\{f'(\xi), f'[\xi(0)]\} = f'\{\max[\xi, \xi(0)]\} < 1. \tag{3.11}$$

Then, from (3.8) and (3.10), we obtain

$$|\xi(m+1) - \xi| \leq \alpha|\xi(m) - \xi|, \quad m = 0, 1, \ldots . \tag{3.12}$$

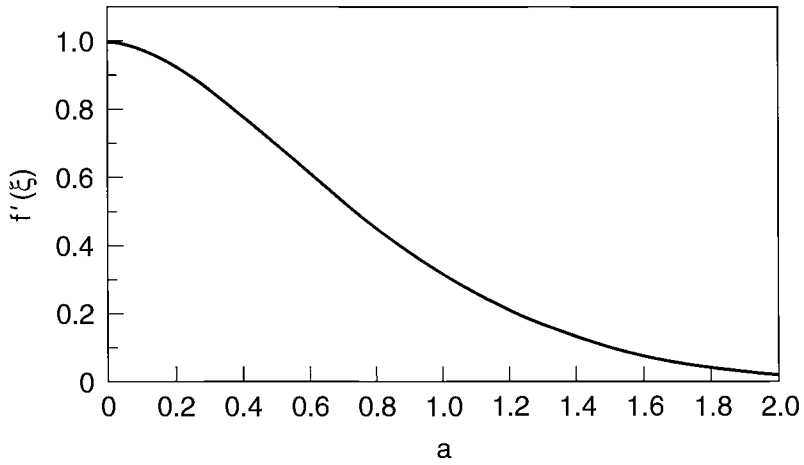


FIG. 3.1. *Plot of the convergence factor $f'(\xi)$ as a function of a.*

Hence $\xi(m)$ converges geometrically and monotonically to ξ. We note, in particular, that if $\xi(0) = -a$, then $\xi > \xi(0)$ and $\alpha = f'(\xi)$. In Figure 3.1, $f'(\xi)$ is depicted as a function of a. Since $f'(\xi) \to 1$ as $a \to 0$, $\xi(m)$ converges slowly for small values of a. Note, from (2.13), that as $a \to 0$ the total carried capacity approaches the capacity of the resource, and we are exiting the critically loaded regime and entering the overloaded one. Consequently, in the next section we analyze the overloaded regime, in which $0 < C - \sum_{s=1}^{S} d_s Y_s = O(1)$.

We conclude this section by establishing the connection between the algorithm (3.2) and the iterative procedure proposed by Mitra [9] for the case of general loading. The general algorithm is

$$\nu_s(m+1) = Y_s + \nu_s(m)\mathbb{L}_s(\mathbf{d}, \boldsymbol{\nu}(m), C), \quad m = 0, 1, \ldots . \tag{3.13}$$

Mitra conjectured that (3.13) leads to a contraction mapping, but its convergence for any initial guess is an open question. If we multiply this equation by d_s and sum on s, we obtain, from (2.1),

$$\begin{aligned} \delta(m+1)\sqrt{C} &= C - \sum_{s=1}^{S} d_s \nu_s(m+1) \\ &= C - \sum_{s=1}^{S} d_s Y_s - \sum_{s=1}^{S} d_s \nu_s(m)\mathbb{L}_s(\mathbf{d}, \boldsymbol{\nu}(m), C). \end{aligned} \tag{3.14}$$

But, from (2.5) and (2.6), to lowest order,

$$\mathbb{L}_s(\mathbf{d}, \boldsymbol{\nu}(m), C) \sim \frac{d_s \beta(m)}{\sqrt{C}}, \qquad \nu_s(m) \sim Y_s, \tag{3.15}$$

and, from (2.2),

$$\sigma(m) \sim \sqrt{\frac{2}{C} \sum_{s=1}^{S} d_s^2 Y_s} \sim \sigma(m+1). \tag{3.16}$$

Hence, asymptotically, from (3.14)–(3.16), we obtain

$$\frac{\delta(m+1)}{\sigma(m+1)} + \frac{1}{2}\sigma(m)\beta(m) \sim \frac{\left(C - \sum_{s=1}^{S} d_s Y_s\right)}{\sqrt{2\sum_{s=1}^{S} d_s^2 Y_s}}. \tag{3.17}$$

From (2.3) and (2.13), we have

$$\frac{\delta(m+1)}{\sigma(m+1)} + \frac{e^{-[\delta(m)/\sigma(m)]^2}}{\sqrt{\pi}\,\mathrm{Erfc}\,[-\delta(m)/\sigma(m)]} \sim a. \tag{3.18}$$

Since $\xi(m) \sim -\delta(m)/\sigma(m)$, this corresponds asymptotically to (3.2), in view of (3.3).

4. Overloaded regime. In the overloaded regime, $0 > C - \sum_{s=1}^{S} d_s\nu_s = O(C)$, and the loss probability $\mathbb{L}_s(\mathbf{d}, \boldsymbol{\nu}, C)$ for service s is asymptotically given by (see [6], [10])

$$\mathbb{L}_s(\mathbf{d}, \boldsymbol{\nu}, C) = 1 - (z^*)^{d_s} + O\left(\frac{1}{C}\right), \tag{4.1}$$

where z^* is the unique positive solution of

$$\sum_{s=1}^{S} d_s\nu_s(z^*)^{d_s} = C, \quad 0 < z^* < 1. \tag{4.2}$$

Hence,

$$Y_s = \nu_s[1 - \mathbb{L}_s(\mathbf{d}, \boldsymbol{\nu}, C)] = \nu_s(z^*)^{d_s} + O(1) \tag{4.3}$$

and

$$0 < \gamma \triangleq C - \sum_{s=1}^{S} d_s Y_s = O(1). \tag{4.4}$$

To determine z^* in terms of Y_s ($s = 1, 2, \ldots, S$), we derive an asymptotic approximation to the correction term in (4.4). In Appendix D we establish that

$$\gamma = C - \sum_{s=1}^{S} d_s Y_s \sim \frac{z^*}{(1 - z^*)}. \tag{4.5}$$

Hence,

$$z^* \sim \frac{\gamma}{(1+\gamma)}, \tag{4.6}$$

and, from (4.3), it follows that

$$\nu_s \sim Y_s \left(\frac{1+\gamma}{\gamma}\right)^{d_s}. \tag{4.7}$$

We will show that this result matches asymptotically with those for the critically loaded regime for $\gamma \gg 1$ and $0 < \gamma/\sqrt{C} \ll 1$. But $\gamma = a\sigma_0\sqrt{C}$, from (2.10), (2.13),

and (4.4), so this corresponds to $0 < a \ll 1$. Since $g(\xi) = a$, it follows from (2.14) and (B.1) that

$$\xi = \frac{1}{2a}[1 - 4a^2 + O(a^4)]\,, \tag{4.8}$$

so that

$$1 - 2a^2 - 2a\xi = 2a^2 + O(a^4). \tag{4.9}$$

Hence, from (2.10) and (2.27), we obtain for small a

$$\nu_s \sim Y_s \left\{ 1 + \frac{d_s}{a\sigma_0\sqrt{C}}[1 + O(a^2)] + \frac{d_s}{2a^2\sigma_0^2 C}[d_s - 1 + O(a^2)] \right\}. \tag{4.10}$$

This matches with (4.7) for $\gamma = a\sigma_0\sqrt{C} \gg 1$.

5. Numerical results. The CLAA to the traffic intensities ν_s, in terms of the carried loads Y_r $(r = 1, \ldots, S)$, is given by (2.16), where $\xi > -a$ satisfies $f(\xi) - \xi = a$, and a and $f(x)$ are given by (2.13) and (3.3). Numerical values of ξ were obtained by the iterative refinement procedure (3.2), with $\xi(0) = -a$. The stopping criterion $|[\xi(m+1) - \xi(m)]/[\xi(m+1) + a]| \le 10^{-4}$ was used. The RCLAA is given by (2.27).

We present numerical results for the two examples considered in [14]. The first example has capacity $C = 96$ and two services with call capacities $d_1 = 1$ and $d_2 = 6$. In Table 1, the CLAA and RCLAA to the traffic intensities ν_1 and ν_2 are compared with the values from which the carried loads Y_1 and Y_2 were calculated, using the exact values of the loss probabilities tabulated in [14]. These values are $\nu_1 = 0.86\nu/1.7$ and $\nu_2 = 0.14\nu/1.7$, where the total traffic intensity $\nu = \nu_1 + 6\nu_2 = 96 + \beta\sqrt{96}$, and $\beta = -6(1)3$, ranging from an underloaded resource to an overloaded one. The CLAA gives moderately accurate values in the critically loaded regime, but less so in the overloaded regime. The RCLAA, however, gives significantly better results in these regimes. Errors of more than 1% are indicated in parentheses in Table 1. The results become quite accurate in the underloaded regime, since the loss probabilities, although not well approximated by CLAA or RCLAA there, are exponentially small. Also listed in Table 1 is the number of iterations required to obtain $\xi + a$ to the desired accuracy, which increases significantly with the load.

Similar comparisons are made in Table 2 for the second example, which has capacity $C = 130$, and three services with call capacities $d_1 = 1$, $d_2 = 3$, and $d_3 = 10$. The traffic intensities from which the carried loads Y_1, Y_2, and Y_3 were calculated exactly are $\nu_1 = 2\nu/13 = \nu_2$ and $\nu_3 = \nu/26$, where the total traffic intensity is $\nu = \nu_1 + 3\nu_2 + 10\nu_3$, and $\nu = 46.8(20.8)171.6$, again ranging from an underloaded resource to an overloaded one. (For $\nu = 67.6$, the exact value of $\mathbb{L}_3$ in [14] should be 4.9623×10^{-3}.) Similar comments apply to the CLAA and RCLAA here as for the first example. Errors of more than 1% are indicated in parentheses in Table 2. Also listed in Table 2 is the number of iterations required to obtain $\xi + a$ to the desired accuracy. All computations were performed with double precision.

The overloaded asymptotic approximation (4.7) gives less accurate results than the RCLAA, even for the heaviest loads considered for the two examples. Thus, for $\nu = 125.4$ in Table 1, (4.7) gives the approximations 64.9 and 11.3 for ν_1 and ν_2, respectively. For $\nu = 171.6$ in Table 2, (4.7) gives the approximations 26.8, 27.6, and 7.23 for ν_1, ν_2, and ν_3, respectively.

TABLE 1

Comparisons of the CLAA and RCLAA to the traffic intensities ν_1 and ν_2 with the exact results for two services with $C = 96$, $d_1 = 1$ and $d_2 = 6$, and total traffic intensity $\nu = \nu_1 + 6\nu_2$.

Carried loads	CLAA	(% err.)	RCLAA	(% err.)	Exact	ν	# of iter.
18.8249	18.8249		18.8249		18.8250	37.2	2
3.06434	3.0643		3.0644		3.0645		
23.779	23.779		23.781		23.782	47.0	3
3.8668	3.867		3.869		3.871		
28.709	28.72		28.74		28.74	56.8	4
4.6372	4.65		4.67		4.68		
33.546	33.67		33.70		33.69	66.6	6
5.2964	5.42	(1.3)	5.48		5.49		
38.180	38.64		38.65		38.65	76.4	8
5.7476	6.16	(2.1)	6.28		6.29		
42.529	43.63		43.60		43.61	86.2	12
5.9537	6.88	(3.1)	7.07		7.10		
46.574	48.63		48.55		48.56	96.0	17
5.9496	7.52	(4.9)	7.84		7.91		
50.337	53.62		53.50		53.52	105.8	24
5.7967	8.07	(7.3)	8.59	(1.4)	8.71		
53.848	58.61		58.45		58.48	115.6	31
5.5503	8.50	(10.7)	9.30	(2.3)	9.52		
57.137	63.60		63.40		63.43	125.4	39
5.2494	8.81	(14.7)	9.98	(3.4)	10.33		

We also performed some numerical experiments using the algorithm (3.13) proposed by Mitra [9]. The loss probabilities $\mathbb{L}_s$ were evaluated by means of the refined uniform asymptotic approximation (RUAA) derived in [14], which was shown to be very accurate for the two examples considered. The initial values $\nu_s(0)$, $s = 1, 2, \ldots, S$, were chosen either as Y_s (calculated exactly) or as $Z_s = Y_s[(1+\gamma)/\gamma]^{d_s}$, where γ is given by (4.4), corresponding to the approximation (4.7) for the overloaded regime.

In Table 3 we compare the results of the iterative procedure with the (more precisely listed) exact values of the traffic intensities ν_1 and ν_2 for the first example. The corresponding carried loads are given in Table 1. Also listed is the number of iterations required to obtain the desired accuracy. As indicated by our critically loaded asymptotic analysis, the number of iterations increases significantly with the load. For the most heavily loaded case considered, 8 fewer iterations were required when $\nu_s(0) = Z_s$ than when $\nu_s(0) = Y_s$, $s = 1, 2$.

In Table 4 we compare the results of the iterative procedure with the exact values of the traffic intensities ν_1, ν_2, and ν_3 for the second example. The corresponding carried loads are given in Table 2. Similar comments apply to the number of iterations required.

Appendix A. We derive here the refined asymptotic approximation (2.5) to the loss probability $\mathbb{L}_s(\mathbf{d}, \boldsymbol{\nu}, C)$ for service s, by specializing the uniform asymptotic

TABLE 2

Comparisons of the CLAA and RCLAA to the traffic intensities ν_1, ν_2, and ν_3 with the exact results for three services with $C = 130$, $d_1 = 1$, $d_2 = 3$, and $d_3 = 10$, and total traffic intensity $\nu = \nu_1 + 3\nu_2 + 10\nu_3$.

Carried loads	CLAA	(% err.)	RCLAA	(% err.)	Exact	ν	# of iter.
7.19997	7.19997		7.19997		7.2	46.8	2
7.19989	7.19989		7.19990		7.2		
1.79983	1.79983		1.79984		1.8		
10.3970	10.3980		10.3998		10.4	67.6	3
10.3898	10.3928		10.3986		10.4		
2.58710	2.5896		2.5959		2.6		
13.5563	13.59		13.601		13.6	88.4	6
13.5491	13.57		13.603		13.6		
3.24937	3.34	(1.8)	3.396		3.4		
16.5875	16.80		16.798		16.8	109.2	10
16.1393	16.76		16.797		16.8		
3.58302	4.04	(3.8)	4.180		4.2		
19.4315	20.01		19.994		20.0	130	17
18.2869	19.93		19.986		20.0		
3.57130	4.64	(7.2)	4.929	(1.4)	5.0		
22.0957	23.23		23.192		23.2	150.8	26
19.9652	23.03		23.177		23.2		
3.35269	5.07	(12.6)	5.620	(3.1)	5.8		
24.6070	26.44		26.391		26.4	171.6	37
21.2824	26.03	(1.4)	26.364		26.4		
3.04643	5.31	(19.5)	6.229	(5.6)	6.6		

approximation [10] to the critically loaded regime. Let

$$\nu_s = \alpha_s C, \quad s = 1, 2, \ldots, S, \tag{A.1}$$

where $C \gg 1$ and $\alpha_s > 0$ is $O(1)$ and bounded away from zero. Also, let

$$f(z) = \sum_{s=1}^{S} \alpha_s (z^{d_s} - 1) - \log z. \tag{A.2}$$

There is a unique positive solution z^* of $f'(z) = 0$ so that

$$\sum_{s=1}^{S} \alpha_s d_s (z^*)^{d_s} = 1, \quad z^* > 0. \tag{A.3}$$

It follows from (A.2) and (A.3) that

$$v \triangleq (z^*)^2 f''(z^*) = \sum_{s=1}^{S} \alpha_s d_s^2 (z^*)^{d_s}. \tag{A.4}$$

We define

$$K = \frac{1}{(1 - z^*)} - \frac{\sqrt{v}\,\mathrm{sgn}\,(1 - z^*)}{\sqrt{-2f(z^*)}}, \quad z^* \neq 1, \tag{A.5}$$

TABLE 3

Comparison of the final iterated values of the traffic intensities ν_1 and ν_2 with the exact results for two services with $C = 96$, $d_1 = 1$ and $d_2 = 6$, and total traffic intensity $\nu = \nu_1 + 6\nu_2$.

ν	Exact	$\nu_s(0) = Y_s$	# of iter.	$\nu_s(0) = Z_s$	# of iter.
37.2	18.8250 3.06454	18.8250 3.06454	3	18.8250 3.06454	4
47.0	23.7816 3.87143	23.7819 3.87141	3	23.7819 3.87141	5
56.8	28.7383 4.67832	28.7381 4.67828	5	28.7381 4.67829	6
66.6	33.6949 5.48521	33.6947 5.48515	7	33.6947 5.48515	8
76.4	38.6515 6.29210	38.6515 6.29210	10	38.6515 6.29215	11
86.2	43.6081 7.0990	43.6081 7.0988	15	43.6083 7.0991	16
96.0	48.5647 7.9059	48.5645 7.9055	22	48.5653 7.9065	22
105.8	53.5213 8.7128	53.5208 8.7112	31	53.5233 8.7141	30
115.6	58.478 9.520	58.475 9.517	43	58.481 9.523	39
125.4	63.435 10.327	63.429 10.320	57	63.442 10.333	49

$$K = \frac{1}{2} + \frac{1}{6v}\sum_{s=1}^{S} \alpha_s d_s^3, \quad z^* = 1, \tag{A.6}$$

and

$$M = \frac{1}{2}\,\mathrm{Erfc}\left[\mathrm{sgn}\,(1-z^*)\sqrt{-Cf(z^*)}\right] + \frac{Ke^{Cf(z^*)}}{\sqrt{2\pi Cv}}\,. \tag{A.7}$$

Then (see [10]),

$$\mathbb{L}_s(\mathbf{d},\boldsymbol{\nu},C) = \frac{e^{Cf(z^*)}[(z^*)^{d_s}-1]}{\sqrt{2\pi Cv}(z^*-1)M}\left[1 + O\left(\frac{1}{C}\right)\right]. \tag{A.8}$$

Since $f(1) = 0$ and $f'(z^*) = 0$, it follows from (A.4) that the expression for K in (A.5) remains finite as $z^* \to 1$, and its limiting value is given by (A.6).

In the critically loaded regime corresponding to (2.1), z^* is close to 1 and we expand in powers of $1/\sqrt{C}$ and let

$$z^* \sim 1 + \frac{c_1}{\sqrt{C}} + \frac{c_2}{C} + \cdots. \tag{A.9}$$

TABLE 4

Comparison of the final iterated values of the traffic intensities ν_1, ν_2, and ν_3 with the exact results for three services with $C = 130$, $d_1 = 1$, $d_2 = 3$, and $d_3 = 10$, and total traffic intensity $\nu = \nu_1 + 3\nu_2 + 10\nu_3$.

ν	Exact	$\nu_s(0) = Y_s$	# of iter.	$\nu_s(0) = Z_s$	# of iter.
46.8	7.2	7.20000	3	7.20000	4
	7.2	7.20000		7.20000	
	1.8	1.80000		1.80000	
67.6	10.4	10.40003	4	10.40003	5
	10.4	10.40000		10.40000	
	2.6	2.60000		2.60000	
88.4	13.6	13.60003	7	13.60003	8
	13.6	13.60005		13.60005	
	3.4	3.40000		3.40000	
109.2	16.8	16.79996	13	16.79998	14
	16.8	16.79998		16.80006	
	4.2	4.19991		4.19999	
130	20.0	19.9998	22	20.0001	23
	20.0	19.9996		20.0003	
	5.0	4.9994		5.0002	
150.8	23.2	23.1995	37	23.2005	35
	23.2	23.1987		23.2019	
	5.8	5.7982		5.8013	
171.6	26.4	26.3985	57	26.4017	49
	26.4	26.395		26.405	
	6.6	6.595		6.604	

Then, from (2.1), (A.1), and (A.3), we obtain

$$\frac{\delta}{\sqrt{C}} \sim \frac{c_1}{\sqrt{C}} \sum_{s=1}^{S} \alpha_s d_s^2 + \frac{1}{C} \sum_{s=1}^{S} \alpha_s d_s^2 \left[c_2 + \frac{1}{2} c_1^2 (d_s - 1) \right] + \cdots . \tag{A.10}$$

Hence, from (2.2), we find that

$$c_1 = \frac{2\delta}{\sigma^2}, \qquad c_2 = -\frac{4\delta^2}{\sigma^6} \left(\eta - \frac{1}{2}\sigma^2 \right) . \tag{A.11}$$

Also, from (A.4) and (A.6),

$$v \sim \frac{1}{2}\sigma^2 + \frac{2\delta\eta}{\sigma^2\sqrt{C}} + \cdots, \qquad K \sim \frac{1}{2} + \frac{\eta}{3\sigma^2} + \cdots . \tag{A.12}$$

Since $f'(z^*) = 0$,

$$0 = f(1) = f(z^*) + \frac{1}{2}(1 - z^*)^2 f''(z^*) + \frac{1}{6}(1 - z^*)^3 f'''(z^*) + \cdots . \tag{A.13}$$

However, from (A.2), we obtain

$$z f''(z) + f'(z) = \sum_{s=1}^{S} \alpha_s d_s^2 z^{d_s - 1} \tag{A.14}$$

and

$$zf'''(z) + 2f''(z) = \sum_{s=1}^{S} \alpha_s d_s^2 (d_s - 1) z^{d_s - 2} . \tag{A.15}$$

From (2.2), (A.4), (A.9), (A.11), and (A.15), after some algebra, it is found that

$$f''(z^*) \sim \frac{1}{2}\sigma^2 + \frac{2\delta}{\sigma^2\sqrt{C}}(\eta - \sigma^2) + \cdots \tag{A.16}$$

and

$$f'''(z^*) \sim \eta - \frac{3}{2}\sigma^2 + \cdots . \tag{A.17}$$

It then follows from (A.13) that

$$-Cf(z^*) \sim \frac{\delta^2}{\sigma^2} - \frac{4\delta^3\eta}{3\sigma^6\sqrt{C}} + \cdots . \tag{A.18}$$

Hence, since $z^* < 1$ if $\delta < 0$,

$$\operatorname{sgn}(1 - z^*)\sqrt{-Cf(z^*)} \sim -\frac{\delta}{\sigma}\left(1 - \frac{2\delta\eta}{3\sigma^4\sqrt{C}} + \cdots\right) . \tag{A.19}$$

It follows from (A.18) that

$$e^{Cf(z^*)} \sim e^{-(\delta/\sigma)^2}\left(1 + \frac{4\delta^3\eta}{3\sigma^6\sqrt{C}} + \cdots\right) . \tag{A.20}$$

Also, from (2.4) and (A.19), we obtain

$$\operatorname{Erfc}\left[\operatorname{sgn}(1 - z^*)\sqrt{-Cf(z^*)}\right] \sim \operatorname{Erfc}\left(-\frac{\delta}{\sigma}\right) - \frac{4\delta^2\eta}{3\sigma^5\sqrt{\pi C}} e^{-(\delta/\sigma)^2} + \cdots , \tag{A.21}$$

and hence, from (A.7) and (A.12),

$$M \sim \frac{1}{2}\operatorname{Erfc}\left(-\frac{\delta}{\sigma}\right) + \frac{e^{-(\delta/\sigma)^2}}{\sigma\sqrt{\pi C}}\left[\frac{1}{2} + \frac{\eta}{3\sigma^2}\left(1 - \frac{2\delta^2}{\sigma^2}\right)\right] + \cdots . \tag{A.22}$$

Consequently, with β given by (2.3),

$$\frac{e^{-(\delta/\sigma)^2}}{M\sigma\sqrt{\pi}} \sim \beta\left\{1 - \frac{\beta}{2\sqrt{C}}\left[1 + \frac{2\eta}{3\sigma^2}\left(1 - \frac{2\delta^2}{\sigma^2}\right)\right] + \cdots\right\} . \tag{A.23}$$

Next, from (A.12),

$$\frac{1}{\sqrt{2v}} \sim \frac{1}{\sigma}\left(1 - \frac{2\delta\eta}{\sigma^4\sqrt{C}} + \cdots\right) . \tag{A.24}$$

Finally, from (A.9) and (A.11),

$$\begin{aligned} \frac{\left[(z^*)^{d_s} - 1\right]}{(z^* - 1)} &= \sum_{n=0}^{d_s - 1} (z^*)^n \\ &\sim \sum_{n=0}^{d_s - 1}\left(1 + \frac{2n\delta}{\sigma^2\sqrt{C}} + \cdots\right) \\ &\sim d_s + \frac{\delta}{\sigma^2\sqrt{C}} d_s(d_s - 1) + \cdots . \end{aligned} \tag{A.25}$$

From (A.8), (A.20), and (A.23)–(A.25), we obtain the approximation (2.5).

Appendix B. We here establish that $g(-\infty) = \infty$, $g(\infty) = 0$, and $g'(x) < 0$ for $x < \infty$, where $g(x)$ is given by (2.14), and $\mathrm{Erfc}\,(x)$ by (2.4). First (see [8]), $\mathrm{Erfc}\,(-\infty) = 2$, so that $g(-\infty) = \infty$, and

$$\mathrm{Erfc}\,(x) \sim \frac{e^{-x^2}}{\sqrt{\pi}x}\left[1 - \frac{1}{2x^2} + \frac{3}{4x^4} + O\left(\frac{1}{x^6}\right)\right], \quad x \gg 1. \tag{B.1}$$

Hence $g(\infty) = 0$. Next, we define

$$\omega(x) = \frac{\sqrt{\pi}}{2} e^{x^2}\,\mathrm{Erfc}\,(x) = e^{x^2}\int_x^\infty e^{-u^2}du = \int_0^\infty e^{-v^2}e^{-2xv}dv. \tag{B.2}$$

Then, $\omega(x) > 0$ for $x < \infty$, and

$$g(x) = \frac{1}{2\omega(x)} - x, \qquad g'(x) = -\frac{\omega'(x)}{2[\omega(x)]^2} - 1. \tag{B.3}$$

From (B.2) we obtain

$$[\omega(x)]^2 = \int_0^\infty \int_0^\infty e^{-(v^2+w^2)}\, e^{-2x(v+w)}\, dv\, dw. \tag{B.4}$$

If we make the transformation of variables

$$v = \frac{1}{\sqrt{2}}(\eta - \zeta), \qquad w = \frac{1}{\sqrt{2}}(\eta + \zeta), \tag{B.5}$$

then

$$[\omega(x)]^2 = \int_0^\infty \int_{-\eta}^{\eta} e^{-(\eta^2+\zeta^2)}\, e^{-2\sqrt{2}\,x\eta}\, d\zeta\, d\eta. \tag{B.6}$$

Also, from (B.2),

$$\frac{1}{2}\omega'(x) = -\int_0^\infty ve^{-v^2}\, e^{-2xv}dv = -2\int_0^\infty \eta e^{-2\eta^2}\, e^{-2\sqrt{2}\,x\eta}\, d\eta. \tag{B.7}$$

Hence,

$$\frac{1}{2}\omega'(x) + [\omega(x)]^2 = \int_0^\infty e^{-\eta^2}\, e^{-2\sqrt{2}\,x\eta}\, \psi(\eta)d\eta, \tag{B.8}$$

where

$$\psi(\eta) = \int_{-\eta}^{\eta} e^{-\zeta^2}d\zeta - 2\eta\, e^{-\eta^2} > 0, \quad \eta > 0\,. \tag{B.9}$$

It follows, from (B.3), (B.8), and (B.9), that $g'(x) < 0$ for $x < \infty$.

Appendix C. Here we prove Lemma 3.1. We consider $-\infty < x < \infty$. Then, from (3.3) and (B.2),

$$f(x) = \frac{1}{2\omega(x)}, \qquad f'(x) = -\frac{\omega'(x)}{2[\omega(x)]^2} > 0. \tag{C.1}$$

Next,

$$f''(x) = \frac{[\omega'(x)]^2}{[\omega(x)]^3} - \frac{\omega''(x)}{2[\omega(x)]^2}\,. \tag{C.2}$$

But, from (B.7),

$$\frac{1}{2}\omega''(x) = 2\int_0^\infty v^2\, e^{-v^2}\, e^{-2xv}\, dv. \tag{C.3}$$

Hence, from (B.2) and (B.7),

$$\begin{aligned}[\omega'(x)]^2 - \frac{1}{2}\omega(x)\omega''(x) \\ = \int_0^\infty \int_0^\infty (4vw - v^2 - w^2) e^{-(v^2+w^2)}\, e^{-2x(v+w)}\, dv\, dw,\end{aligned} \tag{C.4}$$

where we have expressed $\omega(x)\omega''(x)$ in a symmetric form.

If we make the transformation of variables (B.5), then

$$[\omega'(x)]^2 - \frac{1}{2}\omega(x)\omega''(x) = \int_0^\infty e^{-\eta^2}\, e^{-2\sqrt{2}\,x\eta}\, \chi(\eta)\, d\eta, \tag{C.5}$$

where

$$\chi(\eta) = \int_{-\eta}^{\eta} (\eta^2 - 3\zeta^2)\, e^{-\zeta^2}\, d\zeta. \tag{C.6}$$

Hence, from (B.9),

$$\chi'(\eta) = 2\eta\, \psi(\eta) > 0, \quad \eta > 0. \tag{C.7}$$

Since $\chi(0) = 0$, it follows that $\chi(\eta) > 0$ for $\eta > 0$. Hence, from (C.2) and (C.5), since $\omega(x) > 0$, we have established that $f''(x) > 0$ for $-\infty < x < \infty$, and the proof of the lemma is complete.

Appendix D. We here establish the asymptotic approximation in (4.5). Let

$$G(\mathbf{d}, \boldsymbol{\nu}, C - n) = \frac{1}{2\pi i} \int_{|z|<1} \frac{z^{n-1} e^{Cf(z)}}{(1-z)} dz \tag{D.1}$$

for integer values of n, where the integral is taken in a counterclockwise direction around a circle of radius less than 1, and, from (A.1) and (A.2),

$$Cf(z) = \sum_{s=1}^{S} \nu_s (z^{d_s} - 1) - C \log z. \tag{D.2}$$

Then (see [14]),

$$1 - \mathbb{L}_s(\mathbf{d}, \boldsymbol{\nu}, C) = \frac{1}{2\pi i G(\mathbf{d}, \boldsymbol{\nu}, C)} \int_{|z|<1} \frac{z^{d_s - 1} e^{Cf(z)}}{(1-z)} dz. \tag{D.3}$$

From (D.2) it follows that

$$\sum_{s=1}^{S} d_s \nu_s z^{d_s} = C[zf'(z) + 1]. \tag{D.4}$$

Hence, since $Y_s = \nu_s[1 - \mathbb{L}_s(\mathbf{d}, \boldsymbol{\nu}, C)]$,

$$C - \sum_{s=1}^{S} d_s Y_s = -\frac{C}{2\pi i G(\mathbf{d}, \boldsymbol{\nu}, C)} \int_{|z|<1} \frac{f'(z)}{(1-z)} e^{Cf(z)} dz. \tag{D.5}$$

However,

$$\frac{d}{dz}\left[\frac{e^{Cf(z)}}{(1-z)}\right] = \left[\frac{Cf'(z)}{(1-z)} + \frac{1}{(1-z)^2}\right] e^{Cf(z)}. \tag{D.6}$$

It follows that

$$C - \sum_{s=1}^{S} d_s Y_s = \frac{1}{2\pi i G(\mathbf{d}, \boldsymbol{\nu}, C)} \int_{|z|<1} \frac{e^{Cf(z)}}{(1-z)^2} dz. \tag{D.7}$$

From (4.2) and (D.4), $f'(z^*) = 0$. Moreover (see [10]), $|z| = z^*$ is a saddle-point contour, and if $h(z)$ is analytic in a domain containing $|z| = z^*$, then

$$\frac{1}{2\pi i} \int_{|z|=z^*} h(z) e^{Cf(z)} dz = \frac{e^{Cf(z^*)}}{\sqrt{2\pi C f''(z^*)}} \left[h(z^*) + O\left(\frac{1}{C}\right)\right]. \tag{D.8}$$

If we deform the contour of integration to $|z| = z^*$ and set $n = 0$ in (D.1), we obtain the asymptotic approximations

$$G(\mathbf{d}, \boldsymbol{\nu}, C) \sim \frac{e^{Cf(z^*)}}{\sqrt{2\pi C f''(z^*)}\ z^*(1 - z^*)} \tag{D.9}$$

and

$$\frac{1}{2\pi i} \int_{|z|<1} \frac{e^{Cf(z)}}{(1-z)^2} dz \sim \frac{e^{Cf(z^*)}}{\sqrt{2\pi C f''(z^*)}\ (1 - z^*)^2}. \tag{D.10}$$

The result in (4.5) follows from (D.7), (D.9), and (D.10).

Acknowledgments. The authors are indebted to Debasis Mitra for posing the general problem and suggesting this investigation. They are also grateful to the referees for suggestions that improved the presentation.

REFERENCES

[1] E. Bouillet, D. Mitra, and K. G. Ramakrishnan, *Design-assisted, real time, measurement-based network controls for management of service level agreements,* in Proceedings of the EURANDOM Workshop on "Stochastics of Integrated-Services Communications Networks," Eindhoven, The Netherlands, 1999.

[2] E. Bouillet, D. Mitra, and K. G. Ramakrishnan, *The structure and management of service level agreements in networks,* IEEE J. Sel. Areas Commun., 20 (2002), pp. 691–699.

[3] S. P. Evans, *Optimal bandwidth management and capacity provision in a broadband network using virtual paths,* Perform. Eval., 13 (1991), pp. 27–43.

[4] J. S. Kaufman, *Blocking in a shared resource environment,* IEEE Trans. Commun., COM-29 (1981), pp. 1474–1481.

[5] F. P. Kelly, *Reversibility and Stochastic Networks*, John Wiley, New York, 1980.

[6] F. P. Kelly, *Loss networks,* Ann. Appl. Probab., 1 (1991), pp. 319–378.

[7] S. S. Lam, *Queueing networks with population size constraints,* IBM J. Res. Develop., 21 (1977), pp. 370–378.

[8] W. Magnus, F. Oberhettinger, and R. P. Soni, *Formulas and Theorems for the Special Functions of Mathematical Physics*, Springer-Verlag, New York, 1966.

[9] D. Mitra, *Personal communication.*

[10] D. Mitra and J. A. Morrison, *Erlang capacity and uniform approximations for shared unbuffered resources,* IEEE/ACM Trans. Networking, 2 (1994), pp. 558–570.

[11] D. Mitra, J. A. Morrison, and K. G. Ramakrishnan, *ATM network design and optimization: A multirate loss network framework,* IEEE/ACM Trans. Networking, 4 (1996), pp. 531–543.

[12] D. Mitra, J. A. Morrison, and K. G. Ramakrishnan, *TALISMAN: An integrated set of tools for ATM network design and optimization,* in Proceedings of the 7th International Network Planning Symposium (NETWORKS '96), Sydney, Australia, Telstra Press, Sydney, pp. 483–488.

[13] D. Mitra, J. A. Morrison, and K. G. Ramakrishnan, *Optimization and design of network routing using refined asymptotic approximations,* Performance Evaluation, 36–37 (1999), pp. 267–288.

[14] J. A. Morrison, K. G. Ramakrishnan, and D. Mitra, *Refined asymptotic approximations to loss probabilities and their sensitivities in shared unbuffered resources,* SIAM J. Appl. Math., 59 (1998), pp. 494–513.

[15] M. I. Reiman, *Some allocation problems for critically loaded loss systems with independent links,* in Proceedings of the 14th IFIP WG7.3 International Symposium on Computer Performance Modelling, Measurement, and Evaluation (PERFORMANCE '90), Edinburgh, Scotland, P. J. B. King, I. Mitrani, and R. J. Pooley, eds., Elsevier, North-Holland, New York, Amsterdam, 1990, pp. 145–158.

[16] M. I. Reiman, *A critically loaded multiclass Erlang loss system,* Queueing Systems, 9 (1991), pp. 65–82.

[17] M. Schwartz, *Broadband Integrated Networks*, Prentice–Hall, Englewood Cliffs, NJ, 1996.

SIAM J. APPL. MATH.
Vol. 63, No. 1, pp. 241–258

DERIVATION OF A CONTINUUM MODEL FOR EPITAXIAL GROWTH WITH ELASTICITY ON VICINAL SURFACE*

YANG XIANG†

Abstract. In heteroepitaxial growth, the mismatch between the lattice constants in the film and the substrate causes misfit strain in the film, making a flat surface unstable to small perturbations. This morphological instability is called Asaro–Tiller–Grinfeld (ATG) instability, which can drive the film to self-organize into nanostructures such as quantum wires or quantum dots. In practice, most devices are fabricated on vicinal surfaces which consist of steps and terraces. In this case, the misfit strain causes step bunching, and traditional continuum models for the ATG instability do not apply directly. In this paper, we derive a continuum model for step bunching by taking the continuum limit of the discrete models proposed by Duport, Politi, andVillain [*J. Phys.* I, 5 (1995), pp. 1317–1350] and Tersoff et al. [*Phys. Rev. Lett.*, 75 (1995), pp. 2730–2733].

Key words. epitaxial growth, step bunching, continuum model, elasticity

AMS subject classifications. 35Q72, 74A50, 74A10

PII. S003613990139828X

1. Introduction.

Nanostructures such as quantum wires and quantum dots exhibit novel electronic and optical properties and have important potential applications in semiconductor technology. How to fabricate them efficiently has raised intense interest recently. One promising way is to employ the self-organization process during heteroepitaxial growth of thin films where they are under stress. The mismatch between the lattice constants in the film and in the substrate causes misfit strain and stress in the film, driving the self-organization of surface morphology (see, e.g., [44, 46]). Therefore understanding the mechanism of misfit related self-organization is an important step to make this technology a reality.

The stress-driven morphological instability was first studied by Asaro and Tiller [1] and later independently by Grinfeld [17, 18] and Srolovitz [43]. It is called Asaro–Tiller–Grinfeld (ATG) instability or Grinfeld instability. These authors studied the linear instability of a planar surface of a stressed solid to small perturbations and found that the planar surface is unstable for wavenumbers less than a critical value. This instability is manifested by a mass transport via surface diffusion. The stress in the solid is a destabilizing factor while the surface energy is a stabilizing one. This linear instability was also studied by Gao [15], Spencer, Voorhees, and Davis [42], Freund and Jonsdottir [14], Grilhe [16], and others.

The nonlinear evolution of the stress-driven instability for thick films will result in the formation of cusps. It was studied by Yang and Srolovitz [54], Chiu and Gao [5], Spencer and Meiron [40], and Kassner and Misbah [20]. If the films are thin and wet the substrates, Stranski–Krastanow wetting islands will form. The steady states of island shapes were studied by Spencer and Tersoff [41], Kukta and Freund [23], Spencer [39], Rudin and Spencer [33], and Shanahan and Spencer [36], and others. The nonlinear evolution of the surfaces of thin films and the formation of islands were

*Received by the editors November 14, 2001; accepted for publication (in revised form) March 29, 2002; published electronically September 5, 2002. This work was supported by NSF and DARPA through the VIP program.

http://www.siam.org/journals/siap/63-1/39828.html

†Princeton Materials Institute, Princeton University, Princeton, NJ 08544 (yxiang@princeton.edu).

studied by Chiu and Gao [6], Zhang and Bower [55].

These authors treated the surface as a continuum, neglecting the presence of steps. This can be true only at relatively high temperature above the roughening transition, when the surface can change continuously. The normal temperature for epitaxial growth is below the roughening transition, when the surface will consist of steps and terraces (see, e.g., [30]). In this case, the surface cannot continuously change, and it has been shown that there is an activation barrier for the nucleation of steps [48, 53]. Therefore the continuum theories mentioned above do not apply directly. In practice, most semiconductor devices are fabricated on vicinal surfaces. Such surfaces are cut at a small angle to the atomic planes, creating a succession of terraces separated by atomic-height steps. The self-organization driven by misfit elasticity is achieved by step bunching [26, 28]. These bunches have uniform size and spacing, and they are much straighter than single steps, which tend to meander (Bales–Zangwill instability [3]) due to the Schwoebel barrier [35]. Therefore they can serve as superior templates for growth of quantum wires and nucleation of clusters [32, 45].

The understanding of step bunching induced by misfit strain is not as complete as that of the traditional ATG instability for a continuous surface. One important model was proposed by Tersoff et al. [49] and Liu, Tersoff, and Lagally [24] describing the dynamics of each step based on Burton, Cabrera, and Frank (BCF) theory [4]. In their model, the elastic effect of a step on a thick film is modeled by a force monopole caused by misfit stress in the bulk and a force dipole caused by the step [47]. The force monopole causes attractive interaction between successive steps, which destabilizes a uniform step train. The force dipole causes repulsive interaction between successive steps, which stabilizes a uniform step train. They analyzed the linear instability toward step bunching from a uniform step train with small perturbations and showed that it evolves by progressive coalescence of step bunches [49]. They also studied the kinetic debunching effect and demonstrated numerically how to control the size of bunches [24]. Another important model was proposed by Duport, Nozieres, and Villain [8], Duport, Politi, and Villain [9], and Duport [7]. Besides the two elastic effects between the steps considered by Tersoff et al., they also considered the elastic interaction between adatoms and steps and the Schwoebel barrier. The elastic interaction between adatoms and steps is stabilizing or destabilizing depending on the sign of the misfit. The Schwoebel barrier is always stabilizing. Additional work was done on the shapes of the islands which consist of steps. Duport, Priester, and Villain [10] computed the equilibrium shapes of pyramid-like islands. Kaganer and Ploog [19] studied the two-dimensional island shapes and growth kinetics of step bunches by modifying Tersoff et al.'s model. All these models are two-dimensional based on the observation that the steps are very straight in bunches. Kukta and Bhattacharya [22] proposed a three-dimensional model for step-flow-mediated crystal growth under stress. As far as we know, there is no continuum model for step bunching induced by misfit strain.

There are three different levels of models for epitaxial growth: kinetic Monte Carlo (e.g., [38]), BCF theory [4], and continuum theories, with length scales ranging from atomic size, to terrace width, to mound size ([13], see also [11]). In most cases, a continuum model is desired when we are interested only in the shapes of step bunches or islands. The existing continuum models for the traditional ATG instability follow Mullins' chemical potential argument [27]. They do not incorporate the atomic structure of the underlying crystal which plays an important role in epitaxial growth.

In this paper, we derive a continuum model for the elastically driving step bunch-

ing by taking the continuum limit from the discrete models of Duport, Politi, and Villian [9] and Tersoff et al. [49]. The underlying atomic features of epitaxial growth reflected in the discrete models are kept in our model. This method of deriving continuum models from small scale models has been used in the literature to obtain continuum models for epitaxial growth without elasticity [50, 2, 31, 21, 29] (from BCF models), [34, 12, 25] (from BCF models coupled with adatom density) and [51] (from kinetic Monte Carlo models).

The rest of the paper is organized as follows. In section 2, we review the concept of epitaxial growth on vicinal surfaces and the BCF model, and we present the discrete models of Duport, Politi, and Villian [9] and Tersoff et al. [49]. In section 3, we derive our continuum model from these discrete models. In section 4, we summarize our results.

2. Epitaxial growth on vicinal surface and BCF-like models. In this section, we review the concept of epitaxial growth on vicinal surfaces and BCF theory, and we present the discrete models for elastically driving step bunching.

Epitaxial growth is the growth of crystalline film on a crystalline substrate following the same structure as the substrate. A vicinal surface consists of a succession of terraces separated by atomic-height steps, and the angle between the surface and the crystallographic plane is small. According to the BCF theory [4], the adatoms diffuse on the terraces until they meet the steps and get incorporated into the steps. The surface then grows (see Figure 2.1). The best way to grow a good crystal is to grow it on an infinite vicinal surface with parallel, equidistant steps (uniform step train) [30].

Steps and adatoms can interact elastically. The elastic interaction may be due to different kinds of mechanisms. One mechanism is the broken bond mechanism, which originates from the force dipole exerted by adatoms or steps due to the broken bonds of the adatoms or along the steps. The other mechanism is the misfit mechanism, which originates from the misfit between the lattice constants of the film and the substrate. The misfit strain and stress exist in the bulk of the film. If the modulation of the surface is small, the effect of this mechanism is equivalent to a surface stress

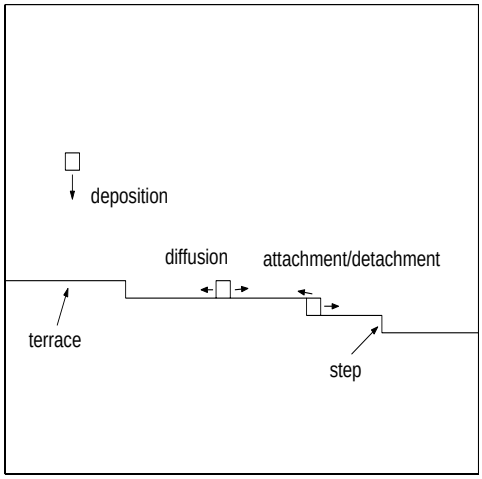


FIG. 2.1. *Schematic picture for the BCF theory.*

acting on a flat surface proportional to the modulation of the surface and the misfit. In the case of isotropic misfit and Hooke's law, the surface stress due to the misfit is given by

$$\eta_{xx} = \eta_{yy} = \frac{\delta a}{a}\frac{E}{1-\nu}\delta h, \tag{2.1}$$

where a is the lattice constant in the substrate, $a + \delta a$ is the lattice constant in the film, E is the Young modulus, ν is the Poisson ratio, and δh is the modulation of the surface.

The elastic interactions between adatoms and steps affect the diffusion of the adatoms on terraces. The elastic interactions between steps modify the equilibrium adatom density on steps and therefore modify the incorporating process of adatoms into the steps if we assume the adatoms incorporated into the steps can detach from them.

As far as diffusion alone is concerned, the adatoms prefer to go to the upper step rather than to the lower step, since an energy barrier exists near the lower step; this is called the Schwoebel barrier [35].

Now we present the model given by Duport, Politi, and Villain in [9] for a 1+1-dimensional vicinal surface based on the BCF theory. The 1+1-dimensional model assumes that the steps are straight and parallel, and the diffusion on the terrace is uniform in the direction parallel to the steps. The vicinal surface is assumed to be infinite and monotonic in the direction perpendicular to the steps. Without loss of generality, it is assumed that the terrace on the left of a step is higher than the terrace on the right. Letting $\{x_n\}$ with $\cdots < x_{n-1} < x_n < x_{n+1} < \cdots$ be the step train, the equations describing the adatom diffusion and step motion can be written as

$$\begin{cases} \dfrac{\partial \rho_n}{\partial t} = D\dfrac{\partial}{\partial x}\left(\dfrac{\partial \rho_n}{\partial x} + \dfrac{\rho_n}{k_B T}\dfrac{\partial U}{\partial x}\right) + F, \quad x_n < x < x_{n+1}, \\ D\left(\dfrac{\partial \rho_n}{\partial x} + \dfrac{\rho_n}{k_B T}\dfrac{\partial U}{\partial x}\right) = k^+(\rho_n - \rho_n^0), \quad x = x_n, \\ D\left(\dfrac{\partial \rho_n}{\partial x} + \dfrac{\rho_n}{k_B T}\dfrac{\partial U}{\partial x}\right) = -k^-(\rho_n - \rho_{n+1}^0), \quad x = x_{n+1}, \\ \dfrac{dx_n(t)}{dt} = a^2[k^+(\rho_n|_{x=x_n^+} - \rho_n^0) + k^-(\rho_{n-1}|_{x=x_n^-} - \rho_n^0)], \end{cases} \tag{2.2}$$

where ρ_n is the adatom density on the terrace between the nth step and the $(n+1)$st step, F is the deposition flux, D is the diffusion constant on terrace, k_B is the Boltzmann constant, T is the temperature, and k^+ and k^- are the hopping rates of an adatom to the upward step and downward step, respectively; the Schwoebel effect stipulates

$$k^+ \geq k^-. \tag{2.3}$$

ρ_n^0 is the equilibrium adatom density on the nth step; it is equal to the equilibrium adatom density on a step in the absence of elastic interactions ρ_0 with a local correction due to elasticity

$$\rho_n^0 = \rho_0 e^{-\frac{1}{k_B T}(U(x_n) - f_n)}. \tag{2.4}$$

The function $U(x)$ is the elastic energy due to the interaction between an adatom and the steps

$$U(x) = -\sum_{m=-\infty}^{+\infty} \frac{\alpha_0}{x_m - x} \tag{2.5}$$

and f_n describes the elastic interaction between the nth step and all other steps

$$f_n = -\sum_{m\neq n}\left(\frac{\alpha_1}{x_m - x_n} - \frac{\alpha_2}{(x_m - x_n)^3}\right). \tag{2.6}$$

The constants $\alpha_1, \alpha_2 > 0$, α_0 may be either positive or negative, and

$$\alpha_0 = \frac{2}{\pi}(1+\nu)\frac{\delta a}{a}\mu_a a^3, \tag{2.7}$$

$$\alpha_1 = \frac{2E}{\pi}\frac{1+\nu}{1-\nu}\left(\frac{\delta a}{a}\right)^2 a^4, \tag{2.8}$$

$$\alpha_2 = \frac{4}{\pi E}(1-\nu^2)\mu_s^2 a^4, \tag{2.9}$$

where $a^2\mu_a$ is the force dipole moment due to the broken bond mechanism for an adatom, $a\mu_s$ is the force dipole moment due to the broken bond mechanism along a step.

The first equation in the system (2.2) describes the deposition and diffusion processes of adatoms on a terrace. The second and third equations describe the incorporating process of adatoms into the steps, which serve as the boundary conditions of the diffusion problem. The fourth equation gives the velocity of the steps.

The function $U(x)$ describes the elastic interaction between one adatom and the steps. In this expression, the adatom has the broken bond effect and the steps have the misfit effect, which is dominant among all possible combinations. The sign of α_0 can be either positive or negative depending on the sign of $\mu_a\delta a$. The case of $\mu_a\delta a > 0$ means that the elastic interaction between an adatom with the broken bond effect and a step with the misfit effect is repulsive for upper steps and is attractive for lower steps. Thus the adatoms on a terrace prefer to go to the lower step than to the upper step. On the other hand, $\mu_a\delta a < 0$ means that the elastic interaction between an adatom with the broken bond effect and a step with the misfit effect is attractive for upper steps and is repulsive for lower steps. Thus the adatoms on a terrace prefer to go to the upper step than to the lower step.

The function f_n describes the elastic interaction between steps. The first term in it comes from the elastic interaction between steps in the step train due to the misfit mechanism. It is an attractive interaction. The second term comes from the elastic interaction between steps in the step train due to the broken bond mechanism. It is a repulsive interaction. Here we do not consider the interactions between a step with the broken bond effect and another step with the misfit effect, since they cancel. In fact, consider two successive steps, one is higher than the other. Due to the misfit mechanism (2.1), the higher step generates a surface stress with a sign the same as δa and the lower step generates a surface stress with a sign the same as $-\delta a$. However, due to the broken bond mechanism, the two steps generate force dipole moments with the same sign. Therefore the elastic interactions between one step with the broken bond mechanism and another step with the misfit mechanism cancel and we do not need to take them into consideration. In Duport, Politi, and Villain's paper [9], they did not notice this and they omitted these interactions by assuming either the broken bond mechanism or the misfit mechanism is dominant.

Now we solve the system (2.2). Usually in epitaxial growth, the deposition process is much slower than the diffusion process, which means that the step velocity is very small compared with the adatom hopping velocity. Therefore the quasi-static approximation can be made:

$$\frac{\partial \rho_n}{\partial t} \approx 0. \tag{2.10}$$

Under this assumption, we can get an explicit solution for the step velocity:

$$\begin{aligned}
\frac{1}{a^2}\frac{dx_n(t)}{dt} &= F\frac{l_n + l_{n-1}}{2} \\
&+ \frac{\rho_n^- e^{\frac{U(x_{n+1})}{k_BT}} - \rho_n^+ e^{\frac{U(x_n)}{k_BT}}}{\frac{e^{\frac{U(x_n)}{k_BT}}}{k^+} + \frac{e^{\frac{U(x_{n+1})}{k_BT}}}{k^-} + \frac{1}{D}\int_{x_n}^{x_{n+1}} e^{\frac{U(y)}{k_BT}}dy} \\
&- \frac{\rho_{n-1}^- e^{\frac{U(x_n)}{k_BT}} - \rho_{n-1}^+ e^{\frac{U(x_{n-1})}{k_BT}}}{\frac{e^{\frac{U(x_{n-1})}{k_BT}}}{k^+} + \frac{e^{\frac{U(x_n)}{k_BT}}}{k^-} + \frac{1}{D}\int_{x_{n-1}}^{x_n} e^{\frac{U(y)}{k_BT}}dy} \\
&+ \frac{\frac{Fl_n}{2}\left(\frac{e^{\frac{U(x_{n+1})}{k_BT}}}{k^-} - \frac{e^{\frac{U(x_n)}{k_BT}}}{k^+}\right)}{\frac{e^{\frac{U(x_n)}{k_BT}}}{k^+} + \frac{e^{\frac{U(x_{n+1})}{k_BT}}}{k^-} + \frac{1}{D}\int_{x_n}^{x_{n+1}} e^{\frac{U(y)}{k_BT}}dy} \\
&- \frac{\frac{Fl_{n-1}}{2}\left(\frac{e^{\frac{U(x_n)}{k_BT}}}{k^-} - \frac{e^{\frac{U(x_{n-1})}{k_BT}}}{k^+}\right)}{\frac{e^{\frac{U(x_{n-1})}{k_BT}}}{k^+} + \frac{e^{\frac{U(x_n)}{k_BT}}}{k^-} + \frac{1}{D}\int_{x_{n-1}}^{x_n} e^{\frac{U(y)}{k_BT}}dy} \\
&+ \frac{\frac{F}{D}\int_{x_n}^{x_{n+1}}\left(y - \frac{x_{n+1}+x_n}{2}\right)e^{\frac{U(y)}{k_BT}}dy}{\frac{e^{\frac{U(x_n)}{k_BT}}}{k^+} + \frac{e^{\frac{U(x_{n+1})}{k_BT}}}{k^-} + \frac{1}{D}\int_{x_n}^{x_{n+1}} e^{\frac{U(y)}{k_BT}}dy} \\
&- \frac{\frac{F}{D}\int_{x_{n-1}}^{x_n}\left(y - \frac{x_n+x_{n-1}}{2}\right)e^{\frac{U(y)}{k_BT}}dy}{\frac{e^{\frac{U(x_{n-1})}{k_BT}}}{k^+} + \frac{e^{\frac{U(x_n)}{k_BT}}}{k^-} + \frac{1}{D}\int_{x_{n-1}}^{x_n} e^{\frac{U(y)}{k_BT}}dy},
\end{aligned} \tag{2.11}$$

where

$$l_n \equiv x_{n+1} - x_n \tag{2.12}$$

is the width of the terrace between the nth step and the $(n+1)$st step.

If it is further assumed that the elastic energies are very small compared to the thermal energy

$$f_n, U(x) << k_BT, \tag{2.13}$$

keeping the leading order terms of $1/k_BT$ for each effect, respectively, we can write the step velocity as

$$
\begin{aligned}
\frac{1}{a^2}\frac{dx_n(t)}{dt} &= F\frac{l_n+l_{n-1}}{2} + \frac{\rho_0}{k_BT}\cdot\frac{l_n}{\frac{1}{k^+}+\frac{1}{k^-}+\frac{l_n}{D}}\cdot\frac{f_{n+1}-f_n}{l_n} \\
&\quad - \frac{\rho_0}{k_BT}\cdot\frac{l_{n-1}}{\frac{1}{k^+}+\frac{1}{k^-}+\frac{l_{n-1}}{D}}\cdot\frac{f_n-f_{n-1}}{l_{n-1}} \\
&\quad + \frac{\frac{Fl_n}{2}\left(\frac{1}{k^-}-\frac{1}{k^+}\right)}{\frac{1}{k^+}+\frac{1}{k^-}+\frac{l_n}{D}} - \frac{\frac{Fl_{n-1}}{2}\left(\frac{1}{k^-}-\frac{1}{k^+}\right)}{\frac{1}{k^+}+\frac{1}{k^-}+\frac{l_{n-1}}{D}} \\
&\quad + \frac{\frac{F}{D}\int_{x_n}^{x_{n+1}}\left(y-\frac{x_{n+1}+x_n}{2}\right)\frac{U(y)}{k_BT}dy}{\frac{1}{k^+}+\frac{1}{k^-}+\frac{l_n}{D}} \\
&\quad - \frac{\frac{F}{D}\int_{x_{n-1}}^{x_n}\left(y-\frac{x_n+x_{n-1}}{2}\right)\frac{U(y)}{k_BT}dy}{\frac{1}{k^+}+\frac{1}{k^-}+\frac{l_{n-1}}{D}}.
\end{aligned}
\tag{2.14}
$$

This step motion equation is simpler than that derived in Duport, Politi, and Villain [9]. We keep only the leading order terms of $1/k_BT$ for each effect, respectively. (We neglect the $O(1/k_BT)$ corrections to the hopping rates $k^{\pm}$, while they kept all $O(1/k_BT)$ terms.)

If we consider only the attractive and repulsive elastic interactions between steps and neglect the elastic interaction between adatoms and steps, as well as the Schwoebel barrier, as was done in Tersoff et al. [49] and Liu, Tersoff, and Lagally [24], then we have

$$
\frac{1}{a^2}\frac{dx_n}{dt} = F\frac{l_n+l_{n+1}}{2} + \frac{\rho_0 D}{k_BT}\left(\frac{f_{n+1}-f_n}{l_n} - \frac{f_n-f_{n-1}}{l_{n-1}}\right). \tag{2.15}
$$

This corresponds to the case $U(y)\equiv 0$ and $k^+ = k^- = +\infty$ in (2.14).

3. Continuum equation. In this section, we derive our continuum model governing the elastically driving step bunching by taking the continuum limit from the modified Duport et al.'s discrete model (2.14), of which Tersoff et al.'s discrete model (2.15) is a special case.

Assume the lattice constant a is very small compared with the length scale in which we are interested; then the surface can be considered as a continuous function $h(x)$. As in the discrete models, we assume that the overall slope of the surface is negative. We also assume that the surface has only a bounded deviation from a flat surface representing a uniform step train, so that those summations of the $1/r$ elastic interactions in (2.5) and (2.6) are defined in the sense of principal value. More precisely, we assume that $h(x)\in C^4(\mathbf{R})$ satisfies the following:

$$
\begin{cases}
h_x, h_{xx}, h_{xxx}, \text{ and } h_{xxxx} \text{ are bounded,} \\
h_x < 0, \\
h(x) = -Ax + \text{a bounded smooth function,}
\end{cases}
\tag{3.1}
$$

where $A > 0$ is a constant.

Due to the monotonicity, we can also consider x as a function of h. Due to the asymptotic property of $h(x)$, we know that $x(h)$ is defined for all h and therefore has a similar asymptotic property. More precisely, $x(h)$ satisfies the following:

$$\begin{cases} x(h) \text{ has up to fourth order bounded derivatives,} \\ x_h < 0, \\ x(h) = -\dfrac{h}{A} + \text{a bounded smooth function.} \end{cases} \tag{3.2}$$

The step motion equation (2.14) can be considered as a numerical scheme for a differential equation of $x(h,t)$ with the grid constant a. We have the following relations:

$$x_n = x(h_n, t) \tag{3.3}$$

$$h_{n+1} - h_n = -a, \tag{3.4}$$

where x_n is the position of the nth step, h_n is the height of the terrace between the $(n-1)$st step and the nth step. See Figure 3.1.

We obtain our continuum equation by letting $a \to 0$ in the step motion equation (2.14). We will first compute the continuum limit of each summation representing each elastic effect and then derive the continuum equation.

3.1. Continuum limit of the $1/r^3$ interaction between steps. In this subsection, we compute the continuum limit of $\sum_{m\neq n} \frac{a^2}{(x_m - x_n)^3}$ as $a \to 0$.

First, we have

$$\begin{aligned} & \sum_{m\neq n} \frac{a^2}{(x_m - x_n)^3} \\ = & \sum_{m=1}^{+\infty} \left(\frac{a^2}{(x_{n+m} - x_n)^3} - \frac{a^2}{(x_n - x_{n-m})^3} \right) \\ = & \sum_{m=1}^{+\infty} \frac{\dfrac{2x_n - x_{n+m} - x_{n-m}}{a^2}}{\left(\dfrac{x_{n+m} - x_n}{a}\right)^3 \left(\dfrac{x_n - x_{n-m}}{a}\right)^3} \\ & \cdot \left(\left(\frac{x_n - x_{n-m}}{a}\right)^2 + \frac{x_n - x_{n-m}}{a} \frac{x_{n+m} - x_n}{a} + \left(\frac{x_{n+m} - x_n}{a}\right)^2 \right) \\ \equiv & \sum_{m=1}^{+\infty} a_m. \end{aligned} \tag{3.5}$$

For each a_m, we can compute its limit

$$\lim_{a\to 0} a_m = -\frac{3}{m^2} \frac{x_{hh}}{x_h^4} = \frac{3}{m^2} h_x h_{xx}. \tag{3.6}$$

By the assumption of $x(h)$ (3.2)

$$a_m = O\left(\frac{1}{m^2}\right). \tag{3.7}$$

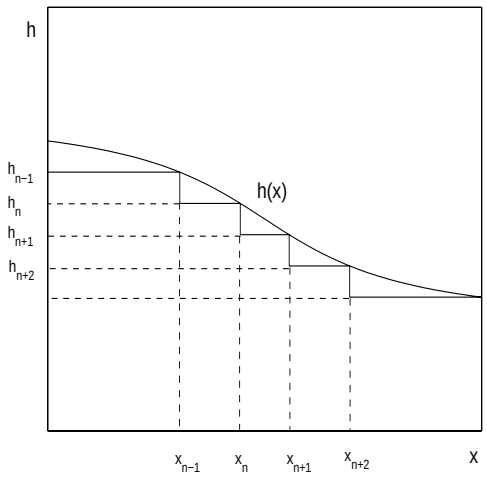


FIG. 3.1. *Continuous surface profile $h(x)$ and positions of steps.*

Therefore the summation converges absolutely, and we can change the order of the summation and the limit to get

$$\lim_{a\to 0}\sum_{m\neq n}\frac{a^2}{(x_m-x_n)^3}=\sum_{m=1}^{+\infty}h_xh_{xx}\frac{3}{m^2}=\frac{\pi^2}{2}h_xh_{xx}. \tag{3.8}$$

3.2. Continuum limit of the $1/r$ interaction between steps. In this subsection, we compute the continuum limit of $\sum_{m\neq n}\frac{a}{x_m-x_n}$ as $a\to 0$.

We use the following theorem obtained by Sidi and Israeli in [37]. They derived it to estimate the error of computing a Cauchy principal value integral using the trapezoidal rule.

THEOREM [37]. *Let $g(x)$ be $2N$ times differentiable on $[a,b]$. The interval $[a,b]$ is divided into n small intervals with $\Delta x=(b-a)/n$, $x_j=a+(j-1)*\Delta x$, $j=0,1,\ldots,n+1$. Let $G(x)=g(x)/(x-t)$, where $t=x_{j_0}$ for some $j_0\neq 0,n+1$. Then as $\Delta x\to 0$,*

$$\begin{aligned}\int_a^b G(x)dx &= \Delta x\left(\frac{1}{2}G(a)+\sum_{x_j\neq t,1\le j\le n}G(x_j)+\frac{1}{2}G(b)\right)+\Delta x g'(t)\\ &\quad+\sum_{\mu=1}^{N-1}\frac{B_{2\mu}}{(2\mu)!}(G^{(2\mu-1)}(a)-G^{(2\mu-1)}(b))\Delta x^{2\mu}+R_{2N}[G;(a,b)],\end{aligned} \tag{3.9}$$

where the B_μ's are the Bernoulli numbers and

$$|R_{2N}[G;(a,b)]|\le M_{2N}\Delta x^{2N}\int_a^b\left|\frac{d^{2N}}{dx^{2N}}\left(\frac{g(x)-g(t)}{x-t}\right)\right|dx \tag{3.10}$$

for a constant M_{2N} not depending on $G(x)$ and $[a,b]$.

Now without loss of generality, assume $x(0)=x_n$ and consider the function $G(h)=1/(x(h)-x(0))$. Choosing the grid constant to be our lattice constant a and

$N = 1$, using the theorem in $(-\infty, +\infty)$, we have

$$\int_{-\infty}^{+\infty} \frac{dh}{x(h) - x(0)} = \sum_{m \neq n} \frac{a}{x_m - x_n} - \frac{a}{2} \frac{x_{hh}}{x_h^2} + O(a^2), \tag{3.11}$$

where the integral on the left is in the sense of the Cauchy principal value at 0 and ∞. It is defined by

$$\int_{-\infty}^{+\infty} \frac{dh}{x(h) - x(0)} \equiv \lim_{H \to +\infty,\ \epsilon \to 0} \left(\int_{-H}^{-\epsilon} \frac{dh}{x(h) - x(0)} + \int_{\epsilon}^{H} \frac{dh}{x(h) - x(0)} \right). \tag{3.12}$$

Here we have used the assumption of $x(h)$ (3.2) to guarantee the existence of the Cauchy principal values at 0 and ∞ and the convergence of the integral in the error term (3.10).

Therefore we have

$$\sum_{m \neq n} \frac{a}{x_m - x_n} = \int_{-\infty}^{+\infty} \frac{dh}{x(h) - x(0)} + \frac{a}{2} \frac{x_{hh}}{x_h^2} + O(a^2). \tag{3.13}$$

Now we show that change of variable from h to x does not affect the Cauchy principal value integral. In fact,

$$\begin{aligned} \int_{\epsilon}^{H} \frac{dh}{x(h) - x(0)} &= \int_{x(\epsilon)}^{x(H)} \frac{h_x}{x - x_n} dx \\ &= -\left(\int_{x(H)}^{x_n - \frac{H}{A}} + \int_{x_n - \frac{H}{A}}^{x_n + x_h(0)\epsilon} + \int_{x_n + x_h(0)\epsilon}^{x(\epsilon)} \right) \frac{h_x}{x - x_n} dx. \end{aligned} \tag{3.14}$$

Since

$$\int_{x_n + x_h(0)\epsilon}^{x(\epsilon)} \frac{h_x}{x - x_n} dx = \int_{x_n + x_h(0)\epsilon}^{x_n + x_h(0)\epsilon + O(\epsilon^2)} \frac{h_x}{x - x_n} dx = O(\epsilon), \tag{3.15}$$

$$\int_{x(H)}^{x_n - \frac{H}{A}} \frac{h_x}{x - x_n} dx = \left| x_n - x(H) - \frac{H}{A} \right| \cdot O\left(\frac{1}{H} \right) = O\left(\frac{1}{H} \right), \tag{3.16}$$

we have

$$\int_{\epsilon}^{H} \frac{dh}{x(h) - x(0)} = -\int_{x_n - \frac{H}{A}}^{x_n + x_h(0)\epsilon} \frac{h_x}{x - x_n} dx + O(\epsilon) + O\left(\frac{1}{H} \right). \tag{3.17}$$

Similarly, we have

$$\int_{-H}^{-\epsilon} \frac{dh}{x(h) - x(0)} = -\int_{x_n - x_h(0)\epsilon}^{x_n + \frac{H}{A}} \frac{h_x}{x - x_n} dx + O(\epsilon) + O\left(\frac{1}{H} \right). \tag{3.18}$$

Therefore a change of variable from h to x does not affect the Cauchy principal value integral. From (3.13) we have as $a \to 0$,

$$\sum_{m \neq n} \frac{a}{x_m - x_n} = -\int_{-\infty}^{+\infty} \frac{h_x}{x - x_n} dx - \frac{a}{2} \frac{h_{xx}}{h_x} + O(a^2). \tag{3.19}$$

The integral is in the sense of the Cauchy principal value at 0 and ∞.

We keep the $O(a)$ term here because it has the same order as the $1/r^3$ interaction.

Summarizing the results, we have shown that as $a \to 0$,

$$a^2 f_n \approx a\alpha_1 \int_{-\infty}^{+\infty} \frac{h_x}{x - x_n} dx + \frac{1}{2} a^2 \alpha_1 \frac{h_{xx}}{h_x} + \frac{\pi^2}{2} \alpha_2 h_x h_{xx}. \tag{3.20}$$

3.3. Continuum limit of the interaction between adatoms and steps. In this subsection, we compute the continuum limit of the interaction between the steps and the adatoms on the terrace between steps x_n and x_{n+1}:

$$\lim_{a\to 0}\frac{1}{a}\int_{x_n}^{x_{n+1}}\left(y-\frac{x_n+x_{n+1}}{2}\right)\sum_m\frac{1}{x_m-y}dy. \tag{3.21}$$

This integral has singularities at the endpoints x_n and x_{n+1}. Duport, Nozieres, and Villain [8], Duport, Politi, and Villain [9], and Duport [7] overcame this difficulty by computing the integral on the interval $[x_n+a, x_{n+1}-a]$. One reason for this truncation of the integration interval is that the nearest lattice site to the step is one lattice constant away from the step. Here we also use this approximation.

First, we compute

$$\int_{x_n}^{x_{n+1}}\left(y-\frac{x_n+x_{n+1}}{2}\right)\frac{1}{x_m-y}dy. \tag{3.22}$$

For $m\neq n, n+1$

$$\int_{x_n}^{x_{n+1}}\left(y-\frac{x_n+x_{n+1}}{2}\right)\frac{1}{x_m-y}dy=-l_n+\left(x_m-\frac{x_n+x_{n+1}}{2}\right)\log\frac{x_m-x_n}{x_m-x_{n+1}}. \tag{3.23}$$

For $m=n$

$$\int_{x_n+a}^{x_{n+1}}\left(y-\frac{x_n+x_{n+1}}{2}\right)\frac{1}{x_n-y}dy=-l_n+\frac{l_n}{2}\log\frac{l_n}{a}. \tag{3.24}$$

For $m=n+1$

$$\int_{x_n}^{x_{n+1}-a}\left(y-\frac{x_n+x_{n+1}}{2}\right)\frac{1}{x_{n+1}-y}dy=-l_n+\frac{l_n}{2}\log\frac{l_n}{a}. \tag{3.25}$$

Therefore we have

$$\begin{aligned}
&\frac{1}{a}\int_{x_n}^{x_{n+1}}\left(y-\frac{x_n+x_{n+1}}{2}\right)\sum_m\frac{1}{x_m-y}dy\\
=\;&\sum_{m=1}^{+\infty}\left(-\frac{l_n}{a}+\left(\sum_{k=1}^{m}\frac{l_{n+k}}{a}+\frac{l_n}{2a}\right)\log\frac{\displaystyle\sum_{k=1}^{m}\frac{l_{n+k}}{a}+\frac{l_n}{a}}{\displaystyle\sum_{k=1}^{m}\frac{l_{n+k}}{a}}\right)\\
&+\sum_{m=1}^{+\infty}\left(-\frac{l_n}{a}+\left(\sum_{k=1}^{m}\frac{l_{n-k}}{a}+\frac{l_n}{2a}\right)\log\frac{\displaystyle\sum_{k=1}^{m}\frac{l_{n-k}}{a}+\frac{l_n}{a}}{\displaystyle\sum_{k=1}^{m}\frac{l_{n-k}}{a}}\right)\\
&-2\frac{l_n}{a}+\frac{l_n}{a}\log\frac{l_n}{a}\\
\equiv\;&\sum_{m=1}^{+\infty}a_m+\sum_{m=1}^{+\infty}b_m-2\frac{l_n}{a}+\frac{l_n}{a}\log\frac{l_n}{a}.
\end{aligned} \tag{3.26}$$

Rewrite a_m as

$$a_m = -\frac{l_n}{a} + \frac{l_n}{2a} \cdot \frac{1}{\theta_m} \log \frac{1+\theta_m}{1-\theta_m}, \tag{3.27}$$

where

$$\theta_m \equiv \frac{\dfrac{l_n}{2a}}{\displaystyle\sum_{k=1}^{m} \frac{l_{n+k}}{a} + \frac{l_n}{2a}}. \tag{3.28}$$

Using the assumption of $x(h)$ (3.2), we can prove that

$$\theta_m = O\left(\frac{1}{m}\right) \tag{3.29}$$

and

$$a_m = O\left(\frac{1}{m^2}\right). \tag{3.30}$$

That means that $\sum_{m=1}^{+\infty} a_m$ is uniformly convergent with respect to a. Therefore

$$\lim_{a\to 0} \sum_{m=1}^{+\infty} a_m = \sum_{m=1}^{+\infty} \lim_{a\to 0} a_m = \sum_{m=1}^{+\infty} \left(1 - \left(m+\frac{1}{2}\right) \log \frac{m+1}{m}\right) \frac{1}{h_x(x_n)}. \tag{3.31}$$

Similarly, we can prove that

$$\lim_{a\to 0} \sum_{m=1}^{+\infty} b_m = \sum_{m=1}^{+\infty} \lim_{a\to 0} b_m = \sum_{m=1}^{+\infty} \left(1 - \left(m+\frac{1}{2}\right) \log \frac{m+1}{m}\right) \frac{1}{h_x(x_n)}. \tag{3.32}$$

It is easy to compute the continuum limit of the other terms in (3.26):

$$\lim_{a\to 0} \left(-2\frac{l_n}{a} + \frac{l_n}{a} \log \frac{l_n}{a}\right) = \frac{2 + \log |h_x(x_n)|}{h_x(x_n)}. \tag{3.33}$$

Using the relation

$$1 + \sum_{m=1}^{+\infty} \left(1 - \left(m+\frac{1}{2}\right) \log \frac{m+1}{m}\right) = \frac{1}{2} \log 2\pi, \tag{3.34}$$

which can be found in Duport, Politi, and Villain's paper [9], we have

$$\lim_{a\to 0} \frac{1}{a} \int_{x_n}^{x_{n+1}} \left(y - \frac{x_n + x_{n+1}}{2}\right) \sum_m \frac{1}{x_m - y} dy = \frac{1}{h_x(x_n)} \log(2\pi |h_x(x_n)|). \tag{3.35}$$

Therefore as $a \to 0$, we have

$$\frac{1}{a} \int_{x_n}^{x_{n+1}} \left(y - \frac{x_n + x_{n+1}}{2}\right) U(y) dy \approx -a\alpha_0 \frac{1}{h_x(x_n)} \log(2\pi |h_x(x_n)|). \tag{3.36}$$

3.4. Derivation of continuum equation. Now we have the continuum limit of f_n (3.20) and the continuum limit of the integral containing $U(x)$ (3.36) as $a \to 0$. To get our continuum limit equation, we approximate the finite differences in the discrete model (2.14) by derivatives:

$$l_n = x_{n+1} - x_n = -x_h a + O(a^2) = -\frac{a}{h_x} + O(a^2), \tag{3.37}$$

$$\begin{aligned} & l_n + l_{n-1} \\ = \;& x_{n+1} - x_{n-1} \\ = \;& -2x_h a - \tfrac{1}{3} x_{hhh} a^3 + O(a^5) \\ = \;& -\frac{2}{h_x} a - \frac{1}{6} \frac{\partial^2}{\partial x^2} \left(\frac{1}{h_x^2} \right) a^3 + O(a^5), \end{aligned} \tag{3.38}$$

$$\frac{G_{n+1} - G_n}{l_n} = \frac{G(x_n + l_n) - G(x_n)}{l_n} = G_x + O(l_n) = G_x + O(a), \tag{3.39}$$

and therefore

$$G_{n+1} - G_n = -a \frac{G_x}{h_x} + O(a^2), \tag{3.40}$$

where G is any smooth function of x and $G_n = G(x_n)$.

We keep only the leading order terms of a in each finite difference, respectively, to keep the main contribution from each effect, except for the term $F(l_n + l_{n-1})/2$. Its leading order term of a is the average growth rate of the surface due to the deposition flux, and we keep a higher order term which has the same order as the term of elastic interactions between steps.

We also use the relation

$$\frac{Dh}{Dt} = \frac{\partial h}{\partial t} + \frac{\partial h}{\partial x} \frac{dx}{dt} = 0, \tag{3.41}$$

where Dh/Dt is the material derivative. It means that adatoms can move only in the horizontal direction.

Now letting $a \to 0$, we get the continuum equation

$$\begin{aligned} h_t \;=\; & a^3 F \left(1 + \frac{a^2}{12} \frac{\partial^2}{\partial x^2} \left(\frac{1}{h_x^2} \right) \right) \\ & + \frac{a^2 \pi \alpha_1 \rho_0 D}{k_B T} \frac{\partial}{\partial x} \left(\frac{1}{1 - (L/a) h_x} \frac{\partial f}{\partial x} \right) \\ & + \frac{1}{2} a^3 F (l_- - l+) \frac{\partial}{\partial x} \left(\frac{1}{1 - (L/a) h_x} \right) \\ & + \frac{a^3 F \alpha_0}{k_B T} \frac{\partial}{\partial x} \left(\frac{\log(2\pi |h_x|)}{1 - (L/a) h_x} \right), \end{aligned} \tag{3.42}$$

where

$$f(x) = -\frac{1}{\pi} \int_{-\infty}^{+\infty} \frac{h_x(y)}{x - y} dy + \frac{a}{2\pi} \frac{h_{xx}}{h_x} + \frac{\pi l_e^2}{2a} h_x h_{xx} \tag{3.43}$$

and parameters

$$l_e = \sqrt{\frac{\alpha_2}{\alpha_1}}, \tag{3.44}$$

$$l_{\pm} = \frac{D}{k^{\pm}}, \tag{3.45}$$

$$L = l_+ + l_-. \tag{3.46}$$

The parameter l_e represents the equilibrium distance between two successive steps under the attractive misfit interaction and the repulsive broken bond interaction, l_+ and l_- represent the strength of the upward and downward step edge barriers, respectively, and

$$l_- \geq l_+ \tag{3.47}$$

due to the Schwoebel effect.

The first term in $f(x)$ is the Hilbert transform of h_x with a negative sign. The Hilbert transform of a function $u(x)$ is defined by

$$H(u) = \frac{1}{\pi} \int_{-\infty}^{+\infty} \frac{u(y)}{x-y} dy. \tag{3.48}$$

We can write our continuum equation in a more convenient form:

$$h_t = \bar{F} + \alpha \frac{\partial^2}{\partial x^2} \left(\frac{1}{h_x^2} \right) + \frac{\partial}{\partial x} \left[\frac{1}{1 - \bar{L} h_x} \left(\beta \frac{\partial f}{\partial x} + \lambda + \sigma \log(2\pi |h_x|) \right) \right], \tag{3.49}$$

$$f = -H(h_x) + \eta \left(\frac{1}{h_x} + \gamma h_x \right) h_{xx}, \tag{3.50}$$

where $H(h_x)$ is the Hilbert transform of h_x and

$$\bar{F} = a^3 F, \tag{3.51}$$

$$\alpha = \frac{a^5 F}{12}, \tag{3.52}$$

$$\bar{L} = \frac{L}{a}, \tag{3.53}$$

$$\beta = \frac{a^2 \pi \alpha_1 \rho_0 D}{k_B T}, \tag{3.54}$$

$$\lambda = \frac{1}{2} a^3 F (l_- - l_+), \tag{3.55}$$

$$\sigma = \frac{a^3 F \alpha_0}{k_B T}, \tag{3.56}$$

$$\eta = \frac{a}{2\pi}, \tag{3.57}$$

$$\gamma = \frac{\pi^2 l_e^2}{a^2}, \tag{3.58}$$

where all the constants are positive except σ, which may be either positive or negative.

The constant $\bar{F}$ and the $\frac{\partial^2}{\partial x^2}(\frac{1}{h_x^2})$ term are due to the deposition flux, where $\bar{F}$ is the average growth rate of the surface and the $\frac{\partial^2}{\partial x^2}(\frac{1}{h_x^2})$ term is the correction due to the local surface profile. The function f represents the elastic interactions between steps. The Hilbert transform term and the h_{xx}/h_x term come from the $1/r$ misfit elastic interaction. The $h_x h_{xx}$ term comes from the $1/r^3$ broken bond elastic interaction. The constant λ and the factor $1/(1-\bar{L}h_x)$ come from the step edge barriers; λ is positive due to the Schwoebel effect. The constant λ also depends on the deposition flux. The $\log(2\pi|h_x|)$ term is due to the elastic interaction between adatoms and steps. Its coefficient σ also depends on the deposition flux. Its sign can be positive or negative depending on the sign of α_0. Recall that $\alpha_0 > 0$ means the elastic interaction between an adatom and a step is repulsive for upper steps and is attractive for lower steps. Thus the adatoms on a terrace prefer to go to the lower step than to the upper step. On the other hand, $\alpha_0 < 0$ means the elastic interaction between an adatom and a step is attractive for upper steps and is repulsive for lower steps. Thus the adatoms on a terrace prefer to go to the upper step rather than to the lower step.

We can also write the continuum equation as

$$h_t = a^3 F - \frac{\partial}{\partial x} J(x,t), \tag{3.59}$$

where the surface flux $J(x,t)$ is

$$J(x,t) = -\alpha \frac{\partial}{\partial x}\left(\frac{1}{h_x^2}\right) + \frac{1}{1-\bar{L}h_x}\left(\beta \frac{\partial f}{\partial x} + \lambda + \sigma \log(2\pi|h_x|)\right). \tag{3.60}$$

If we neglect the Schwoebel barrier and the elastic interaction between adatoms and steps, the continuum equation becomes

$$h_t = \bar{F} + \alpha \frac{\partial^2}{\partial x^2}\left(\frac{1}{h_x^2}\right) + \beta \frac{\partial^2}{\partial x^2}\left[-H(h_x) + \eta\left(\frac{1}{h_x} + \gamma h_x\right) h_{xx}\right]. \tag{3.61}$$

It is the continuum limit of Tersoff et al.'s model (2.15).

If we do not take into consideration the deposition flux, in other words, we consider only the elastic interaction between steps, the continuum equation becomes

$$h_t = \frac{\partial^2}{\partial x^2}\left[-H(h_x) + \eta\left(\frac{1}{h_x} + \gamma h_x\right) h_{xx}\right]. \tag{3.62}$$

Here we have rescaled the time by β.

Equation (3.62) has a variational form. It can be written as

$$h_t = \mu_{xx}, \tag{3.63}$$

where the chemical potential μ is the variation of the total elastic energy

$$\mathcal{E} = \int \left(-\frac{1}{2}\tilde{h}H(h_x) + \eta|h_x|\log|h_x| + \frac{\eta\gamma}{6}|h_x|^3\right) dx. \tag{3.64}$$

Here $\tilde{h} = h - (-Ax)$ is the deviation from the reference planar surface (see assumption (3.1)).

Results on the linear instability, nonlinear evolution, and steady states using our continuum equation are presented in [52].

4. Summary. We have derived a continuum model governing the epitaxial growth with elasticity on a $1+1$-dimensional vicinal surface where the surface profile is monotonic. We obtained our model by taking the continuum limit from the discrete models of Duport, Politi, and Villain [9] and Tersoff et al. [49]. Compared with the existing continuum models for epitaxial growth on a vicinal surface, our model includes the effect of elasticity. Compared with the existing continuum models for surface morphology instability induced by elasticity, our model incorporates the atomic features of the stepped surface.

Acknowledgments. The author would like to thank Professor Weinan E of Princeton University and Professor Robert V. Kohn of New York University for helpful discussions.

REFERENCES

[1] R. J. Asaro and W. A. Tiller, *Interface morphology development during stress-corrosion cracking: Part* 1. *Via surface diffusion*, Metall. Trans., 3 (1972), pp. 1789–1796.
[2] C. Atkinson and P. Wilmott, *Interactions of continuous distributions of ledges*, Proc. Roy. Soc. London Ser. A, 446 (1994), pp. 277–287.
[3] G. S. Bales and A. Zangwill, *Morphological instability of a terrace edge during step-flow growth*, Phys. Rev. B, 41 (1990), pp. 5500–5508.
[4] W. K. Burton, N. Cabrera, and F. Frank, *The growth of crystals and the equilibrium structure of their surfaces*, Phil. Trans. Roy. Soc. London Ser. A, 243 (1951), pp. 299–358.
[5] C. H. Chiu and H. Gao, *Stress singularities along a cycloid rough surface*, Internat. J. Solids Structures, 30 (1993), pp. 2983–3012.
[6] C. H. Chiu and H. Gao, *A numerical study of stress controlled surface diffusion during epitaxial film growth*, Mater. Res. Soc. Symp. Proc., 356 (1995), pp. 33–44.
[7] C. Duport, *Elasticite et croissance cristalline*, Ph.D. thesis, Universite Joseph Fourier, Grenoble, France, 1996.
[8] C. Duport, P. Nozieres, and J. Villain, *New instability in molecular beam epitaxy*, Phys. Rev. Lett., 74 (1995), pp. 134–137.
[9] C. Duport, P. Politi, and J. Villain, *Growth instabilities induced by elasticity in a vicinal surface*, J. Phys. I, 5 (1995), pp. 1317–1350.
[10] C. Duport, C. Priester, and J. Villain, *Equilibrium shapes of a coherent epitaxial cluster*, in Morphological Organization in Epitaxial Growth and Removal, Z. Zhang and M. Lagally, eds., World Scientific, Singapore, 1997, pp. 73–97.
[11] W. E, *Selected problems in material science*, in Mathematics Unlimited—2001 and Beyond, Part I, B. Engquist and W. Schmid, eds., Springer-Verlag, New York, 2001.
[12] W. E and N. K. Yip, *Continuum theory of epitaxial crystal growth* I, J. Statist. Phys., 104 (2001), pp. 221–253.
[13] J. P. Van Der Eerden, *Crystal growth mechanisms*, in Handbook of Crystal Growth 1, D. T. J. Hurle, ed., North-Holland, Amsterdam, 1993, pp. 307–475.
[14] L. B. Freund and F. Jonsdottir, *Instability of a biaxially stressed thin film on a substrate due to material diffusion over its free surface*, J. Mech. Phys. Solids, 41 (1993), pp. 1245–1264.
[15] H. Gao, *Morphological instabilities along surfaces of anisotropic solids*, in Modern Theory of Anisotropic Elasticity and Applications, J. J. Wu, T. C. T. Ting, and D. M. Barnett, eds., SIAM, Philadelphia, 1991, pp. 139–150.
[16] J. Grilhe, *Study of roughness formation induced by homogeneous stress at the free surfaces of solids*, Acta Metall. Mater., 41 (1993), pp. 909–913.
[17] M. A. Grinfeld, *Instability of the separation boundary between a non-hydrostatically stressed elastic body and a melt*, Soviet Phys. Dokl., 31 (1986), pp. 831–834.
[18] M. A. Grinfeld, *The stress driven instability in elastic crystals: Mathematical models and physical manifestations*, J. Nonlinear Sci., 3 (1993), pp. 35–83.
[19] V. M. Kaganer and K. H. Ploog, *Strained islands as step bunches: Shape and growth kinetics*, Solid State Commun., 117 (2001), pp. 337–341.
[20] K. Kassner and C. Misbah, *Nonlinear evolution of a uniaxially stressed solid—a route to fracture*, Europhys. Lett., 28 (1994), pp. 245–250.
[21] J. Krug, *On the shape of wedding cakes*, J. Statist. Phys., 87 (1997), pp. 505–518.
[22] R. V. Kukta and K. Bhattacharya, *A three-dimensional model of step flow mediated crystal*

growth under the combined influences of stress and diffusion, Thin Solid Films, 357 (1999), pp. 35–39.

[23] R. V. Kukta and L. B. Freund, *Minimum energy configuration of epitaxial material clusters on a lattice-mismatched substrate*, J. Mech. Phys. Solids, 45 (1997), pp. 1835–1860.

[24] F. Liu, J. Tersoff, and M. G. Lagally, *Self-organization of steps in growth of strained films on vicinal substrates*, Phys. Rev. Lett., 80 (1998), pp. 1268–1271.

[25] T. S. Lo and R. V. Kohn, *A new approach to the continuum modeling of epitaxial growth: Slope selection, coarsening, and the role of the uphill current*, Phys. D, 161 (2002), pp. 237–257.

[26] T. Marschner, S. Lutgen, M. Volk, W. Stolz, E. O. Gobel, N. Y. Jin-Phillipp, and F. Phillipp, *Strain-induced changes in epitaxial layer morphology of highly strained* III/V-*semiconductor heterostructures*, Superlattices Microstruct., 15 (1994), pp. 183–186.

[27] W. W. Mullins, *Theory of thermal grooving*, J. Appl. Phys., 28 (1957), pp. 333–339.

[28] Y. H. Phang, C. Teichert, M. G. Lagally, L. J. Peticolas, J. C. Bean, and E. Kasper, *Correlated-interfacial-roughness anisotropy in* $Si_{1-x}Ge_x$/Si *superlattices*, Phys. Rev. B, 50 (1994), pp. 14435–14445.

[29] O. Pierre-Louis, C. Misbah, Y. Saito, J. Krug, and P. Politi, *New nonlinear evolution equation for steps during molecular beam epitaxy on vicinal surfaces*, Phys. Rev. Lett., 80 (1998), pp. 4221–4224.

[30] A. Pimpinelli and J. Villain, *Physics of Crystal Growth*, Cambridge University Press, New York, 1998.

[31] P. Politi and J. Villain, *Ehrlich-Schwoebel instability in molecular-beam epitaxy: A minimal model*, Phys. Rev. B, 54 (1996), pp. 5114–5129.

[32] K. Pond, A. C. Gossard, A. Lorke, and P. M. Petroff, *Role of steps in epitaxial growth*, Mater. Sci. Eng. B, 30 (1995), pp. 121–125.

[33] C. D. Rudin and B. J. Spencer, *Equilibrium island ridge arrays in strained solid films*, J. Appl. Phys., 86 (1999), pp. 5530–5536.

[34] T. P. Schulze and W. E, *A continuum model for the growth of epitaxial films*, J. Cryst. Growth, 222 (2001), pp. 414–425.

[35] R. L. Schwoebel, *Step motion on crystal surfaces*, J. Appl. Phys., 40 (1969), pp. 614–619.

[36] L. L. Shanahan and B. J. Spencer, *A codimension-two free boundary problem for the equilibrium shapes of a small three-dimensional island in an epitaxially strained solid films*, Interfaces and Free Boundaries, 4 (2002), pp. 1–25.

[37] A. Sidi and M. Israeli, *Quadrature methods for periodic singular and weakly singular Fredholm integral equations*, J. Sci. Comput., 3 (1988), pp. 201–231.

[38] P. Smilauer and D. D. Vvedensky, *Coarsening and slope evolution during unstable epitaxial growth*, Phys. Rev. B, 52 (1995), pp. 14263–14272.

[39] B. J. Spencer, *Asymptotic derivation of the glued-wetting-layer model and contact-angle condition for Stranski-Krastanow islands*, Phys. Rev. B, 59 (1999), pp. 2011–2017.

[40] B. J. Spencer and D. I. Meiron, *Nonlinear evolution of the stress-driven morphological instability in a two-dimensional semi-infinite solid*, Acta Metall. Mater., 42 (1994), pp. 3629–3641.

[41] B. J. Spencer and J. Tersoff, *Equilibrium shapes and properties of epitaxially strained islands*, Phys. Rev. Lett., 79 (1997), pp. 4858–4861.

[42] B. J. Spencer, P. W. Voorhees, and S. H. Davis, *Morphological instability in epitaxially strained dislocation-free solid films*, Phys. Rev. Lett., 67 (1991), pp. 3696–3699.

[43] D. J. Srolovitz, *On the stability of surfaces of stressed solids*, Acta Metall., 37 (1989), pp. 621–625.

[44] C. Teichert, J. C. Bean, and M. G. Lagally, *Self-organized nanostructures in* $Si_{1-x}Ge_x$ *films on* Si(001), Appl. Phys. A, 67 (1998), pp. 675–685.

[45] C. Teichert, M. G. Lagally, L. J. Peticolas, J. C. Bean, and J. Tersoff, *Stress-induced self-organization of nanoscale structures in SiGe/Si multilayer films*, Phys. Rev. B, 53 (1996), pp. 16334–16337.

[46] J. Tersoff, *Self-organized epitaxial growth of low-dimensional structures*, Phys. E, 3 (1998), pp. 89–91.

[47] J. Tersoff, *Surface stress and self-organization of steps*, Phys. Rev. Lett., 80 (1998), pp. 2018–2018.

[48] J. Tersoff and F. K. LeGoues, *Competing relaxation mechanisms in strained layers*, Phys. Rev. Lett., 72 (1994), pp. 3570–3573.

[49] J. Tersoff, Y. H. Phang, Z. Zhang, and M. G. Lagally, *Step-bunching instability of vicinal surfaces under stress*, Phys. Rev. Lett., 75 (1995), pp. 2730–2733.

[50] J. Villain, *Continuum models of crystal growth from atomic beams with and without desorp-*

tion, J. Phys. I, 1 (1991), pp. 19–42.
[51] D. D. Vvedensky, A. Zangwill, C. N. Luse, and M. R. Wilby, *Stochastic equation of motion for epitaxial growth*, Phys. Rev. E, 48 (1993), pp. 852–862.
[52] Y. Xiang and W. E, *Continuum model for epitaxial growth with elasticity on vicinal surface*, to appear.
[53] Y. H. Xie, G. H. Gilmer, C. Roland, P. J. Silverman, S. K. Buratto, J. Y. Cheng, E. A. Fitzgerald, A. R. Kortan, S. Schuppler, M. A. Marcus, and P. H. Citrin, *Semiconductor surface roughness: Dependence on sign and magnitude of bulk strain*, Phys. Rev. Lett., 73 (1994), pp. 3006–3009.
[54] W. H. Yang and D. J. Srolovitz, *Cracklike surface instabilities in stressed solids*, Phys. Rev. Lett., 71 (1993), pp. 1593–1596.
[55] Y. W. Zhang and A. F. Bower, *Numerical simulations of island formation in a coherent strained epitaxial thin film system*, J. Mech. Phys. Solids, 47 (1999), pp. 2273–2297.

SIAM J. APPL. MATH.
Vol. 63, No. 1, pp. 259–278

DERIVATION OF CONTINUUM TRAFFIC FLOW MODELS FROM MICROSCOPIC FOLLOW-THE-LEADER MODELS*

A. AW†, A. KLAR‡, T. MATERNE‡, AND M. RASCLE†

Abstract. In this paper we establish a connection between a microscopic follow-the-leader model based on ordinary differential equations and a semidiscretization of a macroscopic continuum model based on a conservation law. Naturally, it also turns out that the natural discretization of the conservation law in Lagrangian coordinates is equivalent to a straightforward time discretization of the microscopic model. We also show *rigorously* that, at least in the homogeneous case, the macroscopic model can be viewed as the limit of the time discretization of the microscopic model as the number of vehicles increases, with a scaling in space and time (a zoom) for which the density and the velocity remain fixed. Moreover, a numerical investigation and comparison is presented for the different models.

Key words. microscopic and macroscopic traffic models, Godunov scheme, hydrodynamic limit

AMS subject classifications. 35L45, 90B20

PII. S0036139900380955

1. Introduction.

Microscopic modeling of vehicular traffic is usually based on so-called follow-the-leader models; see [16], [6]. A system of ordinary differential equations is used to model the response of vehicles to their leading vehicle. These models usually consist of a system of second order ordinary differential equations. For instance (a more general nonlinearity could be considered as well), we consider

$$\begin{aligned} \dot{x}_i &= v_i, \\ \dot{v}_i &= C\frac{v_{i+1}-v_i}{(x_{i+1}-x_i)^{\gamma+1}} + A\frac{1}{T_r}\left[V\left(\frac{\Delta X}{x_{i+1}-x_i}\right)-v_i\right], \end{aligned} \tag{1.1}$$

where $x_i(t)$, $v_i(t), i = 1, \ldots,$ are location and speed of the vehicles at time $t \in \mathbb{R}^+$, and ΔX is the length of a car. The basic idea is that the acceleration at time t depends on the relative speeds of the vehicle and its leading vehicle at time t and the distance between the vehicles. The constants $C > 0$, $A > 0$, $\gamma \geq 0$ and the relaxation time T_r are given parameters. In the *homogeneous* case $A = 0$ we recover the usual form of microscopic follow-the-leader models. For $A > 0$ a relaxation term is added, driving the velocity of the car to an equilibrium velocity V, which depends on macroscopic properties of the flow ahead of the driver. T_r is the corresponding relaxation time, different from (and typically much larger than) the reaction time of individual drivers. The constants C, γ are fitted to special situations (see [6]). A common choice is, for example, $\gamma = 0$. This case has to be treated separately from $\gamma > 0$; see below. Initial values $x_i(0) = x_i^0$, $v_i(0) = v_i^0$ have to be described with $v_i^0 \geq 0$ and $x_{i+1}^0 > x_i^0$. Sometimes, a time lag is included in the equations to account for the reaction times of the drivers.

*Received by the editors November 10, 2000; accepted for publication (in revised form) March 26, 2002; published electronically September 5, 2002. This research was supported by the EC-TMR network hyperbolic and kinetic equations (HYKE), by Deutsche Forschungsgemeinschaft (DFG), KL 1105/5, and by the NSF-CNRS contract 5909.
http://www.siam.org/journals/siap/63-1/38095.html

†Laboratoire de Mathematiques, Université de Nice, F-06108 Nice Cedex, France (aaw@math.unice.fr, rascle@math.unice.fr).

‡Fachbereich Mathematik, TU Darmstadt, D-64289 Darmstadt, Germany (klar@mathematik.tu-darmstadt.de, materne@mathematik.tu-darmstadt.de).

Macroscopic modeling of vehicular traffic started with the work of Whitham [21]. He considered the continuity equation for the density ρ, closing the equation by an equilibrium assumption on the mean velocity v. The equation is

$$\partial_t \rho + \partial_x(\rho V(\rho)) = 0,$$

where $V = V(\rho)$ describes the dependence of the velocity with respect to the density for an *equilibrium* situation. An additional velocity equation has been introduced by Payne [15] and Whitham [21] as an analogy to fluid dynamics. Recently Daganzo [5] has pointed out some severe drawbacks of the Payne/Whitham-type models in certain situations. In [2] Aw and Rascle did develop a new heuristic continuum model avoiding these inconsistencies:

$$\begin{aligned} \partial_t \rho + \partial_x(\rho v) &= 0, \\ \partial_t(\rho v) + \partial_x(\rho v^2) - \rho^2 P'(\rho)\partial_x v &= A\,\frac{\rho}{T_r}\left[V(\rho) - v\right], \end{aligned}$$

where $P(\rho)$ is a given function describing the anticipation of road conditions in front of the drivers, and P' denotes its derivative with respect to ρ. In [2], the authors considered the case of the homogeneous system $A = 0$, but one can also consider in particular the case $A > 0$; see [17] and also [8]. Using the new variable

$$w = v + P(\rho)$$

the model can be written in conservative form as follows:

$$\begin{aligned} \partial_t \rho + \partial_x(\rho v) &= 0, \\ \partial_t(\rho w) + \partial_x(v\rho w) &= A\,\frac{\rho}{T_r}\left[V(\rho) - v\right]. \end{aligned}$$

Initial conditions have to be prescribed: $\rho(x,0) = \rho^0(x) \geq 0$ and $v(x,0) = v^0(x) \geq 0$. We note that the coefficients in the above models can be prescribed in an a priori way or derived from microscopic considerations. See, e.g., [11] for a derivation from a kinetic traffic flow equation.

In the present paper we show how the Aw–Rascle model can be viewed as the limit of a time discretization of a microscopic follow-the-leader model. In particular, the macroscopic coefficient $P = P(\rho)$ is determined from the microscopic model. The paper is arranged in the following way: in section 2 microscopic follow-the-leader models and the Aw–Rascle continuum model are considered in more details. In section 3 scaling limits of the microscopic equation are considered and the formal connection between microscopic and macroscopic model is established. Section 4 contains the full space-time discretization of both models and rigorous relations between the models. Section 5 considers numerically the convergence of the discretized system towards the conservation law in the limit of small time steps and large number of vehicles. We refer to [17] for a related discussion. Finally, when finishing this paper, we received from J. Greenberg—whom we thank—a very recent preprint [8], based on quite similar ideas; see section 4.1 for some comments and a comparison of the results. We have also learned about closely related ideas in [22]. Clearly, these kind of ideas are on the rise.

2. The models. In this section we discuss the microscopic and macroscopic models in more detail.

2.1. The microscopic model. We reconsider the microscopic equations (1.1) with constant $C = C_\gamma$. We introduce a new variable, the distance between, say, the tails of two vehicles following each other:

$$l_i = x_{i+1} - x_i.$$

One obtains the system

$$\begin{aligned} \dot{x}_i &= v_i, \\ \dot{v}_i &= C_\gamma \frac{(v_{i+1} - v_i)}{l_i^{\gamma+1}} + A\frac{1}{T_r}(V(\rho_i) - v_i), \end{aligned}$$

where the local "density around vehicle i" and its inverse (the local (normalized) "specific volume") are, respectively, defined by

$$\rho_i = \frac{\Delta X}{l_i} \quad \text{and} \qquad \tau_i = \frac{1}{\rho_i} = \frac{l_i}{\Delta X}.$$

REMARK 1. *The density is often defined as the* number *of cars per unit length; here* $\nu := 1/l_i$*, and therefore has the dimension of the inverse of a length. With our definition, the density is already* normalized, $\rho = \nu.\Delta X := \nu/\nu_m$*, and is therefore dimensionless, so that the* maximal *density is* $\rho_m = 1/\tau_m = 1$*, when cars are "nose to tail." We will often write expressions like* ρ/ρ_m *or* τ/τ_m *to emphasize this normalization.*

Now define the constant C_γ by

$$C_\gamma = v_{ref}(\Delta X/\ \rho_m)^\gamma = v_{ref}(\Delta X\ \tau_m)^\gamma = \ v_{ref}\Delta X^\gamma,$$

where $v_{ref} > 0$ is a reference velocity, and the coefficient $(\Delta X\ \tau_m)^\gamma$ allows us to recognize in (1.1) the derivative of function $\tilde{P}(\tau)$ defined below in (2.2). One obtains the microscopic model

$$\begin{aligned} \dot{x}_i &= v_i, \\ \dot{v}_i &= \frac{1}{\Delta X}\,(v_{i+1} - v_i)\frac{v_{ref}\tau_m^\gamma}{\tau_i^{\gamma+1}} + A\,\frac{1}{T_r}\,(V(\rho_i) - v_i)\,, \end{aligned} \tag{2.1}$$

where again $\tau_m = 1$ with our definition. We have

$$\dot{l}_i = v_{i+1} - v_i \quad \text{or} \qquad \dot{\tau}_i = \frac{1}{\Delta X}\,(v_{i+1} - v_i).$$

Using the new variable

$$w_i :=\ v_i +\ \tilde{P}(\tau_i)\ \text{ with }\ \tilde{P}(\tau_i) := \begin{cases} \frac{v_{ref}}{\gamma}\ (\frac{\tau_m}{\tau_i})^\gamma, & \gamma > 0, \\ -v_{ref}\ln(\frac{\tau_i}{\tau_m}), & \gamma = 0, \end{cases} \tag{2.2}$$

we get

$$\dot{w}_i = A\,\frac{1}{T_r}\,(V(\rho_i) - v_i)\,.$$

Altogether, one notices that (2.1) can be rewritten in the form

$$\begin{aligned} \dot{\tau}_i &= \frac{1}{\Delta X}\,(v_{i+1} - v_i), \\ \dot{w}_i &= A\,\frac{1}{T_r}\,(V(\rho_i) - v_i)\,. \end{aligned} \tag{2.3}$$

The initial conditions are $\tau_i(0) = \tau_i^0 > 0$, $v_i(0) = v_i^0 \geq 0$.

2.2. The macroscopic model. In conservative form, the macroscopic system under consideration is given by the following equations:

$$\begin{aligned} \partial_t \rho + \partial_x \rho v &= 0, \\ \partial_t \rho w + \partial_x v \rho w &= A\, \frac{\rho}{T_r}\, \left[V(\rho) - v\right], \end{aligned} \tag{2.4}$$

where ρ is again defined as the (normalized) density, i.e., the (local) dimensionless fraction of space occupied by the cars, and v denotes the macroscopic velocity of the cars. Moreover, $A = 0$ in the case of the homogeneous model and A is a positive constant, say $A = 1$, for the relaxed model, and

$$w = v + P(\rho). \tag{2.5}$$

The hyperbolic part of the above system is written as

$$\begin{aligned} \partial_t \rho + \partial_x (\rho v) &= 0, \\ \partial_t (\rho w) + \partial_x (v \rho w) &= 0. \end{aligned} \tag{2.6}$$

In the following, we consider a special class of functions $P(\rho) := \tilde{P}(1/\rho)$, where $\tilde{P}$ is defined in (2.2). In other words, for $\rho > 0$,

$$P(\rho) = \begin{cases} \frac{v_{ref}}{\gamma} \left(\frac{\rho}{\rho_m}\right)^{\gamma}, & \gamma > 0, \\ v_{ref} \ln\left(\frac{\rho}{\rho_m}\right), & \gamma = 0, \end{cases} \tag{2.7}$$

where, as in the previous subsection, $\rho_m = 1$ and v_{ref} is a given reference velocity. The function P is not a pressure. In fact, it is homogeneous to a velocity [11], [17]. In the context of gas dynamics—completely irrelevant here; see [5], [2]—this pseudopressure P would be homogeneous to the enthalpy, so that the exponent γ here plays the role of the usual $(\gamma - 1)$. In particular, the case $\gamma = 0$ here would correspond to the isothermal case, with the same mathematical advantages and difficulties; e.g., one of the Riemann invariants is unbounded near regions of local vacuum; see section 4.1. To obtain a well-defined problem for the case with relaxation—$A > 0$—we choose the function $V = V(\rho)$, $\rho > 0$, such that

$$-P'(\rho) \le V'(\rho) \le 0 \tag{2.8}$$

is fulfilled for all $\rho > 0$. This is the so-called *subcharacteristic condition*; see, e.g., [21], [3], [7]. A typical choice would be $V(\rho) = -c\,(P(\rho) - P(\rho_m))$, $0 \le c \le 1$, and a description of the equilibrium curve $v = V(\rho)$ in the (w, v) plane is shown in Figure 2.1 (left) for the case $\gamma > 0$ and in Figure 2.2 (left) for $\gamma = 0$. Of particular interest is the *characteristic* case, where the equality holds in one of the above inequalities. More precisely, if $\gamma > 0$, we assume that

$$V'(\rho) = -P'(\rho) \text{ (resp., 0) for } \rho \le \rho_* \text{ (resp., } \rho_* \le \rho \le \rho_m), \tag{2.9}$$

where ρ_* is some positive intermediate value between $\rho = 0$ and the maximal value ρ_m of ρ. The equilibrium curve is shown in Figure 2.1 (right); see [17].

In contrast, if $\gamma = 0$, then in the characteristic case we will assume that

$$\begin{aligned} &V(\rho) = v_m \text{ (resp., } -P(\rho)) \text{ (resp., 0)} \\ &\text{for } 0 \le \rho \le \rho_* \text{ (resp., } \rho_* \le \rho \le \rho_{**}) \text{ (resp., } \rho_{**} \le \rho \le \rho_m), \end{aligned} \tag{2.10}$$

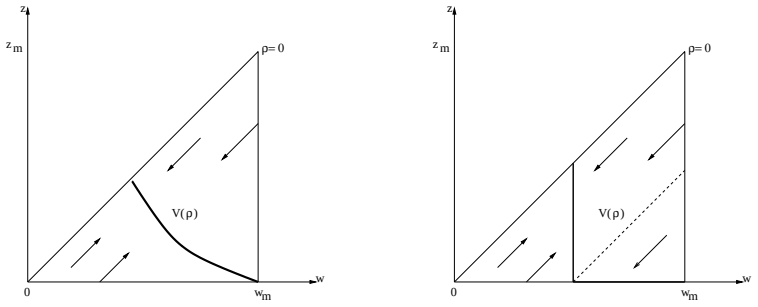


FIG. 2.1. *Invariant region* R *and equilibrium curve* $v = V(\rho)$ *for* $\gamma > 0$*, in the* $(w, z) = (w, v)$ *plane, in the subcharacteristic case (left) and in the characteristic case (right). In the first case, the convexity of the equilibrium curve could be arbitrary.*

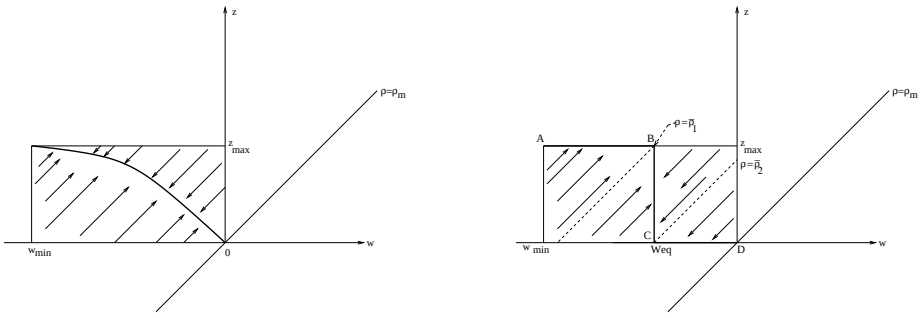


FIG. 2.2. *Invariant region* R *and equilibrium curve* $v = V(\rho)$ *for* $\gamma = 0$*, in the* (w, v) *plane, in the subcharacteristic case (left) and in the characteristic case (right).*

where ρ_* and ρ_{**} are two positive intermediate densities; see Figure 2.2 (right). In this case $\gamma = 0$, Figure 2.3 shows examples of the *same* equilibrium curve and associated bounded region (in the $(\rho,\ \rho\ v)$ plane), which are *invariant* for the *full* system (2.4) in the *subcharacteristic* case Figure 2.3 (left) and in the *characteristic* case Figure 2.3 (right).

Moreover, let $\tau := \rho^{-1}$ be the specific volume, and define the associated functions

$$\tilde{P}(\tau) = P\left(\frac{1}{\tau}\right), \qquad \tilde{V}(\tau) = V\left(\frac{1}{\tau}\right), \qquad w = v + \tilde{P}(\tau). \tag{2.11}$$

We note that

$$\tilde{P}'(\tau) = -\frac{v_{ref}\tau_m^{\gamma}}{\tau^{\gamma+1}}, \qquad \gamma \geq 0,$$

where $\tilde{P}'$ denotes the derivative of $\tilde{P}$ with respect to τ and, as in Remark 1, $\tau_m := \rho_m^{-1} = 1$. For $\rho > 0$ we have $\tilde{P}' < 0$ and $\tilde{P}'' > 0$. For $\rho > 0$, using the specific volume τ, we now transform (2.4): we change the Eulerian coordinates (x, t) into Lagrangian "mass" coordinates (X, T) (see [4]) with

$$\partial_x X = \rho, \qquad \partial_t X = -\rho v, \qquad T = t$$

or

$$\partial_X x = \rho^{-1} = \tau, \qquad \partial_X t = 0, \qquad \partial_T x = v, \qquad \partial_T t = 1.$$

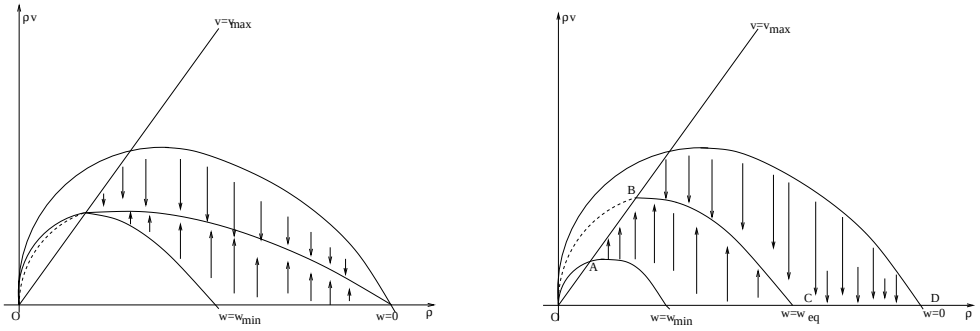


FIG. 2.3. *Invariant region* R *and equilibrium curve* $v = V(\rho)$ *for* $\gamma = 0$*, in the* $(\rho,\ \rho\, v)$ *plane, in the subcharacteristic case (left) and in the characteristic case (right).*

Thus, $X = \int^x \rho(y,t)dy$ is not a mass. In fact, it describes the total space occupied by cars up to point x. We obtain

$$\partial_T \tau - \partial_X v = 0, \tag{2.12}$$
$$\partial_T w = A\,\frac{1}{T_r}\left[V(\rho) - v\right].$$

Since $w = v + \tilde{P}(\tau)$, this is a hyperbolic system to the unknown functions w and τ, with relaxation term if $A > 0$. We add initial conditions $\tau^0(x) > 0$ and $v^0(x) \geq 0$.

We note that, as in the case of gas dynamics [20], even for *weak* (L^∞) solutions, this new system is equivalent to (2.4). This equivalence holds *even* in the vacuum case, where the map $x \to X$ is not invertible, so that $\partial_X x = \tau$ contains a delta-function. However, in this case one must admit for (2.12) a larger class of test-functions, which is an additional difficulty. In the numerical schemes described below, each cell moves between two trajectories, so that the total mass inside this cell remains constant. Therefore in each nonvoid cell, the (usual) weak solutions to (2.4) and (2.12) are the same.

3. Scaling and formal macroscopic limit of the microscopic equations. According to (2.3) the microscopic system can be written as

$$\dot{\tau}_i = \frac{1}{\Delta X}(v_{i+1} - v_i),$$
$$\dot{w}_i = A\,\frac{1}{T_r}\left(V(\rho_i) - v_i\right),$$

where $w_i = v_i + \tilde{P}(\tau_i)$ is defined in (2.2). On the other hand, denoting time by t as in Eulerian coordinates, the Lagrangian macroscopic system (2.12) is rewritten as

$$\partial_t \tau = \partial_X v,$$
$$\partial_t w = A\,\frac{1}{T_r}\left[V(\rho) - v\right],$$

with again $w = v + \tilde{P}(\tau)$. Clearly, (2.3) is at least a rough *semi*discretization of (2.12).

Now let us introduce the scaling. Obviously a macroscopic description for traffic flow is only valid if we consider a large number of vehicles on a long stretch of the highway. Therefore we introduce a scaling such that the size of the domain under

consideration goes to infinity, as well as the number of vehicles, whereas the length of cars shrinks to 0.

In other words, we "make a zoom," i.e., we introduce a small parameter ε, and we multiply the space and time units by $1/\varepsilon$; i.e., we shrink space and time coordinates x and t to

$$x' := \varepsilon\, x \qquad \text{and} \qquad t' := \varepsilon\, t.$$

In particular, the length of a car is now $\Delta X' := \varepsilon.\Delta X$.

Practically, the parameter ε is proportional to the inverse of the maximal possible number of cars per (new) unit length. The space and time derivatives are multiplied by ε. Similarly, since X is the primitive of ρ in x, it is replaced by $X' := \varepsilon\ X$, and therefore the derivatives in X are also multiplied by ε. On the other hand, the normalized density and specific volume are unchanged, as well as the velocity and the other Riemann invariant $w = v + \tilde{P}(\tau)$:

$$\rho' = \rho\ , \qquad \tau' = \tau\ , \qquad v' = v\ , \qquad w' = w.$$

Dropping the primed notation for these unchanged dependent variables, system (2.12) becomes

$$\begin{aligned} \frac{\partial \tau}{\partial t'} &= \frac{\partial v}{\partial X'}\ , \\ \frac{\partial w}{\partial t'} &= A\ \frac{1}{\varepsilon\ T_r}\left[V(\rho) - v\right]. \end{aligned} \tag{3.1}$$

Now let us look at the microscopic system, with the same scaling. The only additional modification is $l_i' = \varepsilon\ l_i$, and the relation $\tau_i = l_i/\Delta X$ is preserved with the primed variables. Again dropping the primes for the unchanged dependent variables, system (2.3) becomes

$$\begin{aligned} \frac{d\tau_i}{dt'} &= \frac{1}{\Delta X'}(v_{i+1} - v_i) \\ \frac{dw_i}{dt'} &= \frac{A}{\varepsilon\ T_r}\ \left(V(\rho_i) - v_i\right). \end{aligned} \tag{3.2}$$

Now let us discuss the consequences of the above scaling on the *equations.* There are two cases.

The homogeneous case $A = 0$. In this case, not surprisingly, the hyperbolic system and the microscopic system remain unchanged with this self-similar scaling. The only (*important*) difference is that in the new coordinates, the mesh size (see the next section) $\Delta X' = \varepsilon \Delta X$ tends to 0 when the zoom parameter ε tends to 0.

Therefore, at least *formally,* the microscopic system "converges" to the macroscopic one when ε tends to 0. More precisely, in this homogeneous case $A = 0$, (3.2) can be viewed as the natural semidiscretization of (3.1); see section 4—finer and finer when ε tends to 0. Obviously, the scaling changes the *initial data*, see Remark 3 below.

The relaxed case $A > 0$. Then there are two possibilities. First, assume that the positive constant A in front of the relaxation time depends nicely on some macroscopic scale and is, in fact, proportional to ε. In other words, let us assume that the relaxation time is comparable in size to the number of cars per (rescaled)

unit length. We note that for numerical purposes, we do not really need to let ε tend to 0, but we need only to consider a "small" ε, so that the semidiscretization (3.2) is "fine enough"; see Remark 2 below. In this case, the conclusion is the same as in the homogeneous case: the macroscopic system is at least the *formal* limit of the microscopic one. Second, on the contrary, assume that this constant A is unchanged in the scaling; then we formally end up with a scalar Lighthill–Whitham–Richards-type equation, but then the limit we are considering is the limit $(\Delta X', \Delta t') = (\Delta X, \Delta t)\ \varepsilon \to\ (0,0)$, with ΔX and Δt constant.

REMARK 2. *The size of the physical quantities allows for various possibilities to scale nicely the equations, with relatively small (but* finite*) values of ε, possibly with* different *scaling constants in x and t. For instance, assume that the "old" units are meter and second. Then, choose as "new" units (or reference length and time)* 1500 *m (or a mile) and* 1 *minute, with $\Delta X = 5$ m. Then we rescale as follows:*

$$x' = \frac{x}{1500}, \qquad t' = \frac{t}{60}.$$

On the other hand, a typical velocity is 90 *km/h, i.e.,* 25 *m/s, or* 1500 *m per mn, i.e.,* 1 *in the new units, which is perfect. Moreover, in these new units, the length of a car is $\Delta X' = 5/1500 = 1/300$, whereas a good time step in the time discretization, of the same order as the reaction time of the drivers, would be $\Delta t = 1/5$ second $= 1/300$ of the new time unit. Thus, in such a system of units, a typical (maximal?) velocity is of order* 1*, as well as the maximal (normalized) density, whereas typical space and time steps are of the order of* 1/300 *of the corresponding unit, subject of course to the CFL condition; see the next section. On the other hand, the relaxation time is typically found to be around* 30 *seconds, i.e.,* 0.5 *in the new units; see, for example,* [12]*. In such a scaling, the rescaled relaxation time, i.e., $\frac{A}{\varepsilon T}$, would still remain* finite*, and therefore we would still be far away from the zero-relaxation limit, i.e., from the Lighthill–Whitham–Richards model.*

REMARK 3. *So far, we have not discussed the problem of the* initial data*. Let us restrict ourselves to the* homogeneous *case $A = 0$, say, in Lagrangian coordinates. (The discussion would be the same in Eulerian coordinates.) In this case, as we said, the scaling preserves the system* (2.12) *(with $A = 0$), which we rewrite in the general form*

$$\frac{\partial U}{\partial t'} + \frac{\partial F(U)}{\partial X'} = 0, \tag{3.3}$$

with $U := U^\varepsilon := (\tau^\varepsilon,\ w^\varepsilon)$. However, this scaling modifies the initial data, where there are obviously (at least) two scales: the microscopic one, i.e., the length of a car (a few meters), and the macroscopic one (say, one kilometer). Therefore, it is natural to extend the microscopic initial data defined in section 2.1 *and to assume, for instance, that in rescaled Lagrangian coordinates the initial data are written as*

$$U_0^\varepsilon(X') = \sum_j \overline{U_j^0}\chi_j(X'), \tag{3.4}$$

where the characteristic function χ_j satisfies $\chi_j(X') = 1$ if and only if $X' \in I_j := (X'_{j-1/2},\ X'_{j+1/2})$, with $X'_l := l\Delta X'$, and $\overline{U_j^0}$ is the average value of a "macroscopic" function U_0 over the same interval.

When $\varepsilon \to 0$, the initial data (3.4) provide initial numerical data to approximate the solution of the initial value problem (3.3), (3.5), with

$$U(X', 0) = U_0(X'). \tag{3.5}$$

4. Rigorous relations between the microscopic and macroscopic equations. In this section microscopic and macroscopic discretizations are discussed as well as different convergence results.

4.1. The discretized models. In this section we will show that a standard explicit Euler discretization of the microscopic model is equivalent to the classical Godunov scheme applied to the macroscopic model. Moreover, this discretization is investigated in more detail.

The discretized microscopic model. We first introduce an explicit Euler time discretization of the *new* microscopic model, (3.2), using the *rescaled* time step $\Delta t'$. With the above scaling we note that the *new* Δt and ΔX tend to zero when ε tends to 0, with a *fixed* ratio $\lambda := \Delta t/\Delta X$. Neglecting the primed notation, i.e., writing Δt and ΔX instead of $\Delta t'$ and $\Delta X'$, we obtain

$$\tau_i^{n+1} = \tau_i^n + \frac{\Delta t}{\Delta X}(v_{i+1}^n - v_i^n), \tag{4.1}$$

with

$$v_i^{n+1} = w_i^{n+1} - \tilde{P}(\tau_i^{n+1}),$$

and if $A > 0$, the relaxation is approximated by

$$w_i^{n+1} = w_i^n e^{\frac{-A\,\Delta t}{\varepsilon T_r}} + (\tilde{V}(\tau_i^{n+1}) + \tilde{P}(\tau_i^{n+1}))(1 - e^{\frac{-A\,\Delta t}{\varepsilon T}}), \tag{4.2}$$

with $\rho_i^n = 1/\tau_i^n$. Of course, (4.2) contains the homogeneous case—if $A = 0$, then

$$w_i^{n+1} = w_i^n. \tag{4.3}$$

On the other hand, if $A > 0$, the relaxation term is correctly treated for ε small, i.e., for small relaxation times, where the equations are becoming stiff. Now let us discretize the macroscopic model.

The discretized macroscopic model. As above, we consider the macroscopic model (3.1) in rescaled variables x', t', with corresponding steps $\Delta t'$, $\Delta X'$, and we again drop the primed notations. Then (3.1) is discretized using a splitting scheme which treats separately the convection and the relaxation terms. Consider

$$\begin{aligned} \partial_t \tau - \partial_X v &= 0, \\ \partial_t w &= 0 \quad \text{if } A = 0, \end{aligned} \tag{4.4}$$

and

$$\partial_t w = \frac{A}{\varepsilon T}\,[V(\rho) - v] \quad \text{if } A \neq 0. \tag{4.5}$$

The most natural discretization to treat the convection part is the Godunov scheme. The relaxation part is treated by the same time discretization as for the microscopic model. Before writing Godunov's method for the hyperbolic equation, we need a brief description of the solution to the Riemann problem.

We consider the system (4.4), or the equivalent system (2.6) in Eulerian coordinates. We recall that $w = v + \tilde{P}(\tau)$. First, the eigenvalues of the system (4.4) are

$\lambda_1 = \tilde{P}'(\tau) < 0$ and $\lambda_2 = 0$. The Riemann invariants are w and $z := v$. They satisfy, for smooth solutions,

$$\partial_t v + \lambda_1 \partial_x v = 0, \qquad \partial_t w + \lambda_2 \partial_x w = 0.$$

Now, since $\tilde{P}''(\tau) > 0$, it turns out that the first eigenvalue λ_1 is genuinely nonlinear (GNL), i.e., for all (w, v), $\partial\lambda_1/\partial v \neq 0$. On the contrary, $\partial\lambda_2/\partial w \equiv 0$; i.e., $\lambda_2 = 0$ is linearly degenerate (LD), as for the original system (2.6) in Eulerian coordinates.

Now let us denote the left and right Riemann data by (w_L, v_L) and (w_R, v_R), respectively. Since the first characteristic speed is genuinely nonlinear, a state (w, v) can be connected on its left (in the (X, T) plane) to (w_L, v_L), either by a backward 1-shock if $v < v_L$, which corresponds to *braking*, or by a backward 1-rarefaction (acceleration) wave if $v > v_L$.

Moreover, see [2], these equivalent systems (2.6) and (4.4) are sometimes called *Temple* systems [19]; see also [10]. Their shock curves and rarefaction curves coincide. Therefore, *even in the case of a shock*, we have $w = w_L$.

On the other hand, (w, v) can be connected on its right to (w_R, v_R) by a stationary 2-contact discontinuity $v = v_L$. Hence, two states (w_L, v_L) and (w_R, v_R) can be connected through a constant intermediate state (w_0, v_0), which is connected to (w_L, v_L) by a 1-shock (braking) if $v_R < v_L$, or by a (continuous) 1-rarefaction wave (acceleration) if $v_R > v_L$ and to (w_R, v_R) by a 2-contact discontinuity. Moreover, w, v are monotone functions of X/t for all values of this variable, whereas τ is only monotone inside *each* elementary wave. In general, $(w_0, v_0) := (w_L, v_R)$, so that we can easily solve graphically the Riemann problem in the coordinates of Figure 2.1 or 2.2.

PROPOSITION 1. *We consider here the system* (4.4), *or the equivalent system* (2.6) *in Eulerian coordinates, with the above data* (w_L, v_L) *and* (w_R, v_R). *Then we have the following:*

1. *No local extremum of* w *or* v *is created for* $t > 0$. *Therefore, the* total variation *in space of* each *Riemann invariant is* nonincreasing *in time. More precisely,*

$$|f_+ - f_0| + |f_0 - f_-| = |f_+ - f_-|, \quad f := w \text{ or } v.$$

2. *In the case* $\gamma > 0$, *the density is nonnegative if and only if* $w - v = P(\rho) \geq 0$. *Consequently, the intermediate state is at* vacuum *if the cars in front are "too fast" with respect to the following cars, namely, if*

$$v_R > w_L = v_L + \tilde{P}(\tau_L).$$

In this case, $\rho = 1/\tau = 0$, v *and* w *are not physically defined, but if we insist and* mathematically *define, for instance,* $v_0 = w_0 = w_L$, *then statement* 1 *remains true.*

3. *In the same case* $\gamma > 0$, *any region defined by the Riemann invariants*

$$\mathcal{T} := \{0 \leq v \leq w \leq w_m = P(\rho_m)\}$$

is bounded *and* invariant *for the Riemann problem and corresponds to bounded nonnegative densities and velocities. An example of such a region is the triangle represented in Figure* 2.1. *Moreover, any rectangle*

$$\mathcal{R} := \{(w, v); 0 \leq w_{min} \leq w \leq w_m = P(\rho_m)\ ,\ 0 \leq v_{min} \leq v \leq v_{max}\} \tag{4.6}$$

inside $\mathcal{T}$ *also is* invariant *for the Riemann problem, away from vacuum if* $w_{min} - v_{max} > 0$.

4. *In the case $\gamma = 0$, vacuum corresponds to $w = -\infty$. Therefore,* any *rectangle*

$$\mathcal{R} := \{-\infty < w_m \le w \le 0 \le v \le v_m\} \tag{4.7}$$

is invariant *for the Riemann problem and corresponds to bounded nonnegative velocities and densities bounded from below by* positive *quantities.*

Proof. Statements 2 and 3 are related to the case $\gamma > 0$, which has been studied in detail in [2]. Statement 1 is then obvious, including if vacuum appears. Now, let us consider the case $\gamma = 0$, which is often considered in the literature on microscopic models; see, e.g., [16], [6]. In this case, it also is easy to check statement 4 again using the relations $w_0 = w_-$, $v_0 = v_+$. So the proof is complete. ∎

REMARK 4. *In the case $\gamma = 0$, when solving the Riemann problem, it is easy to check that the maximal possible speed that cars can reach in an acceleration wave emanating from (w_L, v_L) is infinite. Therefore, the cars behind can always catch up with the cars in front of them, without having to reach the vacuum; compare the numerical results in Figures* 5.2 *and* 5.3 *in section* 5 *below.* ∎

Now (4.4) is discretized using the Godunov method for the hyperbolic problem: We introduce grids in time and mass coordinates with (rescaled) stepsize Δt and ΔX and grid points t_n and $X_{i+1/2}$. Let f_i^n denote the approximation of the function $f(t, X)$ for $X \in [X_{i-1/2}, X_{i+1/2})$, $t \in [t_n, t_{n+1})$. Let $\lambda = \frac{\Delta t}{\Delta X}$ be the grid ratio.

In view of the above discussion, the Godunov method for system (4.4) is given by

$$\begin{aligned} w_i^{n+1} &= w_i^n, \\ \tau_i^{n+1} &= \tau_i^n + \lambda(v_{i+1/2}^n - v_{i-1/2}^n) \\ &= \tau_i^n + \lambda(v_{i+1}^n - v_i^n). \end{aligned} \tag{4.8}$$

And if $A \neq 0$, the full discretization is then

$$\begin{aligned} \tau_i^{n+1} &= \tau_i^n + \lambda(v_{i+1}^n - v_i^n), \\ w_i^{n+1} &= w_i^n e^{\frac{-A\,\Delta t}{\varepsilon T_r}} + (V(\rho_i^{n+1}) + P(\rho_i^{n+1}))(1 - e^{\frac{-A\,\Delta t}{\varepsilon T_r}}), \end{aligned} \tag{4.9}$$

so that we recover exactly the above system (4.2) (or (4.1) when $A = 0$). We have therefore shown the equivalence between the discretizations of the microscopic and the macroscopic system.

By the way, in the macroscopic view of this scheme it will be clear in Theorem 4.1 below that the (sub)characteristic condition is necessary for the stability. This is far from obvious in the microscopic interpretation.

Classically, the above numerical scheme consists of three successive steps, described here in Lagrangian coordinates for nonvoid cells:

1. Starting from piecewise-constant data $U_i^n := (\tau_i^n, w_i^n)$ in each cell, solve the Riemann problem for $t_n < t < t_{n+1}$ assuming that the CFL condition is satisfied. Let $U_h(X, t) := (\tau_h, w_h)(X, t)$ denote the corresponding solution. In fact, the index h stands for the couple $(\Delta X, \Delta t)$, and plays the role of the scaling parameter ε in section 3. We note that the intermediate state $U_{i+1/2}^n$ in the Riemann problem satisfies

$$w_{i+1/2}^n = w_{i+1/2}^n, \qquad v_{i+1/2}^n = v_{i+1}^n. \tag{4.10}$$

2. At time t_{n+1}, average this solution on each cell, i.e., solve (4.1). If $A \neq 0$, let us denote the average values of conservative variables by $(\tau_i^{n+1/2}, w_i^{n+1/2})$.

3. If $A \neq 0$, approximate the ordinary differential equation as above to obtain $(\tau_i^{n+1}, w_i^{n+1})$ from (4.9).

Again, the formulas would be the same in Eulerian coordinates, except that now the cells $x_{i-1/2}, x_{i+1/2}$ would be moving with time. Since $v_{i+1/2}^n = v_{i+1}^n$, we refresh the position by

$$x_{i+1/2}^{n+1} = x_{i+1/2}^n + \Delta t v_{i+1}^n.$$

Now let $I(a,b) := [\min(a,b), \max(a,b))$. Using Proposition 1 and standard ideas, we obtain the first important result.

THEOREM 4.1. *We consider the above algorithm, under the CFL condition, assuming that if $A \neq 0$, the (sub)characteristic condition is satisfied and that the initial data lie in an invariant rectangle $\mathcal{R}$, away from vacuum. Then we have the following.*

1. *We first assume that $A = 0$. Then, as in Proposition* 1, *in each Riemann problem the total variation of w_h is nonincreasing in time. Moreover, in each cell w_h remains constant: $w_h(x,t) \equiv w_i^n$. Consequently, v_h and also τ_h are monotone in each cell. Moreover, $(w_h(X,t), v_h(X,t))$ remains in the invariant region $\mathcal{R}$ for $t_n \leq t \leq t_{n+1}$.*
2. *In step* 2, *$w_i^{n+1/2} = w_i^n$. On the other hand, since in each cell w_h is constant and τ_h monotone, the average $\tau_i^{n+1/2}$ is in $I(\tau_i^n, \tau_{i+1/2}^n)$. By monotonicity the same result is true for the velocity. In fact,*

$$v_i^{n+1/2} \in I(v_i^n, v_{i+1/2}^n) = I(v_i^n, v_{i+1}^n).$$

3. *Finally, the invariant rectangle $\mathcal{R}$ also is invariant for the Godunov scheme. Moreover the total variation of the Riemann invariants, $\sum_j |f_{i+1}^n - f_i^n|$, $f =$ w or v, in space is still decreasing with respect to n. Since w_h is constant and v_h monotone in each cell, the total variation in* time *of w_h and of $\tilde{v}_h$: $(x,t) \to v_i^n + \Delta t(v_i^{n+1} - v_i^n)$ also is controlled from above.*
4. *On any time interval (t_n, t_{n+1}) the solution U_h satisfies the (discrete)* entropy inequality *in the sense of Lax: for any convex entropy $\eta(U) = \eta(\tau, w)$ associated with the entropy flux $q(U)$, and for any n and j,*

$$\eta(U_j^{n+1}) \leq \eta(U_j^n) - (\Delta t/\Delta X)(q(U_{j+1/2}^n) - q(U_{j-1/2}^n)). \tag{4.11}$$

5. *Now we consider the full problem, and we assume that the invariant rectangle $\mathcal{R}$ is constructed as in Figure* 2.1 *or* 2.2, *e.g., in the subcharacteristic case its upper left and lower right corner are at equilibrium. Then the region $\mathcal{R}$ also is invariant by* (2.12, 2) *and by step* 3, *i.e., by* (4.2). *Moreover, under the (sub)characteristic assumption, the* sum *of the total variations in space of the Riemann invariants, $\sum_j (\, |w_{j+1}^n - w_j^n| + |v_{j+1}^n - v_j^n| \,)$, is still nonincreasing in time, and the other conclusions of* (3) *remain valid for* (4.2). *Consequently, since the inverse function $\tilde{P}^{-1}$ is Lipschitz (away from vacuum), the total variation of τ_h also remains uniformly bounded in time.*

Proof. Statements (1) to (3) exploit, in particular, the monotonicity of v between v_j^n and v_{j+1}^n, which is obvious since w_h is constant in each cell, so that there is only one simple wave per cell. Note that v is not a conserved variable, so that an *Eulerian* classical Godunov scheme the averaging step would *not* preserve the total variation and the invariant regions.

Statement (4) is classical: on any time interval (t_n, t_{n+1}) the solution U_h is constructed by the Riemann problem and thus satisfies the *entropy inequality* in the sense

of Lax [13]. Therefore, by the Jensen inequality, the new average values U_j^{n+1} satisfy (4.11).

Finally, besides the above-mentioned references, it is an exercise to show (5). Indeed, in (4.9), compute the differences $(w_{i+1}^{n+1} - w_i^{n+1})$ and $(v_{i+1}^{n+1} - v_i^{n+1})$ in terms of the previous values $f_j^{n+1/2}$, $f = w$ or v, multiply each difference by its sign and add them. Then use the (sub)characteristic assumption to show the result. In the *characteristic* case (see [8]) we note that the evolution of w in each cell does *not* depend on the other cells, so that the total variation of each Riemann invariant w and v is nonincreasing in time, whereas in the subcharacteristic case (see [1], [14]) we can only control the *sum* of these total variations. □

4.2. Convergence results and hydrodynamic limit. There are three levels of description: the fully discrete system (4.9), or (4.1), (4.3), the follow-the-leader model (1.1), and the continuous system (4.4). In this section we discuss first the limit from the fully discrete level (4.8) to the continuous macroscopic model. Moreover, passing from the fully discrete level (4.8) to the semidiscrete one (1.1) and passing from the latter to the continuous level (4.4) are considered.

With the above scaling, we *state* a first *rigorous* result of convergence of the Godunov scheme when $\varepsilon \to 0$; i.e., we state a result dealing with the limit of (4.8) to (4.4) when the rescaled Δt and ΔX tend to 0 with a fixed ratio λ, fulfilling the CFL condition.

For simplicity, we consider the *homogeneous* case $A = 0$, away from vacuum, and we state the result in *rescaled* Lagrangian coordinates, again dropping the primed notations. However, our result also is valid for system (3.1), in rescaled variables, in the (less realistic) case where A is proportional to ε, i.e., $A_0\,\varepsilon$ instead of A. Similarly, the result also is the same for the two corresponding systems in *Eulerian* coordinates. Using the above and standard compactness results, as well as standard results to control the error in the projection steps, we obtain the following theorem.

THEOREM 4.2. *Let us consider the rescaled initial data* (3.5), *and assume that the associated Riemann invariants* w_0 *and* v_0 *are bounded, have a bounded total variation, and lie in an invariant rectangle* $\mathcal{R}$, *away from vacuum.*

Then, using the piecewise-constant initial data (3.4) *as initial data for this scheme, at least a subsequence* $U_h := (w_h, \tau_h)$ *produced by the numerical scheme* (4.1) *converges to a weak entropy solution to the initial value problem* (3.3), (3.5) *as the rescaled* Δt *and* ΔX *tend to* 0 *with a fixed ratio* λ, *fulfilling the CFL condition, as the zoom parameter* $\varepsilon \to 0$.

The above result deals with passing directly from (4.8) to (4.4). It strongly suggests studying the two other natural limits: passing from the fully discrete level (4.8) to the semidiscrete one (1.1), and passing from the latter to the continuous level (4.4). Again we restrict ourselves to the case $A = 0$, away from vacuum.

THEOREM 4.3. *Under the above assumptions, i.e.,* $A = 0$, *and the initial data lie in an invariant rectangle* $\mathcal{R}$, *away from vacuum, we consider in Lagrangian coordinates the values* $U_i^n := (\tau_i^n, w_i^n)$ *constructed by* (4.8) *or* (4.1), (4.3), *but now we rescale only the time step. Therefore the rescaled time step* Δt *vanishes, with a* fixed *space mesh size* ΔX.

Moreover, we assume that the initial data are constant for X *large enough, so that there is a "first" car. Then we have the following:*

1. *The IVP for the (infinite) follow-the-leader system* (1.1) *(with* $A = 0$*) has a unique solution* $U(t)$ *defined at least locally in time. Its natural first order*

approximation (4.1), (4.3) *is stable and consistent, and therefore the whole sequence is convergent for any* fixed ΔX.

2. *The values* U_i^n *stay in the invariant bounded region* $\mathcal{R}$ *and satisfy the uniform* BV*-estimates as in Theorem* 4.1. *Consequently, the solution* U *of* (1.1) *is globally defined and satisfies the same uniform estimates.*
3. *Moreover, set* $U_i := (\tau_i, w_i)$ *and let* $F_{i+1/2} := G(U_i, U_{i+1}) := F(U_{i+1/2}) = F(U_{i+1}) := (v_{i+1}(t),\ 0) = (w_{i+1} - \tilde{P}(\tau_{i+1}),\ 0)$ *denote the (well-defined) Godunov flux at the interface* $X = X_{i+1/2}$. *This nonlinear relation is preserved in the limit* $\Delta t \to 0$*: for all* $t \geq 0$, $F_{i+1/2}(t) := G(U_i(t), U_{i+1}(t)) = (v_{i+1},\ 0) = F(U_{i+1}(t))$. *Finally* (1.1) *is the semidiscretization of* (4.4)*: for any* $t \geq 0$,

$$\frac{dU_i}{dt}(t) = -(\Delta X)^{-1}\,(F_{i+1/2}(t) - F_{i-1/2}(t)). \tag{4.12}$$

Proof. The first part of this result can be adapted from standard textbooks (see, e.g., [18]) to the case of infinite-dimensional systems of ordinary differential equations, here with the l^∞ norm. The other results use the discrete BV estimates (in space and in time) inherited from the Godunov scheme. ∎

Now, define $\mathcal{U}_h(X,t) := \sum_j (\tau_i(t), w_i(t))\,\chi_i(X)$, where χ is defined as in (3.4). We have the following theorem.

THEOREM 4.4. *Under the same assumptions as in Theorem* 4.3*, consider the IVP for the follow-the-leader system* (1.1) *(with* $A = 0$*), and let* ΔX *tend to* 0. *Then at least a subsequence of the sequence* $\mathcal{U}_h$ *converges boundedly almost everywhere to an entropy weak solution* $U := (\tau, w)$ *to the macroscopic system,* (4.4) *for any smooth* $\phi(X,t)$ *with compact support,*

$$\begin{aligned}&\int_0^{+\infty} \sum_i \int_{I_i} [\,U(X,t)\,\partial_t\phi(X,t) \;+\; F(U(X,t))\,\partial_X\phi(X,t)]\,dX\,dt\\ &+\sum_i \int_{I_i} U_0(X)\,\phi(X,0)dX := A + B + D \;=\; 0,\end{aligned} \tag{4.13}$$

and similarly the entropy inequality in the sense of Lax holds for any convex entropy.

Proof. Multiply (4.12) by an arbitrary test-function $\phi(X,t)$, make a (discrete) integration by parts in X and t, and let ΔX tend to 0. We obtain, for any smooth function $\phi(X,t)$ with compact support contained in $[-L,L]\mathrm{x}[0,T]$,

$$\begin{aligned}&\int_0^{+\infty} \sum_i \int_{I_i} [\,U_i(t)\partial_t\phi(X,t) \;+ (\Delta X^{-1})F_{i+1/2}(t)(\phi(X+\Delta X,t) - \phi(X,t))]dXdt\\ &+\sum_i \int_{I_j} U_i(0)\phi(X,0)dX \;:= A_h + B_h + D_h = 0.\end{aligned} \tag{4.14}$$

By compactness, A_h and D_h, respectively, tend to A and D when $h \to 0$. As to B_h, with an obvious first order Taylor expansion, we see that for any ϕ, $|B_h - E_h| \leq C\ \Delta X$, where

$$E_h := \int_0^{+\infty} \sum_i \int_{I_i} F_{i+1/2}(t)\partial_X\phi(X,t)dXdt,$$

and C depends on the L^∞ norm of $F(U)$ and on $L\,T||\partial_X^2\phi||_\infty$.

Now (see Theorem 4.3), $F_{i+1/2}(t) = F(U_{i+1}(t))$, and F is Lipschitz continuous. Therefore, adding and subtracting $F(U_i(t))$, we obtain $|E_h - G_h| \le C'\,\Delta X$, where

$$G_h := \int_0^{+\infty} \sum_i \int_{I_i} F(U_i(t))\partial_X\phi(X,t)dXdt = \int_0^{+\infty} \int_{\setminus} F(\mathcal{U}_h)(X,t)\partial_X\phi(X,t)dXdt$$

and $|C'| \le T\,||\partial_X\phi||_\infty.\,||F'||_\infty.\,\sup_t\{\sum_i |U_{i+1}(t) - U_i(t)|\}$.

Finally, by compactness, G_h tends to B when $\Delta X \to 0$, which shows that U is a weak solution of (4.4). We would establish the entropy inequality in a similar way. First, when $\Delta t \to 0$ as in Theorem 4.3, the fully discrete entropy inequality (4.11) is preserved at the limit and provides the *semidiscrete* entropy inequality, i.e., a relation similar to (4.14), with U, $F(U)$, and the equality sign, respectively, replaced by $\eta(U)$, $q(U)$, and the same *inequality* sign as in (4.11), which implies the Lax entropy inequality by compactness as $\Delta X \to 0$. □

REMARK 5. *In other words, at least in the homogeneous case $A = 0$, and away from vacuum, we have shown that the macroscopic system is the limit of a large number of vehicles on a long stretch of a highway and a large time scale of the same order. For a study of more general cases, we refer to* [1]. *We note, by the way, that in this limit situation the time lag mentioned in the introduction vanishes.*

In contrast, in the relaxed case $A > 0$, with a fixed constant A, we have to study the limit of (4.1) *and* (4.2) *when the three parameters Δt, ΔX, and ε tend to* 0 *together, with fixed ratios, and, of course, satisfy the CFL condition. So far, we have not studied this limit.*

REMARK 6. *In terms of modeling, here we explicitly relate the semidiscretization of the macroscopic system to the microscopic system* (1.1) *directly, i.e., without any intermediate kinetic description. Although we already mentioned the derivation of* (2.4) *from kinetic models* [11], *our* direct *derivation is conceptually important: just imagine a similar result in gas dynamics! For the relation between weak solutions in Eulerian and Lagrangian mass coordinates, we again refer to* [20]. *Finally, we have also learned very recently of a preprint* [22] *with exactly the same formal derivation of the same model.*

5. Numerical methods and examples. For numerical investigations we consider the equations in Eulerian and Lagrangian form. The time discretized microscopic equations (4.1), (4.2) or, equivalently, the Godunov method in Lagrangian coordinates (4.9) can be viewed as a particle method for the conservation law. Computing

$$\begin{aligned} \tau_i^{n+1} &= \tau_i^n + \lambda(v_{i+1}^n - v_i^n), \\ w_i^{n+1} &= w_i^n e^{\frac{-\Delta t}{T(\rho_i^{n+1})}} + (V(\rho_i^{n+1}) + P(\rho_i^{n+1}))(1 - e^{\frac{-\Delta t}{T(\rho_i^{n+1})}}), \end{aligned} \tag{5.1}$$

one obtains the location of the vehicles as follows:

$$x_i^{n+1} = x_i^n + \Delta t v_i^n. \tag{5.2}$$

The density in Eulerian coordinates at the point x_i is then given by the interpolation $\rho_i = \frac{\Delta x}{x_{i+1} - x_i}$. For discretization steps Δt and ΔX tending to 0, one obtains an approximation of the conservation law (2.4) in Eulerian coordinates. From the particle point of view this means that we have to increase the number of vehicles to obtain the desired approximation of the conservation law. However, one could as well use any

other methods to resolve the limiting conservation law in Eulerian coordinates. For example, any second order shock capturing scheme could be used for the macroscopic equations (2.4). Using such a scheme will lead to the same results as the solution of the above discrete equations with a large number of vehicles. We use a relaxation method as developed in [9]. The method is adapted to include the relaxation term on the right-hand side of (2.4). This can be done in a straightforward way that preserves the second order approximation.

In the following we will compare the microscopic particle approach and the above second order scheme using two test problems.

In all cases the equilibrium velocity $V = V(\rho)$ is chosen as a function fitting to experimental data:

$$V(\rho) = U\left(\frac{\rho}{\rho_m}\right)$$

with

$$U(\rho) = v_m \frac{\frac{\pi}{2} + \arctan(11\frac{\rho - 0.22}{\rho - 1})}{\frac{\pi}{2} + \arctan(11 \cdot 0.22)},$$

where ρ_m is the maximal density and v_m is the maximal velocity. The function $P(\rho)$ is chosen by setting $\gamma = 0$ and $\gamma = 1$, i.e., $m_2 = 1$ and $m_2 = 2$, respectively. In order to fulfill the subcharacteristic condition (2.8), we choose the function P as $P(\rho) = 2\ln(\rho/\rho_m)$ and $P(\rho) = 6\rho/\rho_m$. The first test problem is the following: We consider normalized quantities with maximal speed v_m equal to 1 and maximal density ρ_m equal to 1. From the macroscopic point of view we consider a Riemann problem with left and right states given by $\rho_L = 0.05$, $\rho_L v_L = 0.0025$, $v_L = 0.05$, and $\rho_R = 0.05$, $\rho_R v_R = 0.025$, $v_R = 0.5$. The discontinuity is located at $x = 0$. Note that $v_R > w_L = v_L + P(\rho_L)$. Thus, vacuum states appear during the evolution for the continuous conservation law for $\gamma = 1$; see [2].

The discretization size is chosen as $\Delta x = \Delta X = \frac{1}{40}$. This leads to an initial number of cars equal to 800. They are initially distributed equally, spaced with the velocities 0.05 or 0.5, respectively. The time step is chosen according to the CFL condition.

Figure 5.1 shows density ρ and flux $q = \rho u$ at a fixed time for the particle and second order methods for $\gamma = 1$ without the relaxation term. Figure 5.2 shows the same for $\gamma = 0$. Figure 5.3 shows the evolution for $\gamma = 1$ with the relaxation term, where V and T are given by $V(\rho) = U(\rho)$ and $T(\rho) = \text{constant} = 20$. Figure 5.4 shows the same for $\gamma = 0$.

Finally, Figure 5.5 shows the results of our second test case, which is a more complicated situation: The evolution at a bottleneck at $x = 0$ is simulated. The number of lanes is reduced from three to two. This is achieved by setting the maximal density equal to the number of lanes. This means that the fundamental diagram is given by $V(\rho) = U(\frac{\rho}{\rho_m})$, where inside the bottleneck the maximal density ρ_m is reduced from 3 to 2. The boundary data on the left-hand side are chosen such that the flux in the three-lane region is slightly above the maximal possible flux in the two-lane region which creates the traffic jam. ΔX and Δx are chosen as $\frac{1}{4}$, which yields a number of cars around 5000 during the evolution. Figure 5.5 shows the evolution for the microscopic particle method at different times. In particular, the development of a traffic jam is observed in the figure. Identical results are obtained if the second order method for the Eulerian equations is used.

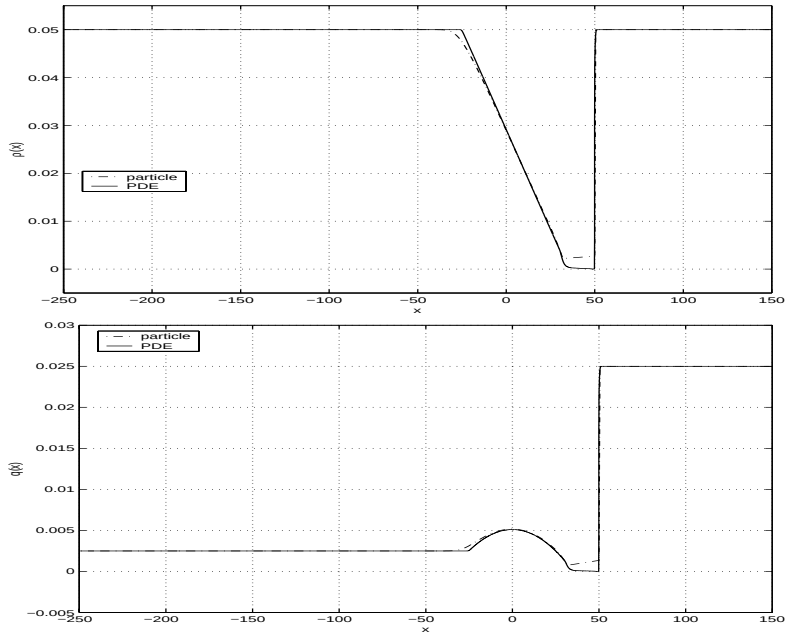


FIG. 5.1. *Time development of density and flux computed by the particle method and the second order method for $\gamma = 1$ without the relaxation term.*

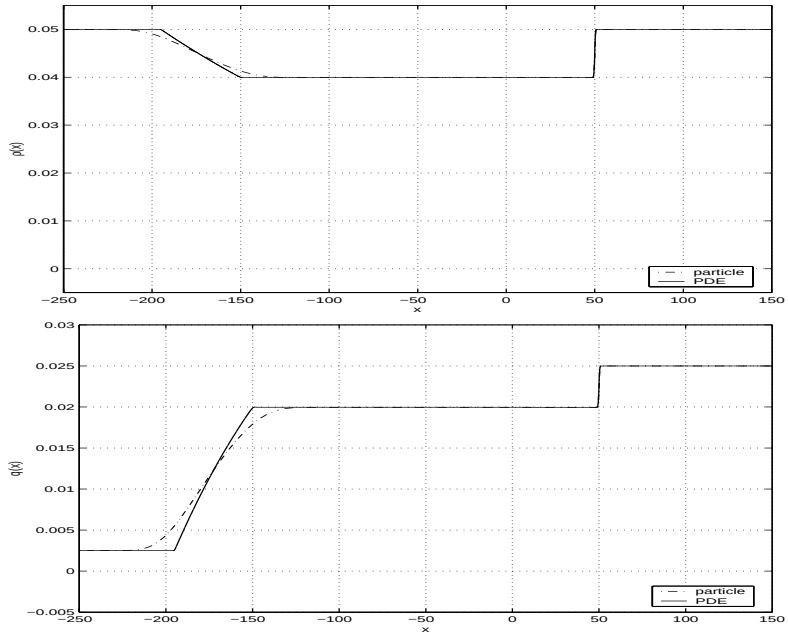


FIG. 5.2. *Time development of density and flux computed by the particle method and the second order method for $\gamma = 0$ without the relaxation term.*

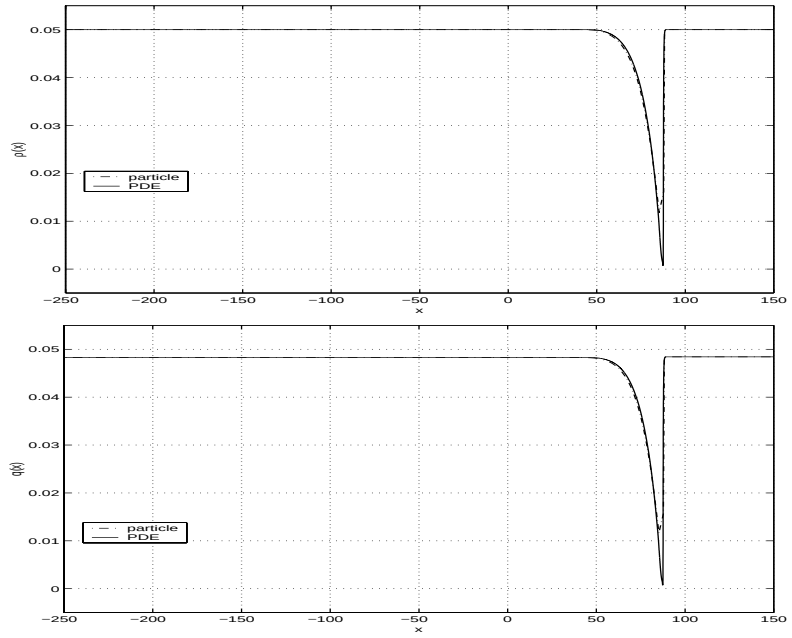


FIG. 5.3. *Time development of density and flux computed by the particle method and the second order method for* $\gamma = 1$ *with the relaxation term.*

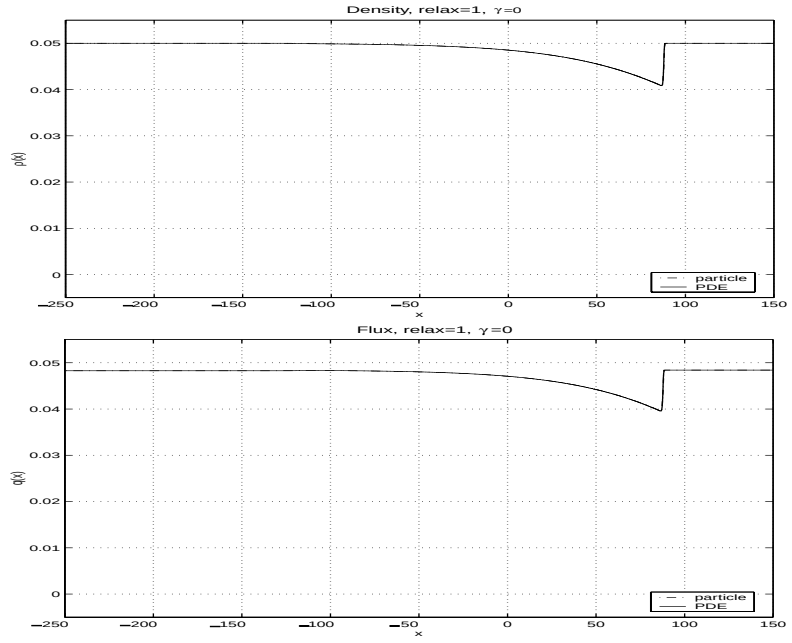


FIG. 5.4. *Time development of density and flux computed by the particle method and the second order method for* $\gamma = 0$ *with the relaxation term.*

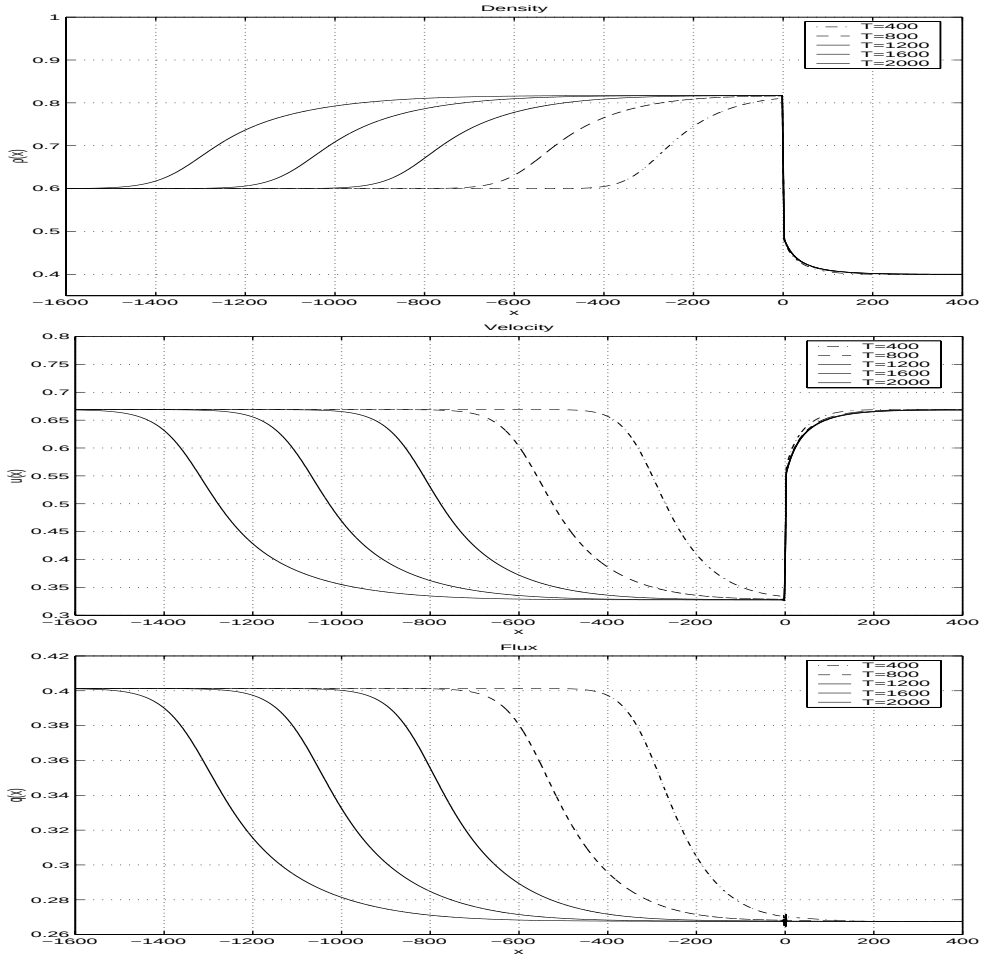


FIG. 5.5. *Time development of density, velocity, and flux. Lane drop from* 3 *to* 2 *lanes at* $x = 0$

REFERENCES

[1] A. AW, *Modèles hyperboliques de trafic automobile*, Ph.D. thesis, University of Nice, Nice, France, 2001.

[2] A. AW AND M. RASCLE, *Resurrection of "second order" models of traffic flow*, SIAM J. Appl. Math., 60 (2000), pp. 916–938.

[3] G. CHEN AND T. LIU, *Zero relaxation and dissipation limits for hyperbolic conservation laws*, Comm. Pure Appl. Math., 46 (1993), pp. 755–781.

[4] R. COURANT AND K. FRIEDRICHS, *Supersonic Flows and Shock Waves*, Springer-Verlag, 1976.

[5] C. DAGANZO, *Requiem for second order fluid approximations of traffic flow*, Transportation Research B, 29B (1995), pp. 277–286.

[6] D. GAZIS, R. HERMAN, AND R. ROTHERY, *Nonlinear follow-the-leader models of traffic flow*, Oper. Res., 9 (1961), p. 545.

[7] G. Q. CHEN, C. LEVERMORE, AND T. LIU, *Hyperbolic conservation laws with stiff relaxation terms and entropy*, Comm. Pure Appl. Math., 47 (1994), pp. 787–830.

[8] J. GREENBERG, *Extensions and amplifications of a traffic model of Aw and Rascle*, SIAM J. Appl. Math., 62 (2001), pp. 729–745.

[9] S. JIN AND Z. XIN, *The relaxation schemes for systems of conservation laws in arbitrary space dimensions*, Comm. Pure Appl. Math., 48 (1995), pp. 235–276.

[10] B. KEYFITZ AND H. KRANZER, *A system of non-strictly hyperbolic conservation laws arising*

in elasticity theory, Arch. Rational Mech. Anal., 72 (1979/1980), pp. 219–241.
[11] A. Klar and R. Wegener, *Kinetic derivation of macroscopic anticipation models for vehicular traffic*, SIAM J. Appl. Math., 60 (2000), pp. 1749–1766.
[12] R. Kühne, *Macroscopic freeway model for dense traffic*, in 9th International Symposium on Transportation and Traffic Theory, N. Vollmuller, ed., VNU Science Press, Utrecht, 1984, pp. 21–42.
[13] P. D. Lax, *Hyperbolic systems of conservation laws and the mathematical theory of shock waves*, in Regional Conf. Ser. in Appl. Math. 11, SIAM, Philadelphia, 1973.
[14] R. Natalini, *Convergence to equilibrium for the relaxation approximations of conservation laws*, Comm. Pure Appl. Math., 49 (1996), pp. 795–823.
[15] H. Payne, *FREFLO: A macroscopic simulation model of freeway traffic*, Transportation Research Record, 722 (1979), pp. 68–75.
[16] I. Prigogine and R. Herman, *Kinetic Theory of Vehicular Traffic*, Elsevier, New York, 1971.
[17] M. Rascle, *An improved macroscopic model of traffic flow: Derivation and links with the Lightill-Whitham model.*, Mat. Comput. Modelling, 35 (2002), pp. 581–590.
[18] M. Schatzman, *Analyse numérique*, InterEditions, Paris, 1991.
[19] B. Temple, *Systems of conservation laws with coinciding shock and rarefaction curves.*, Contemp. Math., 17 (1983), pp. 143–151.
[20] D. Wagner, *Equivalence of the Euler and Lagrangian equations of gas dynamics for weak solutions*, J. Differential Equations, 68 (1987), pp. 118–136.
[21] G. Whitham, *Linear and Nonlinear Waves*, Wiley, New York, 1974.
[22] M. Zhang, *A non-equilibrium traffic model devoid of gas-like behavior*, Transp. Research B, to appear.

SIAM J. APPL. MATH.
Vol. 63, No. 1, pp. 279–298

DIFFUSION APPROXIMATION OF A SCATTERING MATRIX MODEL OF A SEMICONDUCTOR SUPERLATTICE*

PIERRE DEGOND† AND KAIJUN ZHANG‡

Abstract. In this paper we derive a diffusion equation for electron transport in a superlattice. The starting model is a quantum scattering matrix model which relates the phase space density of each superlattice cell to that of the neighboring cells. Then, in the limit of a large number of cells, a diffusion equation for the particle number density in the position-energy space is obtained, which is of the "SHE" (spherical harmonics expansion) type. The diffusion constant retains the memory of the quantum scattering characteristics of the superlattice elementary cell (like, e.g., transmission resonances). An example is treated, for which the diffusion constants are analytically computed.

Key words. superlattices, scattering matrix model, diffusion approximation, spherical harmonics expansion, drift-diffusion, energy transport, transmission resonance

AMS subject classifications. 35Q20, 76P05, 82A70, 78A35

PII. S0036139999360015

1. Introduction.

Semiconductor superlattices are processed by growing periodic layers of two different semiconductor materials like GaAs and GaAlAs [25], [37]. The electronic properties of the two materials result in the establishment of a periodic electrostatic potential in the direction of the growth axis, which is discontinuous at each of the interfaces between the two materials. Superlattices possess a number of interesting physical properties, especially with respect to optoelectronics applications [25], [37]. An application to infrared radiation detection is described in [35].

The electronic properties of solids are characterized by the existence of energy bands [1]: the electron kinetic energy cannot take any arbitrary positive value, like in a vacuum, but may only belong to certain intervals, called energy bands. The bands are separated by forbidden energy gaps. Bloch's theory of bands provides the theoretical framework for this observation: it is a purely quantum mechanical effect originating from the periodic potential created by the regular arrangement of atoms in the crystal. In a superlattice, the periodic alternation of the two materials artificially creates a similar periodicity (although on a larger scale). Therefore, it is natural to expect that electron transport along the periodicity direction will exhibit the same kind of "band splitting." In this case, the bands are called "superlattice minibands," because the width of the energy bands is inversely proportional to the lattice period (see [25], [37]). For a recent account of the mathematical theory of bands, see [27].

Being a quantum mechanical phenomenon, the existence of energy bands is tightly connected with the notion of phase coherence: energy bands (or gaps) result from the constructive (or destructive) interference patterns of the electron wave-functions

*Received by the editors August 3, 1999; accepted for publication (in revised form) April 5, 2002; published electronically September 5, 2002. This work was supported by the TMR network ERB FMBX CT97 0157 on "Asymptotic methods in kinetic theory" of the European Community, by the LIAMA (Laboratoire d'Informatique, Automatique et Mathématiques Appliquées), by the PRA (Programme de Recherches Avancées), and by the Austrian START prize "Nonlinear Schrödinger Equations and Quantum Boltzmann Equations."

http://www.siam.org/journals/siap/63-1/36001.html

†Mathématiques pour l'Industrie et la Physique, UMR CNRS 5640, Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse cedex, France (degond@mip.ups-tlse.fr).

‡Department of Mathematics, Northeast Normal University, Changchun City 130024, P. R. China (zkjjkckj@public.jl.jl.cn).

between neighboring potential periods. A direct observation of the existence of energy minibands in a superlattice can be made through the measurement of Bloch oscillations [25]: electrons subject to a constant electric field undergo a periodic motion in space with zero mean. This is because each time the electron reaches the boundary of the energy band through the action of the field, it reverses direction in position space and goes backwards. Therefore, in theory, no transport in the direction of the field is allowed in a superlattice.

In practice, the situation is not so ideal due to collisions (or scattering). During their motion, electrons can collide against ionized impurities, phonons, etc. [32]. Another source of scattering, which is negligible in bulk semiconductors but which becomes extremely important in superlattices, is scattering by interface roughness [25], [15]. Indeed, at each material interface (and there are at least two of them per superlattice cell), the perfectly periodic arrangement of atoms is broken: the two materials very seldom have the same crystal constants, and this results in crystal defects at the interface. Even if the lattice mismatch (difference of the crystal constants of the two materials) is negligible, the growth process depends highly on the processing conditions and may not produce perfectly clean interfaces. This is particularly true for industrial processes, where cost constraints may contribute to poorer processing conditions.

A first consequence of collisions is that they allow some transport in the field direction, as they allow electrons to jump from one Bloch periodic trajectory to another one. To account for this effect, the reference model for superlattices, developed by Esaki and Tsu [21], is based on the hypothesis that quantum miniband theory applies to electron trajectories between collisions.

However, a second consequence of such collisions, which is not fully accounted for in the Esaki and Tsu model, is the breakdown of the phase coherence of the electron wave-functions. After a few collisions, constructive or destructive interference patterns can no longer be produced because wave-functions between neighboring cells have random phases with respect to each other. Therefore, transport loses its quantum mechanical nature over scales larger than a few mean free paths (the typical distance a particle travels between two collisions) and becomes purely classical; the quantum mechanical miniband theory still remains valid over shorter distances. The larger distance over which quantum mechanical interferences can occur is called the phase-coherence length. If the coherence length is of the order of the superlattice period, or a few superlattice periods, Bloch's theory can no longer be applied straightforwardly. In this case, Esaki and Tsu's model must be corrected to account for the fact that, even between two collisions, electron transport is not purely quantum and does not follow the miniband theory exactly. Instead, transport becomes classical on scales larger than the phase coherence length. The aim of the present paper is to derive a suitable transport model for such a situation.

2. General outline of the present work. Let ℓ denote the superlattice period. For simplicity, we first assume that the phase coherence length is equal to ℓ. We start with a microscopic model that is quantum mechanical over distances less than ℓ and classical over larger distances. For that purpose, we define the electron distribution function $f_{n+\frac{1}{2}}(k,t)$ on a periodic array of points $x_{n+\frac{1}{2}}$ separated by the period ℓ. $f_{n+\frac{1}{2}}(k,t)$ is a purely classical concept and denotes the density of particles at point $x_{n+\frac{1}{2}}$ having wave-vector k at time t. We recall that, in the semiclassical picture of electron kinetics, wave-vectors k and momenta p are related by $p = \hbar k$, where $\hbar$ is the reduced Planck constant.

Between the points $x_{n+\frac{1}{2}}$, the dynamics is quantum mechanical and must be described by means of quantum wave-functions, solutions of the stationary Schrödinger equation in the elementary cell. (We shall comment on this stationarity hypothesis later on.) Among the possible solutions, we focus on the scattering states, which are the quantum mechanical equivalent of the free moving particles. Knowledge of the scattering states gives us access to two important sets of data: the scattering probabilities, i.e., the probabilities with which particles coming into the elementary cell are either reflected or transmitted, and the quantum time delays, i.e., the duration of the reflection or transmission processes. The scattering probabilities reflect the fact that quantum tunneling through the potential structure is enhanced for some specific values of the wave-vector k called transmission resonances. Indeed, for these values of k, the transmission probability becomes close to unity.

The microscopic model consists of relating $f_{n+\frac{1}{2}}(k,t)$ to its neighboring values at previous times $f_{n+\frac{1}{2}\pm 1}(k,t-\tau)$ through the scattering coefficients. The resulting model is referred to in the literature as a "scattering matrix model" [34]. It retains the quantum nature of transport over distances less than or equal to the phase coherence length (i.e., one or a few superlattice periods) through the use of the scattering data, but reduces to a classical model over larger distances through the use of the classical distribution functions.

If the superlattice period ℓ is strictly less than the coherence length, we suppose that the latter is an integer multiple $N\ell$ of the former, and we consider $N\ell$ as the base period rather than ℓ. Therefore, the Schrödinger equation is solved for a potential structure that consists of an array of N elementary superlattice cells. As N increases, the solution comes closer and closer to that of the fully periodic problem. As soon as N passes above unity, the oscillations of the scattering data increase: transmission peaks (i.e., values of k such that the transmission probability is close to unity) due to resonant tunneling grow more numerous. Therefore, even at moderate values of N, the model can capture the highly oscillatory nature of superlattices that is observed in experimental measurements.

From the knowledge of $f_{n+\frac{1}{2}}$, we have access to the electron concentration, to the charge concentration (provided that the positive ion concentration is known), and therefore, to the self-consistent electric potential through the resolution of Poisson's equations. The self-consistent potential, in turn, is involved in the Schrödinger equation which determines the scattering data. In this way, the problem is fully coupled.

It is useful to note that the scattering matrix model is a kind of space and time discrete version of the Boltzmann equation [28]. In the same way that the continuous Boltzmann equation can be viewed as the Liouville equation of a stochastic particle system, the present model can also be viewed as a deterministic version of a random walk process. We shall come back to this point later.

The goal of this paper is to investigate the limit of the scattering matrix model when the total number of cells in the superlattice structure becomes large. Indeed, in practice, superlattices possess a large (but probably not very large) number of cells (on the order of 20), and it can be useful, at least for fast computations, to solve the model obtained by taking the limit of a large number of cells. After space and time rescaling, the scattering matrix problem appears as a perturbation problem which bears strong similarities with diffusion approximation problems in kinetic theory or with diffusion approximations of random walk processes.

We shall show that the limiting equation is a diffusion equation for the electron number density in the position-energy space (or energy distribution function). This

equation belongs to the class of spherical harmonics expansion (SHE) models, which have proved extremely useful in the context of standard semiconductor modeling. The present paper is the first (to our knowledge) to establish the SHE model in the framework of superlattices. Our derivation furnishes a direct connection between the quantum transport characteristics of the superlattice structure and the coefficients of the SHE model. This model provides means to achieve fast simulations of electron transport in superlattices. A related approach based on a continuous (rather than discrete) model can be found in [8].

Alternate macroscopic models for superlattices have also been proposed [25], [12]. They are based on a space-discrete version of pure drift-diffusion in position space only, while the present model deals with diffusion in a combined position-energy space. Our model therefore describes the physics at a more microscopic level and is thus expected to capture the highly oscillating nature of superlattices more accurately. Furthermore, the diffusion constants involved in [12] are phenomenological, while we shall provide an explicit methodology to derive these diffusion constants from the microscopic data (scattering probabilities and time delays). For instance, in [12], the reflection probabilities are ignored; on the other hand, a displacement current term is included in the current equation there. According to the numerical values, it seems that this term is small, and thus it will be neglected in our model. Globally, the present model seems to give access to a finer description of the physics, but it is also more complex in that it involves an additional variable (the particle energy).

We conclude this section with a few bibliographical comments. The diffusion approximation is a theoretical tool which links the evolution of macroscopic quantities like number or energy densities to the microscopic particle dynamics described by a kinetic equation. The diffusion approximation methodology goes back to the work of Hilbert, Chapman, and Enskog (see, e.g., [14] for an introduction to the subject). Its application to bulk semiconductors is reviewed from a physics view point in [32], [16]. The modern mathematical view of the theory was set up in [10] and [2] in the context of neutron transport (the former using a stochastic description of transport, the latter using purely deterministic models), and its application to semiconductors was developed in [30], [24]. In these works, the resulting macroscopic model is the drift-diffusion model, which is the basic tool in semiconductor modeling [28], [33] and which deals with the electron number density in position space.

By analyzing the various collision scales, it has recently been possible to derive a diffusion model for the particle number density in the position-energy space [20], [4], [17]. This model is often referred to in the literature as the SHE model (the term spherical harmonics expansion arises from its early derivation by physicists [36]). It has proved very efficient in semiconductor device modeling [22], [23] and is also used in plasma physics and gas discharge physics (e.g., [18]). An alternate derivation from a stochastic description of individual particle transport is proposed in [13].

The outline of the paper is as follows: the scattering matrix model is proposed in section 3 and appropriately scaled in section 4. Then, the formal diffusion limit is stated in section 5. Comments on the diffusion model and examples are developed in sections 6 and 7. Finally, the derivation of the diffusion model is (formally) proved in the appendix.

3. A scattering-matrix model for superlattices. We now summarize the previous discussion and set up the starting microscopic model. We consider a semiconductor superlattice consisting of layers of several materials periodically arranged in the direction x and generating a permanent periodic potential of period ℓ in this

direction. Since the superlattice period is usually very small (10 to 100 nm), electron motion through such a structure must be described quantum mechanically. However, we assume that various sources of scattering (e.g., interface roughness scattering due to some crystalline disorder at the various interfaces [25], [15]) produce a phase decoherence of the electron wave-functions over distances comparable with the superlattice period. In this context, quantum effects are limited within one superlattice period.

Let us denote the interval occupied by the nth superlattice pattern by $C_n = [(n-\frac{1}{2})\ell, (n+\frac{1}{2})\ell]$. Our assumption is that the state of the electron gas at each pattern boundary $x_{n+\frac{1}{2}} = (n+\frac{1}{2})\ell$ can be described by a classical distribution function $f_{n+\frac{1}{2}}(k,t)$, which represents the number of electrons at time t with wave-vector $k \in \mathbb{R}$ at this point. Then, finding the motion of an electron through the elementary superlattice pattern C_n reduces to a standard quantum mechanical scattering problem. Such a problem is characterized by reflection-transmission coefficients and time delays [29].

To be more specific, we first need to introduce some additional definitions [1]. For simplicity, we restrict our analysis to a one-dimensional geometry and assume that the energy-versus-wave-vector relationship in each material is parabolic, i.e., is of the form

$$\varepsilon(k) = \frac{\hbar^2 k^2}{2m},$$

where ε is the kinetic energy of an electron moving in the crystal with wave-vector $k \in \mathbb{R}$, $\hbar$ is the reduced Planck constant, and m is the electron effective mass. We shall place the points $x_{n+\frac{1}{2}}$ in the middle of the largest layer and denote the corresponding material by A. Unless otherwise specified, m will refer to the electron effective mass in this material, while $m^*(x)$ will refer to the position-dependent effective mass at any point x of the superlattice. (Of course, $m^*(x) = m$ in material A.) The electron velocity $v(k)$ is related to the wave-vector by

$$v(k) = \frac{1}{\hbar}\frac{d\varepsilon}{dk} = \frac{\hbar k}{m}.$$

We assume that the electric potential ϕ is the sum of two contributions $\phi = \phi_{SL} + \phi_{SC}$. The contribution ϕ_{SL} is specific to superlattice structures and derives from the presence of material discontinuities. It is piecewise constant with jump discontinuities at the material interfaces and, of course, has the periodicity of the superlattice structure. To fix it uniquely, we shall assume that it vanishes at the points $x_{n+\frac{1}{2}}$.

The contribution ϕ_{SC} is the self-consistent potential generated by the charges and by the applied bias. ϕ_{SC} is a solution of the Poisson equation

$$-\frac{d}{dx}\left(\epsilon\frac{d\phi_{SC}}{dx}\right) = \rho(x,t), \tag{3.1}$$

where $\rho(x,t)$ is the charge concentration, given in terms of the positive ion concentration $\rho^+(x)$ (supposed known and given) and of the electron concentration $\rho^-(x,t)$ (the expression of which will be given below) by

$$\rho(x,t) = e(\rho^+(x) - \rho^-(x,t)),$$

where e is the elementary (positive) charge. The Poisson equation (3.1) is supplemented by boundary conditions

$$\phi_{SC}(a,t) = 0, \quad \phi_{SC}(b,t) = \phi_{bias}, \tag{3.2}$$

where a and b are the left and right boundaries of the superlattice and ϕ_{bias} is the applied bias. The dielectric constant ϵ is an ℓ-periodic function of x, which may also have jump discontinuities at the material interfaces. ϕ_{SC} is not periodic in general and is time-dependent, as ρ generally is.

Now, we turn to the solution of the Schrödinger equation associated with the nth pattern C_n. Because of our assumption that the correlation length is of the order of ℓ, the wave-functions of electrons belonging to the nth pattern C_n cannot interact with the potential beyond the boundaries of C_n. To take this fact into account, we modify the potential experienced by the electrons in C_n into a constant potential $\bar{\phi}^n(x)$ outside C_n:

$$\bar{\phi}^n(x) = \begin{cases} \phi(x,t) & \forall x \in [x_{n-\frac{1}{2}}, x_{n+\frac{1}{2}}], \\ \phi(x_{n-\frac{1}{2}}) & \forall x \le x_{n-\frac{1}{2}}, \\ \phi(x_{n+\frac{1}{2}}) & \forall x \ge x_{n+\frac{1}{2}}. \end{cases} \tag{3.3}$$

Now we consider t as a frozen time variable, and for $k \in \mathbb{R}$ given we solve the Schrödinger equation on $\mathbb{R}$ for the wave-function $\psi_k(x)$:

$$-\frac{\hbar^2}{2}\frac{d}{dx}\left(\frac{1}{m^*(x)}\frac{d\psi_k}{dx}\right) - e\bar{\phi}^n\psi_k = \left(\frac{\hbar^2 k^2}{2m} - e\Phi_k^n\right)\psi_k, \tag{3.4}$$

where

$$\Phi_k^n = \begin{cases} \phi(x_{n-\frac{1}{2}}) & \text{if } k \ge 0, \\ \phi(x_{n+\frac{1}{2}}) & \text{if } k \le 0. \end{cases}$$

All the present discussion is classical and can be found, for instance, in [29]. We summarize it below for the reader's convenience. To uniquely specify the solution of (3.4), we impose the following additional conditions: for $k > 0$,

$$\psi_k = \begin{cases} e^{ik(x-x_{n-1/2})} + A(k)e^{-ik(x-x_{n-1/2})}, & x < x_{n-\frac{1}{2}}, \\ B(k)e^{ik_+(x-x_{n+1/2})}, & x > x_{n+\frac{1}{2}}, \end{cases}$$

and similarly, for $k < 0$,

$$\psi_k = \begin{cases} e^{-ik(x-x_{n+1/2})} + A(k)e^{ik(x-x_{n+1/2})}, & x > x_{n+\frac{1}{2}}, \\ B(k)e^{-ik_-(x-x_{n-1/2})}, & x < x_{n-\frac{1}{2}}. \end{cases}$$

We have let

$$k_\pm = \sqrt{k^2 \pm \frac{2me}{\hbar^2}\delta\phi^n(t)}, \qquad \delta\phi^n(t) = \phi(x_{n+\frac{1}{2}}, t) - \phi(x_{n-\frac{1}{2}}, t). \tag{3.5}$$

If the quantity inside the square root is negative, the choice of the pure imaginary root is indifferent and leads to the same solution.

These solutions are called the scattering states and describe free moving particles coming from infinity and entering the nth superlattice pattern C_n. The complex numbers A and B are the scattering amplitudes, from which one computes the reflection and transmission probabilities $R(k)$ and $T(k)$ according to

$$R(k) = |A(k)|^2, \qquad T(k) = \begin{cases} \frac{k_+}{k}|B(k)|^2, & k > 0, k_+ \in \mathbb{R}, \\ \frac{k_-}{|k|}|B(k)|^2, & k < 0, k_- \in \mathbb{R}, \\ 0, & \text{otherwise.} \end{cases}$$

The scattering probabilities satisfy the following relations:

$$R+T=1, \qquad \begin{cases} R,T(k)=R,T(-k_+), & k>0, k_+\in\mathbb{R}, \\ R,T(k)=R,T(k_-), & k<0, k_-\in\mathbb{R}. \end{cases} \tag{3.6}$$

Introducing the phases of the complex numbers A and B,

$$A(k)=|A(k)|e^{iS_R(k)}, \qquad B(k)=|B(k)|e^{iS_T(k)},$$

we define the semiclassical reflection and transmission time delays $\tau_R(k)$ and $\tau_T(k)$ according to the following:

$$\tau_R(k)=\frac{1}{v(k)}\frac{dS_R}{dk}, \qquad \tau_T(k)=\frac{1}{v(k)}\frac{dS_T}{dk}.$$

We also have (see [29])

$$\begin{cases} \tau_R,\tau_T(k)=\tau_R,\tau_T(-k_+), & k>0, k_+\in\mathbb{R}, \\ \tau_R,\tau_T(k)=\tau_R,\tau_T(k_-), & k<0, k_-\in\mathbb{R}. \end{cases} \tag{3.7}$$

Of course, all the scattering data depend on the index n, which has been omitted in the previous discussion.

We now establish the dynamics obeyed by the discrete distribution function $f_{n+\frac{1}{2}}(k,t)$. First, consider $f_{n+\frac{1}{2}}(k,t)$ for $k>0$, which corresponds to particles at point $x_{n+\frac{1}{2}}$ moving to the right; let us trace these particles back to previous times. Some of them have been transmitted through the nth pattern and come from point $x_{n-\frac{1}{2}}$ with a momentum $k_-^n>0$ corresponding to the energy shift $-e\delta\phi^n$. The crossing time of the nth pattern being $\tau_T^n(k_-^n)$, the number of these is $T^n(k_-^n)f_{n-\frac{1}{2}}(k_-^n,t-\tau_T^n(k_-^n))$. The other contribution to $f_{n+\frac{1}{2}}(k,t)$ is made of particles reflected by the nth pattern and coming from point $x_{n+\frac{1}{2}}$ with a momentum $-k$. The number of these is $R^n(-k)f_{n+\frac{1}{2}}(-k,t-\tau_R^n(-k))$. Therefore, we obtain the total number $f_{n+\frac{1}{2}}(k,t)$ by summing up these two contributions. For $k>0$, this leads to

$$\begin{aligned} f_{n+\frac{1}{2}}(k,t) &= T^n(k_-^n)f_{n-\frac{1}{2}}(k_-^n,t-\tau_T^n(k_-^n)) \\ &\quad + R^n(-k)f_{n+\frac{1}{2}}(-k,t-\tau_R^n(-k)), \end{aligned} \tag{3.8}$$

and for $k<0$, to

$$\begin{aligned} f_{n-\frac{1}{2}}(k,t) &= T^n(-k_+^n)f_{n+\frac{1}{2}}(-k_+^n,t-\tau_T^n(-k_+^n)) \\ &\quad + R^n(-k)f_{n-\frac{1}{2}}(-k,t-\tau_R^n(-k)). \end{aligned} \tag{3.9}$$

System (3.8)–(3.9) belongs to the class of scattering matrix models which are sometimes used in the literature [34]. These conditions were first proposed by Ben Abdallah [3] as coupling conditions for classical and quantum models (see also [7]).

To preserve causality, we require system (3.8)–(3.9) to be a backwards difference system in time, and consequently we assume that the time delays are positive. We consider that (3.8)–(3.9) describes the evolution of the system for $t>0$, starting from known states for all $t\le 0$, and prescribe the distribution functions for negative times:

$$f_{n-\frac{1}{2}}(k,t)=(f_I)_{n-\frac{1}{2}}(k,t) \qquad \forall t\le 0,\ \forall n\in\mathbb{Z},\ \forall k\in\mathbb{R}, \tag{3.10}$$

where $(f_I)_{n-\frac{1}{2}}$ is given.

For a bounded structure, the incoming distribution function must be prescribed at the boundary. Let us denote by $a = x_{n_a-\frac{1}{2}}$ and $b = x_{n_b+\frac{1}{2}}$ the left and right boundaries of the structure. Then, we prescribe

$$f_{n_a-\frac{1}{2}}(k,t) = f_a(k,t), \quad f_{n_b+\frac{1}{2}}(-k,t) = f_b(-k,t) \qquad \forall t,\ \forall k > 0,$$

where $f_a(k,t)$ and $f_b(-k,t)$ are given functions of $k > 0$. Usually, $f_{a,b}$ are chosen to be equal to thermodynamical equilibrium distribution functions (see the expression in section 6).

In order to complete the model, we must say how the electron concentration is computed. Following [3], we assume that the electron concentration in the nth pattern is the sum of the concentration of the right-going scattering states (for $k > 0$) weighted by the distribution function at the "left entrance" of the cell C_n, i.e., $f_{n-\frac{1}{2}}(k)$, and that of the left-going states (for $k < 0$) weighted by $f_{n+\frac{1}{2}}(k)$. In other words, for $x \in [x_{n-\frac{1}{2}}, x_{n+\frac{1}{2}})$,

$$\rho^-(x,t) = \int_{k>0} f_{n-\frac{1}{2}}(k,t)|\psi_k^n(x,t)|^2 \frac{dk}{\pi} + \int_{k<0} f_{n+\frac{1}{2}}(k,t)|\psi_k^n(x,t)|^2 \frac{dk}{\pi}. \tag{3.11}$$

The factor π^{-1} stands for the one-dimensional momentum density of states (see [1]).

Let us now comment on the hypothesis of frozen time in the solution of the quantum problem. This hypothesis is valid only if the time variation of the potential is slow compared with the transit times of the particles through the pattern C_n. If this is not the case, the potential varies significantly while the particle is crossing the pattern, and the resolution of a time-dependent Schrödinger equation becomes necessary. In such a case, not only would the computational cost increase, but the coupling to the semiclassical distribution function would be much more complicated (see, e.g., [5]). However, the frozen time assumption is acceptable in practice since the self-consistent electric field is a highly averaged, and therefore slowly evolving, quantity. In the present paper, we shall focus on this case. Therefore, all the scattering data in formulas (3.8)–(3.9) must be understood as corresponding to time t.

We note that formula (3.11) also relies on the frozen time assumption: the distribution function evolves on a slower time scale than the time delays, and therefore the statistics of particles in the entire pattern C_n can be approximated by the distribution function at its boundaries at the same time. This approximation is consistent with that made in the resolution of the quantum problem and is not in contradiction with (3.8)–(3.9), where the time delays appear. Indeed, that the time delays are small does not prevent the distribution function from evolving on larger time scales, as we will see in the derivation of the macroscopic model (section 5). Also, removing this assumption is possible, as in [5], but requires a much more complex model, which is beyond our scope here.

Obviously, a stochastic interpretation of the dynamical system (3.8)–(3.9) can be given. At a given time t, a particle sitting at point $x_{n-\frac{1}{2}}$ with momentum $k > 0$ (to fix the ideas) can jump to point $x_{n+\frac{1}{2}}$ with a probability T^n or can reverse momentum with probability R^n. The particle can perform its next jump after a certain time delay $\tau_{R,T}$. Therefore, this model is a dynamical system version of a kind of random walk. The connection between stochastic particle processes and kinetic equations is a very active field, and it is not our aim to elaborate on this relation here. Let us simply mention that, conversely, stochastic particle processes lead to efficient numerical algorithms to solve kinetic equations (referred to as Monte Carlo methods). An introductory monograph for this topic is [26].

In view of this stochastic interpretation, we discuss why the time delays are assigned deterministic values instead of random ones. Semiclassical scattering theory, which is based on a high frequency limit [29], gives very few means for exploring the possible random character of the time delays. Random time delays are likely to appear if the high frequency assumption is removed. However, there exists very little physical information on the probabilistic distribution of time delays in this case. In the absence of such information, the simplest choice seems to be the deterministic semiclassical value. Considering random time delays, however, would not lead to major changes to the present work. An integral over the distribution of time delays would appear in (3.8)–(3.9) and in the expression (5.4) of the density-of-states M of the diffusion model (see section 5).

Finally, let us note that we exclude bound states, which could contribute to trapping particles inside a given superlattice cell. If bound states were considered, additional terms involving the populations of these states would have to appear in (3.11). This would require a model of inelastic collisions, since only such collisions can fill these states. Such collisions are discarded in the present work, as are bound states.

4. Scaling. It is convenient to interpolate the discrete quantities into piecewise continuous functions of the position variable x. We define

$$f(x,k,t) = f_{n+\frac{1}{2}}(k,t), \quad f_I(x,k,t) = (f_I)_{n+\frac{1}{2}}(k,t), \quad x \in [n\ell,(n+1)\ell), \tag{4.1}$$

and

$$(R,T,\tau_R,\tau_T)(x,k) = (R^n,T^n,\tau_R^n,\tau_T^n)(k), \quad x \in [x_{n-\frac{1}{2}},x_{n+\frac{1}{2}}). \tag{4.2}$$

Similarly,

$$\delta\phi(x) = \delta\phi^n,\ k_\pm(x,k) = k_\pm^n(k), \quad x \in [x_{n-\frac{1}{2}},x_{n+\frac{1}{2}}). \tag{4.3}$$

With these definitions, system (3.8)–(3.10) is equivalent to the following system: for $k > 0$,

$$\begin{aligned} f\left(x+\frac{\ell}{2},k,t\right) &= T(x,k_-(x,k))f\left(x-\frac{\ell}{2},k_-(x,k),t-\tau_T(x,k_-(x,k))\right) \\ &\quad + R(x,-k)f\left(x+\frac{\ell}{2},-k,t-\tau_R(x,-k)\right), \end{aligned} \tag{4.4}$$

and for $k < 0$,

$$\begin{aligned} f\left(x-\frac{\ell}{2},k,t\right) &= T(x,-k_+(x,k))f\left(x+\frac{\ell}{2},-k_+(x,k),t-\tau_T(x,-k_+(x,k))\right) \\ &\quad + R(x,-k)f\left(x-\frac{\ell}{2},-k,t-\tau_R(x,-k)\right), \end{aligned} \tag{4.5}$$

with the initial condition

$$f(x,k,t) = f_I(x,k,t) \qquad \forall t \le 0,\ \forall (x,k) \in \mathbb{R}^2. \tag{4.6}$$

The electron concentration takes the form

$$\rho^-(x,t) = \int_{k>0} f(x,k,t)|\psi_k^n(x,t)|^2 \frac{dk}{\pi} + \int_{k<0} f(x+\ell,k,t)|\psi_k^n(x,t)|^2 \frac{dk}{\pi} \quad \forall x \in [x_{n-\frac{1}{2}}, x_n), \tag{4.7}$$

$$\rho^-(x,t) = \int_{k>0} f(x-\ell,k,t)|\psi_k^n(x,t)|^2 \frac{dk}{\pi} + \int_{k<0} f(x,k,t)|\psi_k^n(x,t)|^2 \frac{dk}{\pi} \quad \forall x \in [x_n, x_{n+\frac{1}{2}}). \tag{4.8}$$

We now introduce macroscopic position and time coordinates according to

$$\tilde{x} = \alpha x, \qquad \tilde{t} = \alpha^2 t, \tag{4.9}$$

where $\alpha \ll 1$ is a small parameter describing the ratio between the superlattice cell and the typical size of the device. The time scaling characterizes diffusion phenomena.

Now we make the following assumption, which is classical in the homogenization literature [9]. We suppose that the self-consistent potential can be written as a function of the macroscopic variables $(\tilde{x}, \tilde{t})$ and of the microscopic variable x, and that it is periodic with respect to the latter:

$$\phi_{SC}(x,t) = \tilde{\phi}_{SC}(\tilde{x}, x, \tilde{t}), \tag{4.10}$$

where $\tilde{\phi}_{SC}$ is periodic with respect to its second argument. Indeed, the self-consistent potential may have large scale variations (which extend over the entire structure width, like, e.g., the external potential) as well as subcell variations (related, e.g., to the variations of the electron and ion concentrations or to the dielectric constants). By supposing that the small scale variations are periodic, we assume that a departure from periodicity can occur only on large scales or, equivalently, with small gradients relative to the period. This is the key hypothesis that allows the derivation of a macroscopic regime.

The dielectric constant ϵ as well as the superlattice potential ϕ_{SL} are purely periodic functions and therefore depend only on the periodic variable x: $\epsilon(x)$, $\phi_{SL}(x)$. We could consider large scale variations of these quantities as well, but we discard them here for the sake of clarity.

We now investigate how this assumption translates onto the Schrödinger scattering problem. Using that ϕ_{SL} is periodic and vanishes at the boundary of the elementary cell, and that ϕ_{SC} is periodic with respect to its second argument, the potential (3.3) involved in the Schrödinger equation (3.4) now reads

$$\bar{\phi}^n(x) = \begin{cases} \phi_{SL}(x) + \tilde{\phi}_{SC}(\alpha x, x, \alpha^2 t) & \forall x \in [x_{n-\frac{1}{2}}, x_{n+\frac{1}{2}}], \\ \tilde{\phi}_{SC}(\alpha x_{n-\frac{1}{2}}, x_{-\frac{1}{2}}, \alpha^2 t) & \forall x \le x_{n-\frac{1}{2}}, \\ \tilde{\phi}_{SC}(\alpha x_{n+\frac{1}{2}}, x_{-\frac{1}{2}}, \alpha^2 t) & \forall x \ge x_{n+\frac{1}{2}}. \end{cases} \tag{4.11}$$

We consider this potential the following way: for a given value $\tilde{x}$, we look for the cell C_n such that $\alpha\tilde{x} \in C_n$, i.e., $n(\tilde{x}) = [(\tilde{x}/\alpha\ell) + 1/2]$, where $[\cdot]$ denotes the integer part. Then, we solve the Schrödinger equation (3.4) on $\mathbb{R}$ with the potential (4.11), where x is now a variable independent from $\tilde{x}$, the latter appearing only in $n(\tilde{x})$. Therefore, $\tilde{x}$ and $\tilde{t}$ are *frozen variables* in the Schrödinger problem.

Therefore, the wave-functions are such that $\psi_k = \psi_k(\tilde{x}, x, \tilde{t})$ (but ψ_k is *not* periodic with respect to x), and the scattering data satisfy

$$(R,T) = (\tilde{R},\tilde{T})(\tilde{x},k), \quad (\tau_R,\tau_T) = \alpha^2(\tilde{\tau}_R,\tilde{\tau}_T)(\tilde{x},k),$$
$$\delta\phi = \widetilde{\delta\phi}(\tilde{x},\tilde{t}), \quad k_\pm = \tilde{k}_\pm(\tilde{x},k,\tilde{t}).$$

The factor α^2 in front of the time delays follows from the scaling of time (see the last remark of section 3).

By (4.1), the distribution functions are constant over the period width and therefore do not depend on the fast variable x. However, the electron concentration, because it depends on the wave-functions according to (3.11), may depend on x. Similarly, the ion concentration may have variations within a period, and it also depends on x. Both the distribution function and the concentrations are assumed to be small, of order α^2. This hypothesis is required for consistency with (4.10): if the charge concentration is not small, strong potential gradients can occur on small scales, which we want to avoid. Therefore, we assume the following:

$$f(x,k,t) = \alpha^2 \tilde{f}^\alpha(\tilde{x},k,\tilde{t}),$$
$$(\rho^-,\rho^+,\rho)(x,t) = \alpha^2(\tilde{\rho}^{-\alpha},\tilde{\rho}^{+\alpha},\tilde{\rho}^\alpha)(\tilde{x},x,\tilde{t}).$$

We shall now rename the variables to deal with nicer notations: the periodic variable x will be denoted by y, while the rescaled position and time variables $\tilde{x}$ and $\tilde{t}$ will simply be denoted by x and t. The tildes will be dropped, and the dependence on the scaling parameter α will be recalled only when necessary. With this scaling, the problem (4.4), (4.5), (3.1), (3.11) becomes: for $k > 0$,

$$f\left(x+\alpha\frac{\ell}{2},k,t\right) = R(x,-k)f\left(x+\alpha\frac{\ell}{2},-k,t-\alpha^2\tau_R(x,-k)\right) + T(x,k_-(x,k))f\left(x-\alpha\frac{\ell}{2},k_-(x,k),t-\alpha^2\tau_T(x,k_-(x,k))\right), \tag{4.12}$$

and for $k < 0$,

$$f\left(x-\alpha\frac{\ell}{2},k,t\right) = R(x,-k)f\left(x-\alpha\frac{\ell}{2},-k,t-\alpha\tau_R(x,-k)\right) + T(x,-k_+(x,k))f\left(x+\alpha\frac{\ell}{2},-k_+(x,k),t-\alpha^2\tau_T(x,-k_+(x,k))\right), \tag{4.13}$$

with the initial condition

$$f(x,k,t) = f_I(x,k,t) \qquad \forall t \le 0,\ \forall (x,k) \in \mathbb{R}^2. \tag{4.14}$$

The Poisson equation reads

$$\left(\alpha\frac{\partial}{\partial x}+\frac{\partial}{\partial y}\right)\left(\epsilon(y)\left(\alpha\frac{\partial}{\partial x}+\frac{\partial}{\partial y}\right)\phi_{SC}\right) = \alpha^2\rho(x,y,t), \tag{4.15}$$

$$\rho(x,y,t) = e(\rho^+(x,y)-\rho^-(x,y,t)). \tag{4.16}$$

The electron concentration is given by

$$\forall y \in [x_{-\frac{1}{2}},0), \tag{4.17}$$
$$\rho^-(x,y,t) = \int_{k>0} f(x,k,t)|\psi_k(x,y,t)|^2\frac{dk}{\pi} + \int_{k<0} f(x+\alpha\ell,k,t)|\psi_k(x,y,t)|^2\frac{dk}{\pi},$$

$$\forall y \in [0,x_{\frac{1}{2}}), \tag{4.18}$$
$$\rho^-(x,y,t) = \int_{k>0} f(x-\alpha\ell,k,t)|\psi_k(x,y,t)|^2\frac{dk}{\pi} + \int_{k<0} f(x,k,t)|\psi_k(x,y,t)|^2\frac{dk}{\pi}.$$

The goal of the present paper is to find the formal limit of model (4.12)–(4.18) when α tends to zero. We note that, in this problem, all quantities depend on α. We shall use a superscript α when we want to emphasize this dependence. We state the main theorem in the next section.

5. The diffusion limit. The aim of this paper is to prove the following theorem.

THEOREM 5.1 (formal). *Assume that R, the reflection coefficient associated with potential ϕ_{SL}, is nonzero almost everywhere. Then, in the limit $\alpha \to 0$, the solution f^α, ϕ^α_{SC}, $\rho^{-,\alpha}$ converges, at least formally, to f, ϕ_{SC}, ρ^- such that*

(i) *There exists a function $F(x,\varepsilon(k),t)$, which depends only on k through the energy $\varepsilon(k)$, such that $f(x,k,t) = F(x,\varepsilon(k),t)$. Furthermore, F satisfies the following diffusion equation (SHE model):*

$$M(\varepsilon)\frac{\partial F}{\partial t} + \left(\frac{\partial}{\partial x} - eE\frac{\partial}{\partial \varepsilon}\right) J = 0, \tag{5.1}$$

$$J(x,\varepsilon,t) = -D(\varepsilon)\left(\frac{\partial}{\partial x} - eE\frac{\partial}{\partial \varepsilon}\right) F, \tag{5.2}$$

$$F(x,\varepsilon,t=0) = F_I(x,\varepsilon,0), \tag{5.3}$$

where $E = -\partial\phi_{SC}/\partial x$ is the self-consistent electric field. $M(\varepsilon)$ and $D(\varepsilon)$ are defined by

$$M(\varepsilon) = \frac{1}{\ell}[T(k)\tau_T(k) + R(k)\tau_R(k)], \tag{5.4}$$

$$D(\varepsilon) = \frac{\ell}{2}\frac{T(k)}{R(k)}, \tag{5.5}$$

where $k = \sqrt{2m\varepsilon}/\hbar$ and T, R, τ_T, τ_R refer to the scattering data of the Schrödinger problem with potential $\phi = \phi_{SL}$ on the cell C_0 only.

(ii) *$\phi_{SC} = \phi_{SC}(x,t)$ does not depend on the fast variable y and is a solution of the Poisson equation*

$$-\frac{d}{dx}\left(\bar{\epsilon}\frac{d\phi_{SC}}{dx}\right) = \rho(x,t), \tag{5.6}$$

where the charge concentration $\rho(x,t)$ is given by

$$\rho(x,t) = e(\rho^+(x) - \rho^-(x,t)),$$
$$\rho^+(x) = \int_{-\ell/2}^{\ell/2} \rho^+(x,y)\frac{dy}{\ell}\ , \qquad \rho^-(x,t) = \int_{-\ell/2}^{\ell/2} \rho^-(x,y,t)\frac{dy}{\ell},$$

and the average dielectric constant $\bar{\epsilon}$ by

$$\bar{\epsilon}^{-1} = \int_{-\ell/2}^{\ell/2} \frac{1}{\epsilon(y)}\frac{dy}{\ell}.$$

(iii) *The positive ion concentration $\rho^+(x,y)$ is a datum; the electron concentration is given by*

$$\rho^-(x,y,t) = \int_{k\in\mathbb{R}} F(x,\varepsilon(k),t)|\psi_k(x,y,t)|^2\frac{dk}{\pi}. \tag{5.7}$$

The coefficients M and D are, respectively, referred to as the "density-of-states" and the "diffusivity."

Let us comment on boundary conditions. The boundary conditions (3.2) may also be used in conjunction with the Poisson equation (5.6). However, in order to account for boundary layers in the homogenization process, more accurate boundary conditions can be imposed, of Robin (or Fourier) type, namely,

$$(\phi_{SC} - \alpha\lambda_a\phi'_{SC})(a,t) = 0, \qquad (\phi_{SC} + \alpha\lambda_b\phi'_{SC})(b,t) = \phi_{bias}, \tag{5.8}$$

where λ_a, λ_b are positive extrapolation lengths (for expressions of λ_a and λ_b, see, e.g., [9]) and α is the scaling parameter (4.9). Similarly, considerations of kinetic layers lead to the following boundary conditions for F:

$$(F + \alpha\Lambda_a J)(a,\varepsilon,t) = F_a(\varepsilon,t), \qquad (F - \alpha\Lambda_b J)(b,\varepsilon,t) = F_b(\varepsilon,t), \tag{5.9}$$

where $F_a(\varepsilon,t) = f_a(k,t)$, $F_b(\varepsilon,t) = f_b(-k,t)$, and $\Lambda_{a,b} > 0$. We refer to [19] for the theory of boundary layers in the framework of the SHE model and for the computation of $\Lambda_{a,b}$. We shall not dwell on this point here.

The proof of Theorem 5.1 proceeds in three steps. The first consists of showing that f^α formally converges to a function of $(x,\varepsilon(k),t)$ only. The second and third steps correspond to the derivations of the continuity and current equations (5.1), (5.2). To achieve these goals, two methods can be utilized: the Hilbert expansion method [2], [17] and the moment method [24]. We shall choose the latter because it involves more straightforward computations. We shall defer the details of the proof to the appendix. We shall take the existence of solutions for the original discrete model (4.12)–(4.18) as well as the convergence of f^α, ϕ^α_{SC}, $\rho^{-,\alpha}$ for granted, and we shall focus solely on the establishment of the limit model. In fact, proving convergence is a very challenging mathematical problem which is far beyond the scope of the present paper. In the next sections, we comment further on the model obtained and deal with some practical examples.

6. Comments on the diffusion model (5.1)–(5.2). We refer to the introduction for references on the SHE model (5.1)–(5.2). This model is of great practical interest for semiconductor device simulations because it provides information about the electron energy distribution function at a much lower cost than a Monte Carlo simulation of the Boltzmann equation [22], [23]. To our knowledge, the present paper provides the first derivation of this model in the framework of superlattices, when the diffusion is induced by the scattering properties of the quantum potential structure itself.

Another interesting property of this model is that it gives rise to a hierarchy of moment models [4], [17], including the usual drift-diffusion model [28], [33] and the so-called energy-transport model, which is an extension of the drift-diffusion model with an additional energy balance equation (see [6] and references therein). We refer to [4], [17] for the derivation of these models. Here, we first want to discuss the relation between the number and energy densities $n(x,t)$ and $\mathcal{E}(x,t)$ in position space, which are obviously important macroscopic quantities, and the energy distribution F. This relation takes the form

$$\begin{pmatrix} n \\ \mathcal{E} \end{pmatrix} = \int_{\mathcal{R}} F(x,\varepsilon,t) \begin{pmatrix} 1 \\ \varepsilon \end{pmatrix} M(x,\varepsilon)d\varepsilon, \tag{6.1}$$

where, in drift-diffusion or energy-transport models, F is approximated by a Fermi–Dirac distribution (i.e., the quantum thermodynamical equilibrium distribution function of electrons; see [11]) $F_{\mu,T}(\varepsilon) = (\exp((\varepsilon-\mu)/k_BT)+1)^{-1}$. The thermodynamic parameters $\mu = \mu(x,t)$ and $k_BT = k_BT(x,t)$ are the chemical potential and thermal energy and characterize the state of the electron gas. (6.1) furnishes a local relationship between the pairs $(n,\mathcal{E})$ and (μ,T).

Here, the particular form of the density-of-states M makes this relationship fairly different than in bulk semiconductor materials. M can be viewed as an "averaged density-of-states" over the superlattice elementary period. It reduces to the usual one-dimensional density-of-states (up to a constant factor) $M = v(k)^{-1}$ if the potential ϕ_{SL} is constant in the elementary cell (since in this case $T = 1$, $R = 0$, and τ_T is equal to the classical transit time $\ell/v(k)$). But M may significantly differ from this value if ϕ_{SL} is not constant (see section 7).

Because of this particular form of M, the relationship between $(n,\mathcal{E})$ and (μ,T) in the superlattice can significantly depart from that of bulk materials. In particular, at resonant energies (see the example of square potentials below) the time delays are significantly longer than the classical time delays. Symmetrically, at energies away from resonances, they can be shorter. Therefore, in (6.1), M weights the resonant energies more strongly and the nonresonant ones less strongly. The relation $(\mu,T) \to (n,\mathcal{E})$ reflects these effects.

Next we discuss the value of the electron mobility in the superlattice, given by [4], [17]:

$$\mu_{\mathrm{SL}} = \frac{e}{nk_BT}\int_{\mathcal{R}} D(x,\varepsilon)F_{\mu,T}(1-F_{\mu,T})d\varepsilon. \tag{6.2}$$

We recall that the mobility is the coefficient of Ohm's law $j_n = \mu_{\mathrm{SL}} nE$, which is the expression of the drift-diffusion law when the chemical potential μ and the temperature T are constant in space. The mobility is one of the most important transport parameters in semiconductors as it is easily accessible to measurements and characterizes the ability of electrons to react to an external electric field. In the present case, the mobility of the electrons is induced solely by their scattering by the superlattice potential pattern. A realistic expression of the mobility must also include the influence of "bulk interactions" like phonon or impurity interactions. As already pointed out, this is not yet done in full rigor in the present work, but a rough estimate of the total mobility μ_{tot} can be obtained from

$$\frac{1}{\mu_{\mathrm{tot}}} = \frac{1}{\mu_{\mathrm{bulk}}} + \frac{1}{\mu_{\mathrm{SL}}}, \tag{6.3}$$

where μ_{bulk} is the mobility under the influence of bulk collisions alone (see [32], [16]). Formula (6.3) can be understood by a circuit analogy, in which the resistances (proportional to the reciprocal of the mobility) of the bulk and of the superlattice add up in series. It provides a first answer to the problem of determining the electron mobility in the superlattice. This question was left open, for instance, in [35], where by default, the superlattice mobility was assumed to be equal to that of the bulk.

As a conclusion, the diffusion model (5.1)–(5.2) can be used in two ways: either for direct simulations or as a way to access analytical values for the parameters of macroscopic (drift-diffusion or energy-transport) models. In the latter case, the analytical values of M and D give rise to easily computable values of the drift-diffusion mobilities or density versus chemical potential relationships. (See the example in the

next section.) Through the specific values (5.4), (5.5), the obtained parameters of the drift-diffusion model retain information about the quantum nature of transport in the superlattice base cell in a rigorous way.

7. Example: Square wells or barriers. In this section, we analyze the simple case in which ϕ_{SL} is equal to a simple square well or barrier. More precisely, we suppose that the presence of a second material (called B) results in potential and effective mass jumps at the $A-B$ interfaces. We denote by m_A and m_B the effective masses in materials A and B. We denote by V the potential ϕ_{SL} in material B, with $V>0$ if B produces an (energy) potential well, and $V<0$ if it produces a potential barrier. We assume that the B layer has width a. Therefore,

$$\phi_{SL}(x)=\begin{cases} V, & x\in\left(-\frac{a}{2},\frac{a}{2}\right),\\ 0, & x\notin\left(-\frac{a}{2},\frac{a}{2}\right),\end{cases}\qquad m(x)=\begin{cases} m_B, & x\in\left(-\frac{a}{2},\frac{a}{2}\right),\\ m_A, & x\notin\left(-\frac{a}{2},\frac{a}{2}\right).\end{cases}\tag{7.1}$$

We note that

$$\varepsilon(k)=\frac{\hbar^2k^2}{2m_A},\quad \kappa=\left(\frac{m_B}{m_A}(k^2+k_V^2)\right)^{1/2},\quad k_V^2=\frac{2em_A}{\hbar^2}V.\tag{7.2}$$

Note that $k_V^2>0$ and $\kappa^2>k^2$ in the case of a well, and $k_V^2<0$, $\kappa^2<k^2$ in the case of a barrier. In the barrier case, κ^2 can be negative, in which case κ is a pure imaginary number $\kappa=i\tilde{\kappa}$.

The resolution of (3.4) for the scattering states in this case is an elementary quantum mechanics problem, the solution of which can be found in any textbook (see [29], for instance). Here we just note that, to account for the effective mass discontinuity, the continuity of ψ and $m^{-1}\psi'$ must be enforced at the interfaces $x=\pm a/2$.

With (7.1), we find for a well ($k_V^2>0$) and any value of $k>0$, or for a barrier ($k_V^2<0$) and incident energies above the barrier energy $k^2+k_V^2>0$, that

$$T=\frac{1}{1+\chi_0\sin^2(\kappa a)},\qquad \chi_0=\frac{\left((m_B-m_A)k^2-m_Ak_V^2\right)^2}{4m_Am_Bk^2(k^2+k_V^2)}\tag{7.3}$$

and

$$\tau_T=\tau_R:=\tau_{sq}=\frac{1}{v(k)}(\ell-a+d_{sq}),\tag{7.4}$$

with

$$d_{sq}=\frac{a\chi_1(1+\cot^2(\kappa a))-\kappa^{-1}\chi_2\cot(\kappa a)}{\chi_3+\cot^2(\kappa a)},\qquad \chi_1=\frac{(m_A+m_B)k^2+m_Ak_V^2}{2m_A(k^2+k_V^2)},$$
$$\chi_2=\frac{\left((m_A-m_B)k^2+m_Ak_V^2\right)k_V^2}{2m_A(k^2+k_V^2)k^2},\qquad \chi_3=\frac{\left((m_A+m_B)k^2+m_Ak_V^2\right)^2}{4m_Am_Bk^2(k^2+k_V^2)}.$$

The value of d_{sq} has to be compared to the corresponding value d_{cl} for classical motion:

$$d_{cl}=\frac{k}{m_A}\frac{m_B}{(k^2+k_V^2)^{1/2}}a.$$

At resonant energies (i.e., when $\sin(\kappa a)=0$), $\tau_{sq}>\tau_{cl}$, while the reverse inequality holds at nonresonant energies (i.e., when $\cos(\kappa a)=0$); see [29].

In the case of a barrier ($k_V^2 < 0$) and incident energies below the barrier energy $k^2 + k_V^2 < 0$, we find

$$T = \frac{1}{1 + \chi_0 \sinh^2(\tilde{\kappa} a)}, \qquad d_{sq} = \frac{a\chi_1(\coth^2(\tilde{\kappa} a) - 1) + \tilde{\kappa}^{-1}\chi_2 \coth(\tilde{\kappa} a)}{\chi_3 + \coth^2(\tilde{\kappa} a)},$$

with χ_0, χ_1, and χ_3 obtained by changing $k^2 + k_V^2$ into $|k^2 + k_V^2|$ in the denominators of the formulas above and with χ_2 unchanged.

With (7.3) and (7.4), equations (5.4), (5.5) give the expression of the density-of state and diffusivity in the case of square wells or barriers:

$$D = \frac{\ell}{\hbar\pi} \frac{1}{\chi_0 \sin^2(\kappa a)} \quad \text{for } \kappa^2 > 0, \tag{7.5}$$

$$D = \frac{\ell}{\hbar\pi} \frac{1}{\chi_0 \sinh^2(\tilde{\kappa} a)} \quad \text{for } \kappa^2 = -\tilde{\kappa}^2 < 0, \tag{7.6}$$

$$M = \frac{2\tau_{sq}}{\hbar\pi\ell}. \tag{7.7}$$

As already pointed out, the diffusivity has infinite peaks at resonant energies ($\sin(\kappa a) = 0$). These peaks are due to resonant tunneling and are reminiscent of the conduction minibands in an infinite superlattice.

If more than one well or one barrier is present within a period (i.e., when the coherence length is larger than the period), the scattering data become increasingly complicated (but still can be computed numerically) and resemble even more closely the energy band structure of the infinite superlattice.

This example shows that the coefficients M and D of the diffusion model can be computed easily from the resolution of the Schrödinger equation in a single cell. The most important characteristics of quantum transport (like the existence of transmission resonances) translate into specific behaviors (like singularities at resonance energy) of these constants.

8. Conclusion. In this paper, we have presented a scattering matrix model describing electron transport in semiconductor superlattices when the electron phase coherence length is of the order of the superlattice period. Then, we have investigated the limit of a large number of superlattice cells. We have shown that, at the diffusion time scale, the scattering matrix model formally converges to a diffusion model in the position-energy space, the so-called SHE model, and have explained how it can be used to model electron transport in superlattices. In particular, the model takes into account the electric potential self-consistency through a quantum model of the charge concentration. An analytical example shows how the diffusion constants can be explicitly computed from the scattering data of the base cell.

The scientific merit of this model can only be assessed in light of numerical simulations and comparisons with other experiments or numerical models. In the present paper, we have focused on the derivation of the diffusion model and will defer its numerical validation to future work. However, we feel that the present model has promising capabilities. It will be easy to solve numerically, while giving access to finer physical details than conventional drift-diffusion models. In particular, we have shown how the diffusion constants retain some of the essential features of quantum transport in the superlattice base cell. Beyond their use in the SHE model, these diffusion constants can be used in turn to improve our knowledge of diffusion constants of conventional models of the drift-diffusion type when applied to superlattices.

Appendix. Proof of Theorem 5.1. We shall give only the formal calculations below, supposing that all unknowns have a limit in a convenient function space. The proof of this point is a very difficult one (given the nonlinearity of the problem and its highly oscillating character) and will not be tackled here.

A.1. The potential and the Schrödinger equation. It is a classical matter in homogenization theory [9] that

$$\phi^\alpha_{SC}(x,y,t) = \phi_{SC}(x,t) + O(\alpha) \quad \text{as } \alpha \to 0,$$

where $\phi_{SC}(x,t)$ does not depend on y and satisfies the homogenized Poisson equation (5.6).

Then, standard perturbation theory for the Schrödinger equation applies (see [31]), and we get, at least formally,

$$\psi^\alpha_k(x,y,t) \to \psi_k(x,y,t), \quad R^\alpha, T^\alpha, \tau^\alpha_R, \tau^\alpha_T \to R, T, \tau_R, \tau_T, \dots,$$

where ψ_k, R, T, τ_R, τ_T are the wave-function and scattering data associated with the potential $\phi = \phi_{SC}(x,t) + \phi_{SL}(y)$. However, since ϕ_{SC} does not depend on y, the wave-function is deduced from the solution of the Schrödinger equation with potential $\phi_{SL}(y)$ by multiplication by a constant (in y) phase factor (Gauge transformation), and the scattering data are unchanged. Furthermore, a simple coordinate translation allows us to consider the problem in the elementary cell C_0 again without changing the scattering data. Therefore, R, T, τ_R, τ_T do not depend on x. From now on, ψ_k will denote the wave-function associated with the cell C_0, which also does not depend on x.

Now we estimate the potential shift $\delta\phi^\alpha$ across a cell. Following section 4, we consider a point x and let $n = n^\alpha(x) = [(x/\alpha\ell) + 1/2]$, where $[\cdot]$ is the integer part. According to (4.11), across the cell $C_{n^\alpha(x)}$ associated with x, $\delta\phi^\alpha$ is equal to

$$\delta\phi^\alpha = \phi^\alpha_{SC}(\alpha x_{n+\frac12}, x_{-\frac12}, t) - \phi^\alpha_{SC}(\alpha x_{n-\frac12}, x_{-\frac12}, t).$$

We can define δ_- and δ_+ such that

$$\alpha x_{n-\frac12} = \alpha\left(n - \frac12\right)\ell = x + \alpha\delta_-, \qquad \alpha x_{n+\frac12} = \alpha\left(n + \frac12\right)\ell = x + \alpha\delta_+.$$

We obviously have $-\ell \le \delta_- \le 0$, $0 \le \delta_+ \le \ell$, $\delta_+ - \delta_- = \ell$. Therefore,

$$\delta\phi^\alpha = \phi^\alpha_{SC}(x + \alpha\delta_+, x_{-\frac12}, t) - \phi^\alpha_{SC}(x + \alpha\delta_-, x_{-\frac12}, t) = \alpha\ell\frac{\partial\phi^\alpha_{SC}}{\partial x} + O(\alpha^2). \tag{A.1}$$

In particular, $\delta\phi^\alpha \to \delta\phi = 0$. From this, we note that

$$k^\alpha_\pm(x,k) = |k| \pm \alpha\frac{em\ell}{\hbar^2}\frac{1}{|k|}\frac{\partial\phi^\alpha_{SC}}{\partial x} + O(\alpha^2). \tag{A.2}$$

Finally, taking the limit $\alpha \to 0$ in (4.17) and (4.18) leads to (5.7), provided that f is a function of $\varepsilon(k)$ only. This point is proved in the next section.

A.2. f depends only on the energy. We consider problem (4.12), (4.13) and formally let $\alpha \to 0$. We have, using (A.2),

$$f(x,k,t) = R(-k)f(x,-k,t) + T(k)f(x,k,t), \quad k > 0. \tag{A.3}$$

Now, considering that R and T are the scattering data associated with the limit potential, which has no potential shift ($\delta\phi = 0$), formulas (3.6) lead to

$$R + T = 1, \quad R, T(k) = R, T(-k), \quad \tau_R, \tau_T(k) = \tau_R, \tau_T(-k) \qquad \forall k \in \mathbb{R}. \tag{A.4}$$

Using these relations, (A.3) leads to

$$R(k)(f(x,k,t) - f(x,-k,t)) = 0, \quad k > 0. \tag{A.5}$$

With the assumption that $R > 0$ almost everywhere, we deduce that f is even with respect to k or, equivalently, that it is a function of $\varepsilon(k)$. Therefore, we have $f(x,k,t) = F(x,\varepsilon(k),t)$.

A.3. Current equation (5.2). We introduce the current

$$J^\alpha(x,\varepsilon,t) = \frac{1}{2\alpha}\left(f^\alpha(x,k,t) - f^\alpha(x,-k,t)\right), \quad k = \frac{\sqrt{2\varepsilon m}}{\hbar}. \tag{A.6}$$

We show that $J^\alpha \to J$, where J is given by (5.2).

For $k > 0$, we rewrite (4.12) by shifting the position variable by a half-period:

$$\begin{aligned} f^\alpha(x,k,t) = {} & R^\alpha\left(x - \alpha\frac{\ell}{2}, -k\right) f^\alpha(x,-k,t-\alpha^2\tau_R^\alpha) \\ & + T^\alpha\left(x - \alpha\frac{\ell}{2}, k_-^\alpha\right) f^\alpha(x-\alpha\ell, k_-^\alpha, t-\alpha^2\tau_T^\alpha), \end{aligned} \tag{A.7}$$

where we have not repeated the arguments of k_-^α and $\tau_{R,T}^\alpha$. We rewrite (A.7) according to

$$\begin{aligned} f^\alpha(x,k,t) = {} & R^\alpha\left(x - \alpha\frac{\ell}{2}, -k\right) f^\alpha(x,-k,t) + T^\alpha\left(x - \alpha\frac{\ell}{2}, k_-^\alpha\right) f^\alpha(x,k_-^\alpha,t) \\ & + T^\alpha\left(x - \alpha\frac{\ell}{2}, k_-^\alpha\right)\left(-\alpha\ell\frac{\partial f^\alpha}{\partial x}(x,k,t) + (k_-^\alpha - k)\frac{\partial f^\alpha}{\partial k}(x,k,t)\right) + O(\alpha^2). \end{aligned} \tag{A.8}$$

We note that, by (3.6), $T^\alpha(x - \alpha\frac{\ell}{2}, k_-^\alpha) = T^\alpha(x - \alpha\frac{\ell}{2}, -k)$. Now, using (A.2), we deduce from (A.8) that

$$\begin{aligned} & R^\alpha\left(x - \alpha\frac{\ell}{2}, -k\right)\left(f^\alpha(x,k,t) - f^\alpha(x,-k,t)\right) \\ & \quad = -\alpha\ell T^\alpha\left(x - \alpha\frac{\ell}{2}, k_-^\alpha\right)\left(\frac{\partial f^\alpha}{\partial x}(x,k,t) + \frac{em}{\hbar^2}\frac{1}{k}\frac{\partial\phi_{SC}^\alpha}{\partial x}\frac{\partial f^\alpha}{\partial k}(x,k,t)\right) + O(\alpha^2), \end{aligned}$$

or, dividing by R^α,

$$J^\alpha(x,\varepsilon(k),t) = -\frac{\ell}{2}\frac{T^\alpha(x - \alpha\frac{\ell}{2}, k_-^\alpha)}{R^\alpha(x - \alpha\frac{\ell}{2}, k_-^\alpha)}\left(\frac{\partial f^\alpha}{\partial x}(x,k,t) + \frac{em}{\hbar^2}\frac{1}{k}\frac{\partial\phi_{SC}^\alpha}{\partial x}\frac{\partial f^\alpha}{\partial k}(x,k,t)\right) + O(\alpha).$$

Now, taking the limit $\alpha \to 0$ and noting that

$$\frac{\partial f}{\partial k} = \frac{\partial F}{\partial\varepsilon}\frac{\hbar^2 k}{m}, \quad k > 0,$$

we easily obtain (5.2).

A.4. Continuity equation (5.1). To prove the continuity equation (5.1), we evaluate

$$I^\alpha = \frac{1}{\alpha\ell}\left(J^\alpha\left(x+\alpha\frac{\ell}{2},\varepsilon,t\right) - J^\alpha\left(x-\alpha\frac{\ell}{2},\varepsilon-e\delta\phi^\alpha,t\right)\right).$$

First, using (A.1) together with a Taylor expansion, we readily see that

$$I^\alpha = \left(\frac{\partial J^\alpha}{\partial x} + e\frac{\partial\phi^\alpha_{SC}}{\partial x}\frac{\partial J^\alpha}{\partial\varepsilon}\right)(x,\varepsilon,t) + O(\alpha).$$

Now, using (4.12), (4.13), we compute

$$\begin{aligned}
2\alpha^2\ell I^\alpha &= T(x,k^\alpha_-)f\left(x-\alpha\frac{\ell}{2},k^\alpha_-,t-\alpha^2\tau^\alpha_T(x,k^\alpha_-)\right)\\
&\quad + R(x,-k)f\left(x+\alpha\frac{\ell}{2},-k,t-\alpha^2\tau^\alpha_R(x,-k)\right) - f\left(x+\alpha\frac{\ell}{2},-k,t\right)\\
&\quad + T(x,-k)f\left(x+\alpha\frac{\ell}{2},-k,t-\alpha^2\tau^\alpha_T(x,-k)\right)\\
&\quad + R(x,k^\alpha_-)f\left(x-\alpha\frac{\ell}{2},k^\alpha_-,t-\alpha^2\tau^\alpha_R(x,k^\alpha_-)\right) - f\left(x-\alpha\frac{\ell}{2},k^\alpha_-,t\right).
\end{aligned}$$

By Taylor expanding with respect to t and using (3.6), we finally get

$$\begin{aligned}
2\alpha^2\ell I^\alpha &= -\alpha^2\left(T(x,k^\alpha_-)\tau^\alpha_T(x,k^\alpha_-) + R(x,-k)\tau^\alpha_R(x,-k)\right.\\
&\qquad \left. + T(x,-k)\tau^\alpha_T(x,-k) + R(x,k^\alpha_-)\tau^\alpha_R(x,k^\alpha_-)\right)\frac{\partial f^\alpha}{\partial t}(x,k,t) + O(\alpha^3).
\end{aligned}$$

Taking the limit $\alpha\to 0$ and using (A.4) then easily leads to (5.1).

This conclude the proof of Theorem 5.1. □

Acknowledgments. The first author warmly thanks B. Vinter for stimulating discussions and encouragement.

REFERENCES

[1] N. W. Ashcroft and N. D. Mermin, *Solid State Physics*, Saunders College Publishing, Fort Worth, TX, 1976.

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

[3] N. Ben Abdallah, *A hybrid kinetic-quantum model for stationary electron transport in resonant tunneling diodes*, J. Statist. Phys., 90 (1998), pp. 627–662.

[4] N. Ben Abdallah and P. Degond, *On a hierarchy of macroscopic models for semiconductors*, J. Math. Phys., 37 (1996), pp. 3306–3333.

[5] N. Ben Abdallah, P. Degond, and I. M. Gamba, *Coupling one-dimensional time-dependent classical and quantum transport models*, J. Math. Phys., 43 (2002), pp. 1–24.

[6] N. Ben Abdallah, P. Degond, and S. Génieys, *An energy-transport model for semiconductors derived from the Boltzmann equation*, J. Statist. Phys., 84 (1996), pp. 205–231.

[7] N. Ben Abdallah, P. Degond, and P. A. Markowich, *On a one-dimensional Schrödinger-Poisson scattering model*, Z. Angew. Math. Phys., 48 (1997), pp. 135–155.

[8] N. Ben Abdallah, P. Degond, A. Mellet, and F. Poupaud, *Electron transport in semiconductor multiquantum well structures*, Quart. Appl. Math., to appear.

[9] A. Bensoussan, J. L. Lions, and G. Papanicolaou, *Asymptotic Analysis for Periodic Structures*, North–Holland, Amsterdam, 1978.

[10] A. Bensoussan, J. L. Lions, and G. Papanicolaou, *Boundary layers and homogenization of transport processes*, Publ. Res. Inst. Math. Sci. Kyoto Univ., 15 (1979), pp. 53–157.
[11] J. S. Blakemore, *Semiconductor Statistics*, Pergamon, Oxford, U.K., 1962.
[12] L. L. Bonilla, J. Galaàn, J. A. Cuesta, F. C. Martínez, and J. M. Molera, *Dynamics of electric-field domains and oscillations of the photocurrent in a simple superlattice model*, Phys. Rev. B, 50 (1994), pp. 8644–8657.
[13] E. Bringuier, *Kinetic theory of high-field transport in semiconductors*, Phys. Rev. B, 57 (1995), pp. 2280–2285.
[14] C. Cercignani, *The Boltzmann Equation and Its Applications*, Springer, New York, 1998.
[15] A. Chomette and J. F. Palmier, Solid State Commun., 43 (1982), p. 157.
[16] E. M. Conwell, *High-Field Transport in Semiconductor*, Solid State Physics, Vol. G, Academic Press, New York, 1967.
[17] P. Degond, *Mathematical modeling of microelectronics semiconductor devices*, AMS/IP Studies in Advanced Mathematics 15, AMS Society and International Press, 2000, pp. 77–109.
[18] P. Degond, *A model of near-wall conductivity and its application to plasma thrusters*, SIAM J. Appl. Math., 58 (1998), pp. 1138–1162.
[19] P. Degond and C. Schmeiser, *Kinetic boundary layers and fluid-kinetic coupling in semiconductors*, Transport. Theory Statist. Phys., 28 (1999), pp. 31–55.
[20] P. Dmitruk, A. Saul, and L. Reyna, *High electric field approximation in semiconductor devices*, Appl. Math. Lett., 5 (1992), pp. 99–102.
[21] L. Esaki and R. Tsu, *Superlattice and negative differential conductivity in semiconductors*, IBM J. Res. Develop., 14 (1970), p. 61.
[22] A. Gnudi, D. Ventura, G. Baccarani, and F. Odeh, *Two-dimensional MOSFET simulation by means of a multidimensional spherical harmonic expansion of the Boltzmann transport equation*, Solid State Electron., 36 (1993), pp. 575–581.
[23] N. Goldsman, L. Henrickson, and J. Frey, *A physics-based analytical numerical solution to the Boltzmann transport equation for use in device simulation*, Solid State Electron., 34 (1991), pp. 389–396.
[24] F. Golse and F. Poupaud, *Limite fluide des équations de Boltzmann des semiconducteurs pour une statistique de Fermi-Dirac*, Asymptot. Anal., 6 (1992), pp. 135–160.
[25] H. T. Grahn, ed., *Semiconductor Superlattices, Growth and Electronic Properties*, World Scientific, Singapore, 1995.
[26] B. Lapeyre, E. Pardoux, and R. Sentis, *Méthodes de Monte-Carlo pour les Équations de Transport et de Diffusion*, Springer, Berlin, 1998.
[27] P. Markowich, N. Mauser, and F. Poupaud, *A Wigner function approach to (semi)classical limits: Electrons in a periodic potential*, J. Math. Phys., 35 (1994), pp. 1066–1094.
[28] P. A. Markowich, C. Ringhofer, and C. Schmeiser, *Semiconductor Equations*, Springer, Vienna, 1990.
[29] A. Messiah, *Quantum Mechanics* I *and* II, Halsted, New York, 1961, 1962.
[30] F. Poupaud, *Diffusion approximation of the linear semiconductor equation: Analysis of boundary layers*, Asymptot. Anal., 4 (1991), pp. 293–317.
[31] M. Reed and B. Simon, *Methods of Modern Mathematical Physics, Vols.* III *(Scattering Theory)* and IV (*Analysis of Operators*), Academic Press, New York, 1979.
[32] D. L. Rode, *Low-field electron transport*, in Semiconductor and Semimetals, Vol. 10, Academic Press, New York, 1967.
[33] S. Selberherr, *Analysis and Simulation of Semiconductor Devices*, Springer, Vienna, 1984.
[34] M. A. Stettler and M. S. Lundstrom, *Self-consistent scattering matrix calculation of the distribution function in semiconductor devices*, Appl. Phys. Lett., 60 (1992), pp. 2908–2910.
[35] L. Thibaudeau, *Théorie et Modélisation de Détecteurs Infrarouge à Puits Quantiques*, Ph.D. dissertation, University Paris 7, Paris, 1995.
[36] D. Ventura, A. Gnudi, G. Baccarani, and F. Odeh, *Multidimensional spherical harmonics expansion of Boltzmann equation for transport in semiconductors*, Appl. Math. Lett., 5 (1992) pp. 85–90.
[37] C. Weisbuch and B. Vinter, *Quantum Semiconductor Structures, Fundamentals and Applications*, Academic Press, Boston, 1991.

SIAM J. APPL. MATH.
Vol. 63, No. 1, pp. 299–323

RELAXATION OSCILLATIONS IN A CLASS OF DELAY DIFFERENTIAL EQUATIONS*

A. C. FOWLER† AND MICHAEL C. MACKEY‡

Abstract. We study a class of delay differential equations which have been used to model hematological stem cell regulation and dynamics. Under certain circumstances the model exhibits self-sustained oscillations, with periods which can be significantly longer than the basic cell cycle time. We show that the long periods in the oscillations occur when the cell generation rate is small, and we provide an asymptotic analysis of the model in this case. This analysis bears a close resemblance to the analysis of relaxation oscillators (such as the Van der Pol oscillator), except that in our case the slow manifold is infinite dimensional. Despite this, a fairly complete analysis of the problem is possible.

Key words. relaxation oscillations, delay differential equations, hematopoiesis, stem cells, chronic myelogenous leukaemia

AMS subject classifications. 34K99, 92A07, 34C15, 34E05

PII. S0036139901393512

1. Introduction.

The understanding of periodic behavior in nonlinear ordinary differential equations is reasonably complete. Near Hopf bifurcation, periodic solutions are generically of small amplitude and can be analyzed using the methods of multiple scales. At more extreme parameter values, oscillations are often strongly nonlinear, and it is frequently the case that the dynamics are relaxational, in which case they can be understood through the existence of slow manifolds in phase space and the associated asymptotic analysis of the resulting relaxation oscillators. The classic example is the relaxation oscillation of the Van der Pol oscillator, whose analysis is ably expounded by Kevorkian and Cole (1981).

The situation is much less satisfactory for delay differential equations, which are frequently used to model populations, for example, in ecology (Gurney, Blythe, and Nisbet (1980)) or physiology (Mackey (1997)). One example is the delay recruitment equation

$$\varepsilon \dot{x} = -x + f(x_1), \tag{1.1}$$

where $x_1 = x(t-1)$. For unimodal f (i.e., $f(0) = 0$, $(x - x^*)f'(x) < 0$ for some $x^* > 0$), periodic oscillations can occur for sufficiently small ε. In some circumstances, a singular perturbation analysis of periodic solutions when $\varepsilon \ll 1$ is possible (Chow and Mallet-Paret (1982); Chow, Lin, and Mallet-Paret (1989)), but the results have been limited in scope.

Although linear and weakly nonlinear stability methods are straightforward for delay differential equations, singular perturbation methods appear difficult to implement in general. Much of the work that has been done, such as Chow and Mallet-Paret's work cited above, is concerned with systems with large delay (thus (1.1) or

*Received by the editors August 6, 2001; accepted for publication (in revised form) March 7, 2002; published electronically September 12, 2002.
http://www.siam.org/journals/siap/63-1/39351.html
†Mathematical Institute, Oxford University, 24-29 St. Giles', Oxford OX1 3LB, England (fowler@maths.ox.ac.uk).
‡Departments of Physiology, Physics and Mathematics and Centre for Nonlinear Dynamics, McGill University, Montreal, QC, Canada (mackey@cnd.mcgill.ca).

its generalizations (Chow and Huang (1994); Hale and Huang (1996))). Artstein and Slemrod (2001) place their discussion of relaxation oscillations in the context of slow and fast manifolds familiar from ordinary differential equations and draw a distinction between systems where the delay is "fast" or "slow." (In this context we will find that the delay in our system is fast.)

Actual constructive asymptotic methods are less common. Fowler (1982) analyzed the delayed logistic equation $\varepsilon\dot{x} = x(1 - x_1)$, and Bonilla and Liñan (1984) analyzed a more general system having distributed delay and with diffusion. In a sequence of papers, Lange and Miura (e.g., 1982, 1984) provided asymptotic analyses of models with delays and exhibited boundary layer behavior, although they were exclusively concerned with boundary value problems, and their systems were linear. More recently, Pieroux et al. (2000) analyzed a laser system when the delay was large but dependence on the delayed variable was weak, using multiple scale techniques. In this paper, we show how a constructive relaxational perturbation analysis can be carried out for a particular class of delay differential equations describing stem cell dynamics, when the net proliferation rate is small.

2. A mathematical model of stem cell dynamics. Hematological diseases are interesting and have attracted a significant amount of modeling attention because a number of them are periodic in nature (Haurie, Dale, and Mackey (1998)). Some of these diseases involve only one blood cell type and are due to the destabilization of peripheral control mechanisms, e.g., periodic auto-immune hemolytic anemia (Bélair, Mackey, and Mahaffy (1995); Mahaffy, Bélair, and Mackey (1998)) and cyclical thrombocytopenia (Swinburne and Mackey (2000); Santillan et al. (2000)). Typically, periodic hematological diseases of this type involve periodicities between two and four times the bone marrow production/maturation delay (which is different from the delay considered in this paper).

Other periodic hematological diseases involve oscillations in all of the blood cells (white cells, red blood cells, and platelets). Examples include cyclical neutropenia (Haurie, Dale, and Mackey (1999); Haurie et al. (1999); Haurie et al. (2000)) and periodic chronic myelogenous leukemia (Fortin and Mackey (1999)). These diseases involve very long period dynamics (on the order of weeks to months) and are thought to be due to a destabilization of the pluripotential stem cell (PPSC) compartment from which all of these mature blood cell types are derived.

In Figure 2.1 we have given a pictorial representation of the PPSC compartment and defined the important variables. The dynamics of this PPSC population are governed (Mackey (1978), (1997), (2001)) by the pair of coupled differential delay equations

$$\frac{dP}{d\hat{t}} = -\gamma P + \beta(N)N - e^{-\gamma\tau}\beta(N_\tau)N_\tau \tag{2.1}$$

for the dynamics of the proliferating phase cells and

$$\frac{dN}{d\hat{t}} = -[\beta(N) + \delta]N + 2e^{-\gamma\tau}\beta(N_\tau)N_\tau \tag{2.2}$$

for the nonproliferating (G_0) phase cells. In these equations, $\hat{t}$ is time, τ is the time required for a cell to traverse the proliferative phase, $N_\tau = N(\hat{t} - \tau)$, and the resting to proliferative phase feedback rate β is taken to be a Hill function of the form

$$\beta(N) = \frac{\beta_0\theta^n}{\theta^n + N^n}. \tag{2.3}$$

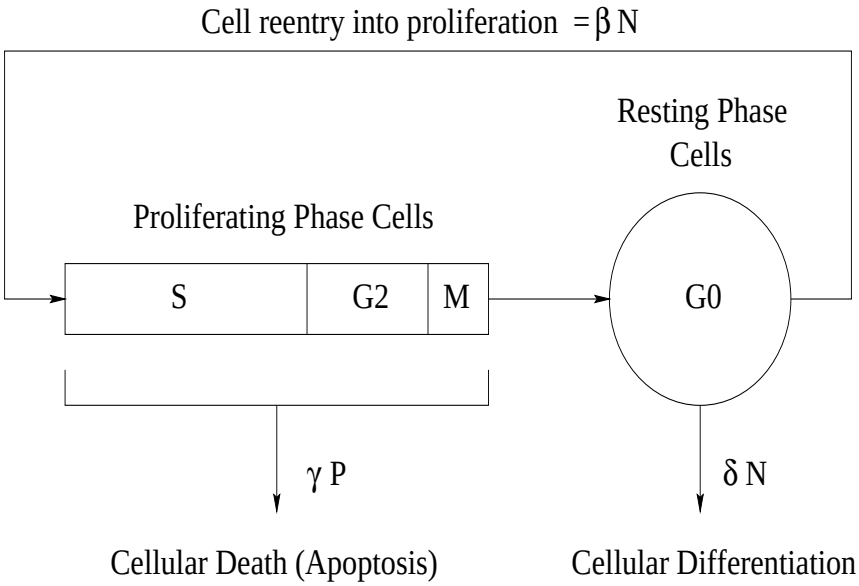


FIG. 2.1. *A schematic representation of the G_0 stem cell model. Proliferating phase cells (P) include those cells in S (DNA synthesis), G_2, and M (mitosis) while the resting phase (N) cells are in the G_0 phase. δ is the rate of differentiation into all of the committed stem cell populations, while γ represents a loss of proliferating phase cells due to apoptosis. β is the rate of cell reentry from G_0 into the proliferative phase, and τ is the duration of the proliferative phase. See Mackey* (1978), (1979), (1997) *for further details.*

The origin of the terms in these equations is fairly obvious. For example, the first term of (2.2) represents the loss of proliferating cells to cell division ($\beta(N)N$) and to differentiation (δN). The second term represents the production of proliferating stem cells, with the factor 2 accounting for the amplifying effect of cell division while $e^{-\gamma\tau}$ accounts for the attenuation due to apoptosis (programmed cell death) at rate γ. It is clear that in investigating the dynamics of the PPSC we need only understand the dynamics of the G_0 phase resting cell population since the proliferating phase dynamics are driven by the dynamics of N.

Typical values of the parameters for humans are given by Mackey (1978), (1997) as

$$\delta = 0.05\,\mathrm{d}^{-1}, \quad \beta_0 = 1.77\,\mathrm{d}^{-1}, \quad \tau = 2.2\,\mathrm{d}, \quad n = 3. \tag{2.4}$$

(The value of θ is 1.62×10^8 cells kg^{-1}, but this is immaterial for dynamic considerations.) For values of γ in the range $0.2\,\mathrm{d}^{-1}$, the consequent steady state is unstable and there is a periodic solution whose period P at the bifurcation ranges from 20–40 days. It is the observation that $P \gg \tau$, which arouses our curiosity, and which we wish to explain. (In differential delay equations, periodic oscillations have periods bounded below by 2τ and under certain circumstances the period may be in the range 2τ to 4τ.)

We rewrite (2.2) in a standard form as follows. First scale the nonproliferating phase cell numbers by θ and the time by τ so that

$$N \to \theta N, \quad \hat{t} = \tau t^*, \tag{2.5}$$

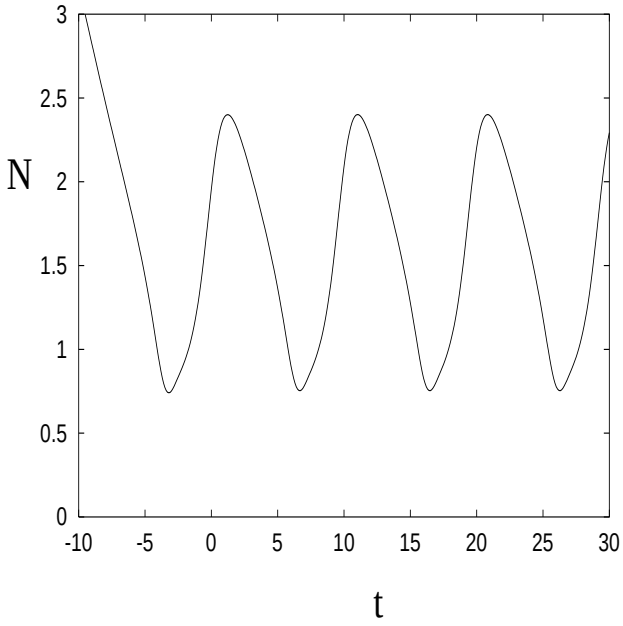


FIG. 2.2. *Solution of* (2.6) *with* $\varepsilon = 0.11$, $b = 3.9$, *and* $\mu = 1.2$.

and (2.2) becomes

$$\dot{N} = g(N_1) - g(N) + \varepsilon[\mu g(N_1) - N], \tag{2.6}$$

where $\dot{N} = dN/dt^*$, $N_1 = N(t^* - 1)$,

$$g(N) = \frac{bN}{1 + N^n}, \tag{2.7}$$

and the parameters are defined by

$$b = \beta_0 \tau, \quad \varepsilon = \delta\tau, \quad \mu = \frac{2e^{-\gamma\tau} - 1}{\delta\tau}. \tag{2.8}$$

The biological interpretation of these is as follows: b represents the rate at which cells migrate round the loop in Figure 2.1, ε represents the rate of loss through differentiation, and μ represents the net proliferation rate round the loop. The dimensionless time t^* is measured in units of the proliferative time spent in the loop. If we take $\gamma \sim 0.2\,\mathrm{d}^{-1}$, then typical values of the parameters are

$$b \sim 3.9, \quad \mu \sim 2.6, \quad \varepsilon \sim 0.11. \tag{2.9}$$

On this basis, we suppose $b, \mu = O(1)$. The long periods are associated with the relatively small value of ε, and so the aim of our analysis is to solve (2.6) when $\varepsilon \ll 1$. Figure 2.2 shows the periodic behavior when $\varepsilon = 0.11$, $b = 3.9$, and $\mu = 1.2$ (the steady state is stable when $\mu = 2.6$).

3. Singular perturbation analysis. The first order delay differential equation (2.6) is an infinite dimensional system. For example, defining the function

$$u_{t^*}(s) = N(t^* + s), \quad s \in [-1, 0], \tag{3.1}$$

we can consider (2.6) as a sequence of ordinary differential equations on the Banach space $C[-1,0]$ of continuous functions on $[-1,0]$. Singular perturbation analysis is therefore not necessarily straightforward, but we shall see that a formal procedure is indeed possible.

The key observation for our investigation is that a solution of (2.6) can be slowly varying, on a slow time scale

$$t = \varepsilon t^*, \tag{3.2}$$

or on a rather loosely defined "slow manifold" on which $N \approx N_1$. In terms of t, which represents time measured in units of the slower differentiation time scale, we have $N(t^* - 1) = N(t - \varepsilon)$; thus (2.6) is ($N' = dN/dt$)

$$N' = \frac{g(N_\varepsilon) - g(N)}{\varepsilon} + \mu g(N_\varepsilon) - N. \tag{3.3}$$

Also, by expanding N_ε for small ε, we have

$$\begin{aligned} N_\varepsilon &= N - \varepsilon N' + \tfrac{1}{2}\varepsilon^2 N'' \dots, \\ g(N_\varepsilon) &= g(N) - [\varepsilon N' - \tfrac{1}{2}\varepsilon^2 N'' + \tfrac{1}{6}\varepsilon^3 N''' \dots]g'(N) + [\tfrac{1}{2}\varepsilon^2 N'^2 \dots]g''(N) + \cdots; \end{aligned} \tag{3.4}$$

note that $N' = dN/dt$, while $g'(N) = dg/dN$. We thus have

$$[1 + g'(N)]N' = \mu g(N) - N + \varepsilon[-\mu g' N' + \tfrac{1}{2}N'' g' + \tfrac{1}{2}N'^2 g''] + \cdots, \tag{3.5}$$

and successive terms in the expansion

$$N \sim N_0 + \varepsilon N_1 + \cdots \tag{3.6}$$

satisfy the equations

$$N_0' = \frac{\mu g(N_0) - N_0}{1 + g'(N_0)}, \tag{3.7}$$

$$\begin{aligned} &[1 + g'(N_0)]N_1' + g''(N_0)N_0'N_1 \\ &\qquad = \mu g'(N_0)N_1 - N_1 + [-\mu g'(N_0)N_0' + \tfrac{1}{2}N_0'' g'(N_0) + \tfrac{1}{2}N_0'^2 g''(N_0)], \end{aligned} \tag{3.8}$$

and so on. Note particularly that in this slow region N_1 denotes the second term in the expansion for N and does *not* represent $N(t^* - 1)$; it will revert to the former meaning when we consider the dynamics in the fast "shock" layer (when the expansion will use u and v as first and second order terms). Equation (3.7) states that the rate of change of the resting stem cell population is due to net proliferation (the first term in the numerator) and loss by differentiation (the second). The effect of the delay in the proliferative cycle is to mediate the rate by the denominator. In our procedure we now begin to follow Kevorkian and Cole's (1981) exposition (pp. 67 and the following ones) quite closely.

The function $g = bN/(1 + N^n)$ is unimodal. If $g' > -1$ everywhere, then N will evolve on the slow time scale to a steady state. Suppose now that

$$b > b_c = \frac{4n}{(n-1)^2}, \tag{3.9}$$

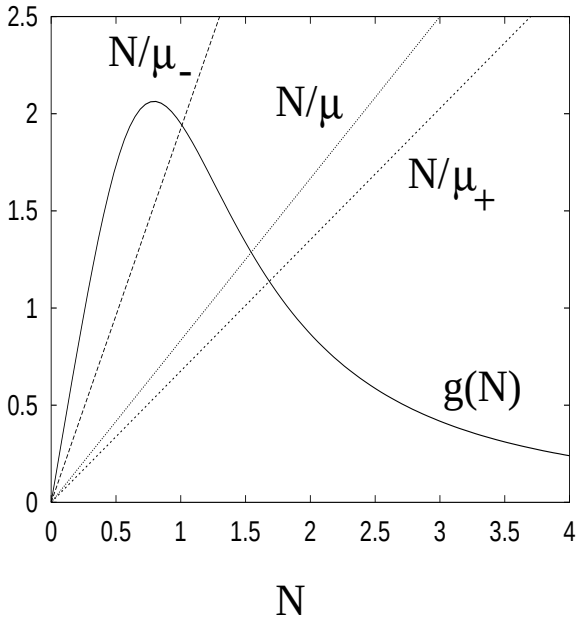


FIG. 3.1. *Graphs of $g(N)$ and N/μ for $b = 3.9$, $n = 3$, and $\mu = 1.2$ in the range $(\mu_-, \mu_+) = (0.52, 1.48)$. Also shown are graphs of N/μ_- and N/μ_+.*

which is the criterion for g' to reach -1. Then there are two values $N_- < N_+$ at which $g' = -1$; for (2.7), we have, explicitly,

$$N_\pm^n = \tfrac{1}{2}(n-1)[b \pm (b^2 - b_c b)^{1/2}] - 1. \tag{3.10}$$

If $\mu b < 1$, $N = 0$ is stable, by consideration of (3.7). If $\mu b > 1$, then there is a positive steady state N^* in which $N^* = \mu g(N^*)$. We define the two values of μ where $N^* = N_\pm$ as $\mu_\pm$; thus,

$$\mu_\pm = \frac{N_\pm}{g(N_\pm)}, \quad \mu_- < \mu_+. \tag{3.11}$$

Using (3.10), we have, explicitly,

$$\mu_\pm = \tfrac{1}{2}(n-1)\left[1 \pm \left(1 - \frac{b_c}{b}\right)^{1/2}\right]. \tag{3.12}$$

The situation which is of interest is when $\mu_- < \mu < \mu_+$, and this is depicted in Figure 3.1. In this situation, the graph of N_0' versus N_0 is as shown in Figure 3.2, and it is apparent that the fixed point in (N_-, N_+) is unstable, because the slope of the graph at the fixed point (where $N' = 0$) is positive. (Conversely, there is a stable fixed point when μ is outside this range.)

Suppose that $N > N_+$ initially. Then N_0 decreases and reaches N_+ at finite time. Define this time to be when $t = 0$; then

$$\int_{N_+}^{N_0} \left\{ \frac{1 + g'(N)}{N - \mu g(N)} \right\} dN = -t. \tag{3.13}$$

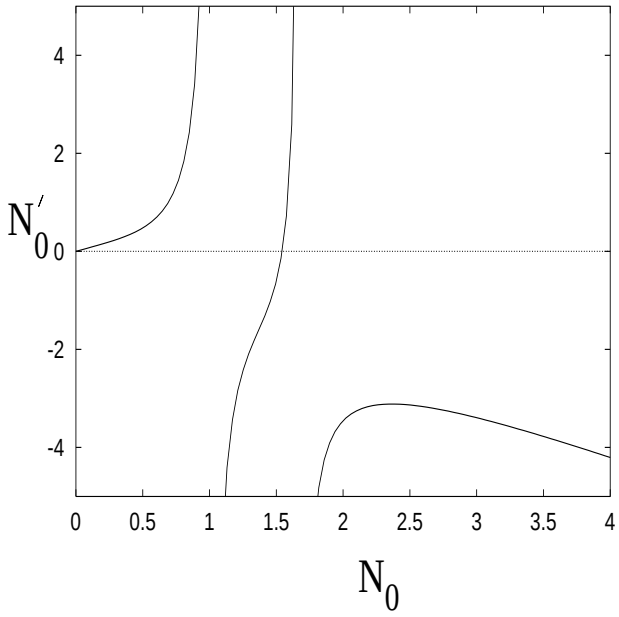


FIG. 3.2. *Graph of* $N_0'(N_0)$ *given by* (3.7) *when* $n = 3$, $b = 3.9$, *and* $\mu = 1.2$.

Since $1+g'(N_+) = 0$, the first term in the expansion of the integral in (3.13) for small $N_0 - N_+$ is quadratic, and from this we find, as $-t \to 0+$,

$$N_0 \sim N_+ + q_1(-t)^{1/2} + q_2(-t) + O[(-t)^{3/2}]. \tag{3.14}$$

Detailed expressions for the coefficients are given in the appendix.

Rearrangement of (3.8) using (3.7) allows N_1 to be obtained in the form

$$N_1 = \left(\frac{N_0 - \mu g(N_0)}{1 + g'(N_0)}\right)\left[A_1 - \frac{g'(N_0)}{2\{1 + g'(N_0)\}} + \int_{N_+}^{N_0} k(N)\, dN + h_+ \ln(N_0 - N_+)\right], \tag{3.15}$$

where $h_+ = h(N_+)$,

$$h(N) = -\frac{g'(N)\{1 - \mu g'(N)\}(N - N_+)}{2\{N - \mu g(N)\}\{1 + g'(N)\}} \tag{3.16}$$

(with the singularity at N_+ removed), and

$$k(N) = \frac{h(N) - h(N_+)}{N - N_+} - \frac{\mu g'(N)}{N - \mu g(N)}. \tag{3.17}$$

In particular,

$$h_+ = \frac{1 + \mu}{2g_+''(N_+ - \mu g_+)}, \tag{3.18}$$

where $g_+ = g(N_+)$, etc. Higher order terms can be obtained in a similar way. Note that, since $N_0 - N_+ \sim (-t)^{1/2}$ as $-t \to 0$, and $g'(N_+) = -1$, it follows that $1+g'(N_0) \sim (-t)^{1/2}$ as $-t \to 0$, and therefore (3.15) implies that $N_1 = O(1/(-t))$ as $-t \to 0+$, and the validity of the expansion breaks down when $(-t)^{1/2} \sim \varepsilon/(-t)$, i.e., when $-t \sim \varepsilon^{2/3}$.

3.1. Transition layer. The solution becomes disordered as $-t \to 0$, and specifically when $-t \sim \varepsilon^{2/3}$. In this section we analyze this "transition" layer. In addition, we might anticipate the existence of a region in which N changes on the fast (delay) time scale t^*, and this will indeed turn out to be the case. However, it transpires that such a fast region cannot be matched directly to the slow outer region, and, just as for the Van der Pol oscillator, the inability to match slow and fast regions also suggests that there is a transition region which joins the two. In terms of the outer time scale t, we shall find that the slow solution is valid for $-t \sim O(1)$, the transition region for $-t \sim O(\varepsilon^{2/3})$, and the fast "shock" layer for $-t \sim O(\varepsilon)$. Indeed, the dynamics of these three regions are essentially the same as those of the corresponding regions in the analysis of the Van der Pol equations, and we follow the exposition in Kevorkian and Cole (1981) closely. In particular, consultation of this book is strongly recommended for those less familiar with the basic procedure of matched asymptotic expansions. (Note that there are some algebraic errors in Kevorkian and Cole's exposition.)

A distinguished limit exists in which we put

$$t = \rho(\varepsilon) + \left(\frac{\varepsilon^{2/3}}{\Omega}\right)\tilde{t}, \tag{3.19}$$

where we assume $\tilde{t}$ is $O(1)$. The definition of Ω is

$$\Omega = (g_+'' q_1)^{2/3}, \tag{3.20}$$

and $\rho(\varepsilon)$ is a (small) origin shift which is introduced to allow matching to be carried out. Since $N - N_+ \sim (-t)^{1/2}$ as $-t \to 0+$, this requires $N - N_+ \sim \varepsilon^{1/3}$, and we define f via

$$N = N_+ + \left(\frac{\varepsilon^{1/3}\Omega}{g_+''}\right) f. \tag{3.21}$$

It is still appropriate to expand the delay term, and we find, from (3.3), that $f(\tilde{t})$ satisfies

$$f'' + 2ff' + 1 = \varepsilon^{1/3}[-\kappa f + \tfrac{1}{3}\Omega f''' + \Omega(f'^2 + ff'') - \lambda f^2 f'] + O(\varepsilon^{2/3}), \tag{3.22}$$

where

$$\lambda = \frac{\Omega g_+'''}{g_+''^2}, \quad \kappa = \frac{2}{\Omega^2}(1+\mu). \tag{3.23}$$

We expand f in powers of $\varepsilon^{1/3}$, thus

$$f \sim f_1 + \varepsilon^{1/3} f_2 + \cdots; \tag{3.24}$$

then from (3.22) we find that

$$\begin{aligned} f_1'' + 2f_1 f_1' + 1 &= 0, \\ f_2'' + 2(f_1 f_2)' &= -\kappa f_1 + \tfrac{1}{3}\Omega f_1''' + \Omega(f_1 f_1')' - \lambda f_1^2 f_1', \end{aligned} \tag{3.25}$$

and so on. The first of these may be integrated to yield

$$f_1' + f_1^2 + \tilde{t} = 0, \tag{3.26}$$

where the constant of integration is absorbed into the time shift $\rho(\varepsilon)$ in (3.19). The solution of the Riccati equation (3.26) is

$$f_1 = \frac{V'(\tilde{t})}{V(\tilde{t})}, \tag{3.27}$$

where V satisfies the modified Airy equation

$$V'' + \tilde{t}V = 0. \tag{3.28}$$

The solutions of (3.28) are $\mathrm{Ai}\,(-\tilde{t})$ and $\mathrm{Bi}\,(-\tilde{t})$, whose leading order behaviors as $\tilde{t} \to -\infty$ are $V \sim \exp\left[\pm\frac{2}{3}(-\tilde{t})^{3/2}\right]$ (minus for Ai). Thus if V contains any Bi, it will dominate as $\tilde{t} \to -\infty$, and hence $f_1 = V'(\tilde{t})/V(\tilde{t}) \sim -(-\tilde{t})^{1/2}$ in this limit. Therefore, in order to obtain $f_1 \sim (-\tilde{t})^{1/2}$ as $\tilde{t} \to -\infty$, which is required for matching purposes, we must suppress the Bi component and choose

$$V(\tilde{t}) = 2\sqrt{\pi}\mathrm{Ai}\,(-\tilde{t}), \tag{3.29}$$

where the premultiplicative constant is chosen for later algebraic convenience (it does not affect the definition of f_1). Since $f_1 \sim (-\tilde{t})^{1/2}$ as $\tilde{t} \to -\infty$, f_1 is monotonically decreasing for large $-\tilde{t}$, and hence from (3.26) $f_1 > (-\tilde{t})^{1/2}$. If f_1' first reaches zero for some value of $\tilde{t} = \tilde{t}_c < 0$, then at that point (3.26) implies that $f_1 = (-\tilde{t})^{1/2}$ and also that (since f_1' is continuous and $f_1 > (-\tilde{t})^{1/2}$ for $\tilde{t} < \tilde{t}_c$) $f_1' < 0$, which contradicts the assertion. Thus $f_1' < 0$ for all $\tilde{t} < 0$, and (3.26) implies this directly for $\tilde{t} > 0$. Thus we find f_1 is monotonically decreasing while it is finite, which is in the region $\tilde{t} < \tilde{t}_0$, where $\tilde{t}_0 \approx 2.338$ is the first zero of $\mathrm{Ai}\,(-\tilde{t})$. The solution will break down as $\tilde{t} \to \tilde{t}_0$, where it will match to an inner region, or shock layer, in which $t^* = O(1)$ (with a suitably chosen origin for t^*).

The first integral of $(3.25)_2$ is (using $f_1 = V'/V$)

$$f_2' + 2f_1 f_2 = -\kappa \ln V + \tfrac{1}{3}\Omega f_1'' + \Omega f_1 f_1' - \tfrac{1}{3}\lambda f_1^3 + C_2, \tag{3.30}$$

where C_2 is constant. By differentiation of (3.26) we find that $-C_2 f_1'$ is a particular solution for (3.30) when only the C_2 term is present on the right-hand side. Using $f_1 = V'/V$, we have

$$(V^2 f_2)' = C_2 V^2 - \kappa V^2 \ln V + \tfrac{1}{3}\Omega V^2 f_1'' + \Omega V^2 f_1 f_1' - \tfrac{1}{3}\lambda V^2 f_1^3. \tag{3.31}$$

Next we make use of the following identities, which can be obtained by integrating by parts and using (3.27) and (3.28):

$$\begin{aligned}
&\int V^2 f_1 f_1' = \tfrac{1}{2}V^2 f_1^2 - \int V^2 f_1^3, \\
&\int V^2 f_1'' = V^2 f_1' - V^2 f_1^2 + 2\int V^2 f_1^3, \\
&\int V^2 f_1^3 = V'^2 \ln V + (V^2 \ln V - \tfrac{1}{2}V^2)\tilde{t} - \int (V^2 \ln V - \tfrac{1}{2}V^2).
\end{aligned} \tag{3.32}$$

The comment after (3.30) implies that

$$\frac{1}{V^2}\int_{-\infty}^{\tilde{t}} V^2 d\tilde{t} = -f_1', \tag{3.33}$$

and use of (3.26) and integration by parts in (3.32) implies that

$$\frac{1}{V^2}\int V^2 f_1^3 = -f_1' \ln V - \frac{1}{V^2}\int V^2 \ln V + \tfrac{1}{2} f_1^2. \tag{3.34}$$

Hence we obtain the solution

$$\begin{aligned} f_2 = f_1'[-C_2 + \tfrac{1}{3}\Omega + \tfrac{1}{3}(\Omega+\lambda)\ln V] - \tfrac{1}{6}\lambda f_1^2 \\ + [\tfrac{1}{3}(\Omega+\lambda) - \kappa]\,\frac{1}{V^2}\int_{-\infty}^{\tilde{t}} V^2 \ln V\, d\tilde{t}, \end{aligned} \tag{3.35}$$

where we have set the integration constant D_2 (in a term D_2/V^2) to zero to prevent exponential growth as $\tilde{t} \to -\infty$.

3.2. Matching. In order to match the outer solution to the transition solution, we expand the latter for large $-\tilde{t}$ and the former for small $-t$. Equation (3.14) gives the behavior of N_0 for small $(-t)$, while if we expand (3.15) for N_0 near N_+, and use (3.14), we find

$$N_1 \sim \frac{r_1}{(-t)} + \frac{r_{21}(A_1) + r_{22}\ln(-t)}{(-t)^{1/2}} + O(1), \tag{3.36}$$

where the constants r_1, r_{21}, r_{22} are given in the appendix; r_1 and r_{22} are known, while r_{21} involves the unknown constant A_1 in (3.15).

Next we need the behavior of f_1 and f_2 as $\tilde{t} \to -\infty$. The function $V = 2\sqrt{\pi}\,\mathrm{Ai}\,(-\tilde{t})$ has the following asymptotic behavior as $\tilde{t} \to -\infty$:

$$V \sim (-\tilde{t})^{-1/4} \exp\left[-\tfrac{2}{3}(-\tilde{t})^{3/2}\right]\left[1 - \frac{5}{48(-\tilde{t})^{3/2}} + \cdots\right]. \tag{3.37}$$

Since $f_1 = V'/V$, we have

$$f_1 \sim (-\tilde{t})^{1/2} + \frac{1}{4(-\tilde{t})} + O[(-\tilde{t})^{-5/2}], \tag{3.38}$$

and thence we find from (3.35) that

$$f_2 \sim s_1(-\tilde{t}) + \frac{s_{21}(C_2) + s_{22}\ln(-\tilde{t})}{(-\tilde{t})^{1/2}} + O\left[\frac{\ln(-\tilde{t})}{(-\tilde{t})^2}\right], \tag{3.39}$$

and the coefficients s_1, s_{22}, and s_{21} are given in the appendix. Again, s_1 and s_{22} are known, and s_{21} involves the unknown constant C_2 in (3.35).

We match in an intermediate region where

$$t = \eta t_\eta + \rho(\varepsilon), \quad \tilde{t} = \left(\frac{\Omega}{\varepsilon^{2/3}}\right)\eta t_\eta, \tag{3.40}$$

and we take $\varepsilon^{2/3} \ll \eta \ll 1$ and also presume that $\eta \gg \rho$. Writing both expansions (3.6) and (3.24) in terms of t_η, the outer expansion is given by

$$\begin{aligned} N \sim N_+ + q_1(-\eta t_\eta)^{1/2} - \frac{\rho q_1}{2(-\eta t_\eta)^{1/2}} \cdots + q_2(-\eta t_\eta) + \cdots \\ + \frac{\varepsilon r_1}{(-\eta t_\eta)} \cdots + \frac{\varepsilon[r_{21} + r_{22}\ln(-\eta t_\eta)\ldots]}{(-\eta t_\eta)^{1/2}} \cdots, \end{aligned} \tag{3.41}$$

while the transition expansion is

$$N \sim N_+ + q_1(-\eta t_\eta)^{1/2} + \frac{\varepsilon}{4g''_+(-\eta t_\eta)}\dots$$
$$+ \frac{s_1\Omega^2}{g''_+}(-\eta t_\eta) + \frac{\varepsilon\Omega^{1/2}}{g''_+}\frac{[s_{21} + s_{22}\{\ln(\Omega/\varepsilon^{2/3}) + \ln(-\eta t_\eta)\}]}{(-\eta t_\eta)^{1/2}}\dots, \tag{3.42}$$

and matching requires

$$r_1 = \frac{1}{4g''_+}, \quad q_2 = \frac{s_1\Omega^2}{g''_+}, \quad r_{22} = \frac{\Omega^{1/2}}{g''_+}s_{22},$$
$$r_{21} = \frac{\Omega^{1/2}}{g''_+}[s_{21} + s_{22}\ln\Omega],$$
$$\rho = \frac{4\Omega^{1/2}s_{22}}{3g''_+q_1}\,\varepsilon\ln\varepsilon. \tag{3.43}$$

The first three of these are satisfied identically (see the appendix), while the fourth and fifth determine s_{21} and ρ, given r_{21} in the outer solution.

3.3. Matching to the shock layer. The transition solution governed by (3.22) breaks down as $\tilde{t} \to \tilde{t}_0$. Near $\tilde{t}_0$, we have that

$$V \approx -K(\tilde{t} - \tilde{t}_0) + \tfrac{1}{6}K\tilde{t}_0(\tilde{t} - \tilde{t}_0)^3 + O[(\tilde{t} - \tilde{t}_0)^4], \tag{3.44}$$

where $K = 2\sqrt{\pi}\mathrm{Ai}'(-\tilde{t}_0) \approx 2.486$, and thus

$$f_1 \sim -\frac{1}{(\tilde{t}_0 - \tilde{t})} + \tfrac{1}{3}\tilde{t}_0(\tilde{t}_0 - \tilde{t})\dots. \tag{3.45}$$

$N - N_+$ becomes of $O(1)$ when $\tilde{t}_0 - \tilde{t} \sim \varepsilon^{1/3}$ (this follows from (3.45) together with (3.21)), and this suggests that we put

$$\tilde{t} = \tilde{t}_0 + \Omega\{\varepsilon^{1/3}t^* + \sigma(\varepsilon)\}, \tag{3.46}$$

and we anticipate that $\sigma \ll 1$. In terms of t,

$$t = \rho(\varepsilon) + \left(\frac{\varepsilon^{2/3}}{\Omega}\right)\tilde{t}_0 + \varepsilon^{2/3}\sigma(\varepsilon) + \varepsilon t^*, \tag{3.47}$$

so that in the transition layer $N(t^*)$ satisfies (2.6), i.e.,

$$\frac{dN}{dt^*} = g(N_1) - g(N) + \varepsilon[\mu g(N_1) - N], \tag{3.48}$$

and N_1 reverts here to its original meaning as $N(t^* - 1)$. The behavior of f_2 as $\tilde{t} \to \tilde{t}_0$ follows from (3.35), which implies

$$f_2 \sim -\tfrac{1}{3}(\Omega + \lambda)\frac{\ln(\tilde{t}_0 - \tilde{t})}{(\tilde{t}_0 - \tilde{t})^2} + \frac{C_3}{(\tilde{t}_0 - \tilde{t})^2}, \tag{3.49}$$

where

$$C_3 = C_2 - \tfrac{1}{3}\Omega - \tfrac{1}{3}(\Omega + \lambda)\ln K - \tfrac{1}{6}\lambda + [\tfrac{1}{3}(\Omega + \lambda) - \kappa]\frac{I_0}{K^2}, \tag{3.50}$$

$$I_0 = \int_{-\infty}^{\tilde{t}_0} V^2 \ln V\, dV. \tag{3.51}$$

If we expand N in a transition region where $\varepsilon^{1/3}t^* = \eta t_\eta \ll 1$, and we suppose $\sigma \ll \eta$, then from (3.45) and (3.49) we find that

$$N \sim N_+ + \frac{\varepsilon^{1/3}}{g''_+}\left[\frac{1}{\eta t_\eta}\left\{1 - \frac{\sigma}{\eta t_\eta}\right\} - \tfrac{1}{3}\Omega^2 \tilde{t}_0(\eta t_\eta + \sigma)\ldots\right]$$
$$+ \frac{\varepsilon^{2/3}}{\Omega g''_+}\frac{1}{(\eta t_\eta)^2}\left[\{C_3 - \tfrac{1}{3}(\Omega+\lambda)\ln\Omega\} - \tfrac{1}{3}(\Omega+\lambda)\ln(-\eta t_\eta)\right]\ldots. \tag{3.52}$$

The presence of the term in $\varepsilon^{2/3}$ formally requires that we expand (3.48) as

$$N \sim u + \varepsilon^{2/3}v + O(\varepsilon) \tag{3.53}$$

and that u, v satisfy

$$\begin{aligned} u' &= g(u_1) - g(u), \\ v' &= g'(u_1)v_1 - g'(u)v, \end{aligned} \tag{3.54}$$

where the suffix 1 indicates a delayed argument.

Evidently, $u \to N_+$ as $t^* \to -\infty$, and its asymptotic behavior can be determined by writing

$$u = N_+ + \phi \tag{3.55}$$

and expanding for small ϕ, together with a Taylor expansion for $\phi_1 \equiv \phi(t^* - 1)$ as $\phi - \phi' + \cdots$. This leads (with the ansatz $\phi \gg \phi' \gg \phi'' \ldots$) to

$$0 = [-\tfrac{1}{2}\phi'' - g''_+\phi\phi'] + [\tfrac{1}{6}\phi''' + \tfrac{1}{2}g''_+(\phi'^2 + \phi\phi'') - \tfrac{1}{2}g'''_+\phi^2\phi'] + \cdots, \tag{3.56}$$

where the brackets enclose terms of similar order. Two terms of the solution of this as $t^* \to -\infty$ yield

$$\phi \sim \frac{1}{g''_+ t^*} + \frac{[E_1 - E_2\ln(-t^*)]}{t^{*2}} + \cdots, \tag{3.57}$$

where E_1 is an arbitrary constant, and E_2 is defined in the appendix. The equation for ϕ is autonomous, and an arbitrary constant can be added to t^*. It is clear that this is equivalent to changing the value of E_1; therefore the value of E_1 fixes the phase of ϕ.

The asymptotic behavior of v can then be found in a similar way, and we find that

$$0 = [-\tfrac{1}{2}v'' - g''_+(\phi v)'] + [\tfrac{1}{6}v''' + \tfrac{1}{2}g''_+(v\phi)'' - \tfrac{1}{2}g'''_+(\phi^2 v)'] + \cdots, \tag{3.58}$$

whence

$$v \sim -E_3[t^* + g''_+E_2\ln(-t^*) - E_4 + \cdots], \tag{3.59}$$

where E_3 is arbitrary and E_4 is given in the appendix. As in the Van der Pol analysis, v has a "homogeneous" solution $v = g'(u)u'$, which is $O(1/t^{*2})$ as $t^* \to -\infty$, and (3.59) comes from the "particular" solution of $(3.54)_2$, which does not tend to zero at $-\infty$.

The behavior of N as $t^* \to -\infty$ is thus

$$N \sim N_+ + \frac{1}{g''_+ t^*} + \frac{[E_1 - E_2\ln(-t^*)]}{t^{*2}} + \cdots$$
$$- \varepsilon^{2/3}E_3[t^* + g''_+E_2\ln(-t^*) - E_4], \tag{3.60}$$

and putting $\varepsilon^{1/3}t^* = \eta t_\eta$ in the matching region gives

$$N \sim N_+ + \frac{\varepsilon^{1/3}}{g''_+\eta t_\eta} + \frac{\varepsilon^{2/3}[E_1 + E_2 \ln \varepsilon^{1/3} - E_2 \ln(-\eta t_\eta)]}{(\eta t_\eta)^2} \cdots$$
$$- \varepsilon^{1/3}E_3(\eta t_\eta) - \varepsilon^{2/3}E_3[g''_+E_2 \ln(-\eta t_\eta) - g''_+E_2 \ln \varepsilon^{1/3} - E_4] \ldots . \tag{3.61}$$

Terms in (3.52) can be matched to the corresponding terms in (3.61) if

$$\begin{aligned} \sigma &= -\varepsilon^{1/3}\left\{\tfrac{1}{3}g''_+E_2 \ln \varepsilon - E_4\right\}, \\ E_1 &= \frac{C_3 - \frac{1}{3}(\Omega + \lambda)\ln \Omega}{\Omega g''_+} - \frac{E_4}{g''_+}, \\ E_3 &= \frac{\Omega^2 \tilde{t}_0}{3g''_+}, \\ E_2 &= \frac{\Omega + \lambda}{3\Omega g''_+}; \end{aligned} \tag{3.62}$$

these determine E_1, E_3, and σ, while the equation for E_2 is satisfied automatically.

3.4. Shock layer. To compute N for $t^* = O(1)$, we must solve for $N = u + \varepsilon^{2/3}v$ the equations

$$\begin{aligned} &u' = g(u_1) - g(u), \\ &u \sim N_+ + \frac{1}{g''_+t^*} + \frac{[E_1 - E_2 \ln(-t^*)]}{t^{*2}} + \cdots \quad \text{as} \quad t^* \to -\infty, \\ &v' = g'(u_1)v_1 - g'(u)v, \\ &v \sim -E_3[t^* + g''_+E_2 \ln(-t^*) - E_4] \quad \text{as} \quad t^* \to -\infty. \end{aligned} \tag{3.63}$$

The solutions of these must be obtained numerically. Note that the value of E_1 determines the origin of t^*, i.e., varying E_1 in $(3.63)_2$ simply phase shifts the solution.

It is at this point that the solution method deviates significantly from the Van der Pol procedure. The Van der Pol shock layer equation admits a first integral, and the solution can be written as a quadrature. The important point, however, is the existence of this first integral. Remarkably, an analogous procedure can be followed for the delay equations (3.63).

First, numerical integration of (3.63) indicates that u tends to a constant as $t^* \to \infty$. This is shown in Figure 3.3. The phase of the solution depends on the location of the initial interval, as shown in Figure 3.4. For the purposes of our analysis, we need to know this constant, and it can be found as follows. A trivial integration of $(3.63)_1$ shows that

$$u(t^*) + \int_{t^*-1}^{t^*} g[u(s)]ds = N_+ + g_+ \tag{3.64}$$

is constant, where the right-hand side is evaluated from the asymptotic expression for u as $t^* \to -\infty$. This immediately implies u is bounded (by $N_+ + g_+ \pm \max g$) as $t \to \infty$, and if we suppose that u tends to a constant N_L (as in Figure 3.3), then the value of the constant is easily found from (3.64) to satisfy

$$N_L + g_L = N_+ + g_+, \tag{3.65}$$

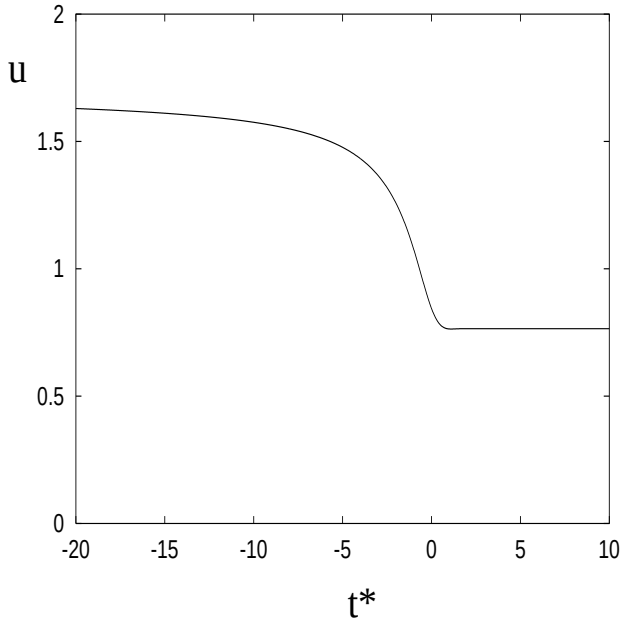


FIG. 3.3. *The solution* $u_\Delta(t^*)$ *of* $(3.63)_1$ *for* u *with the initial data taken from* $(3.63)_2$ *on the interval* $[-\Delta - 1, -\Delta)$. *The solution* $u_{20}(t^*)$ *shown is obtained using* $E_1 = 0$ *and* $\Delta = 20$. *The choice of* Δ *affects the phase of the solution, as indicated in Figure* 3.4. *This phase shift does not affect the analysis since the solution tends to a constant exponentially, so that only exponentially small terms in the slow recovery phase are affected.*

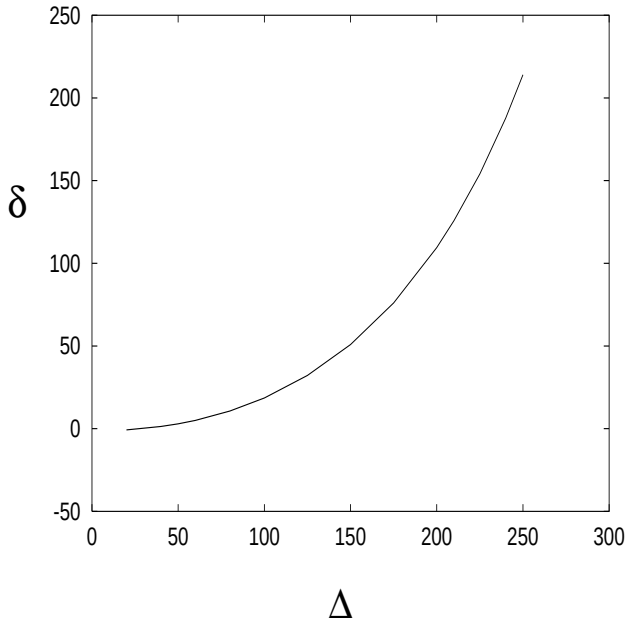


FIG. 3.4. *The variation of the computed value* δ *where* $u = 1$ *(i.e., a measure of the phase of the solution of* $(3.63)_1$*) as a function of the location* Δ *of the initial interval* $[-\Delta - 1, -\Delta)$.

where $g_L = g(N_L)$.

Next, we study the behavior of u near N_L by writing

$$u = N_L + U, \tag{3.66}$$

so that

$$U' \approx g'_L(U_1 - U), \tag{3.67}$$

where $g'_L = g'(N_L)$, and solutions are $e^{-\lambda t^*}$ for a denumerable set $\lambda_1, \lambda_2, \ldots$ of exponents. It is straightforward to show that, if these are assigned in order of increasing real part, then $\operatorname{Re}\lambda_1 > 0$, and $\operatorname{Im}\lambda_k \in ((2k-1)\pi, 2k\pi)$ if $g'_L > 0$, $\operatorname{Im}\lambda_k \in (2(k-1)\pi, (2k-1)\pi)$ if $-1 < g'_L < 0$ (we can assume $g'_L > -1$) except that $\operatorname{Im}\lambda_1 = 0$. In any event $u = N_L$ is stable, and

$$u = N_L + O(e^{-\lambda_1 t^*}) \quad \text{as} \quad t^* \to \infty. \tag{3.68}$$

Integration of $(3.63)_3$ with the matching condition $(3.63)_4$ now shows that

$$N(t) + \int_{t-1}^{t} g'[u(s)]v(s)ds = -E_3\left[\tfrac{3}{2} + g''_+ E_2\right], \tag{3.69}$$

and therefore

$$v = -v_L + O(e^{-\lambda_1 t^*}) \quad \text{as} \quad t^* \to \infty, \tag{3.70}$$

where

$$v_L = \frac{E_3[\frac{3}{2} + g''_+ E_2]}{1 + g'_L}. \tag{3.71}$$

Thus as $t^* \to \infty$,

$$N \sim N_L - \varepsilon^{2/3} v_L + O(\varepsilon, \mathrm{TST}), \tag{3.72}$$

where TST denotes the transcendentally small exponential terms.

3.5. Recovery phase. The second part of the oscillation resembles the first. There follows a slow recovery phase, terminating with transition and shock regions, and then the first slow phase is repeated. As Kevorkian and Cole (1981) point out, it is not worth the effort to compute the $O(\varepsilon \ln \varepsilon)$ terms without also computing the $O(\varepsilon)$ terms, which requires solving for further terms in the expansions. Having shown that the matching procedure does indeed work, we now abandon the $O(\varepsilon \ln \varepsilon)$ terms, and thus we do not require all the detail presented previously. Since the details of the recovery phase are similar to those of the preceding (initiation) phase, we summarize the relevant results much more briefly.

In the recovery phase, we revert to the slow time defined by (3.47):

$$t = \alpha + \varepsilon t^*, \tag{3.73}$$

where

$$\alpha = \frac{\varepsilon^{2/3}\tilde{t}_0}{\Omega} + O(\varepsilon \ln \varepsilon), \tag{3.74}$$

bearing in mind the definitions of ρ and σ. As before, N satisfies (3.5), although the $O(\varepsilon^{2/3})$ term in the shock layer requires a corresponding term in the expansion. However, it is convenient (since there is no forcing term at $O(\varepsilon^{2/3})$) to lump this correction into the $O(1)$ term, accommodating the $O(\varepsilon^{2/3})$ correction by a further phase shift in the time origin. Specifically,

$$N \sim N_0 + \varepsilon N_1 + \cdots, \tag{3.75}$$

and the solution for N_0 can be written as

$$\int_{N_0}^{N_-} \left\{ \frac{1+g'}{\mu g - N} \right\} dN = t_- - t. \tag{3.76}$$

Note that $N_0 \to N_L$ as $t \to \alpha$, and (cf. Figure 3.2) $N_L < N_-$; thus in the recovery phase $1 + g' > 0$ and $\mu g > N$. In (3.70), t_- is the time when the second transition region occurs.

We match (3.76) to the preceding shock layer by writing $N \sim N_0 \sim N_L - \varepsilon^{2/3} v_L$, $t = \alpha + \varepsilon t^*$ in (3.76), and we find that matching requires that

$$t_- = \int_{N_L}^{N_-} \left\{ \frac{1+g'(N)}{\mu g(N) - N} \right\} dN + \varepsilon^{2/3} \left[\frac{\tilde{t}_0}{\Omega} + v_L \left\{ \frac{1+g'_L}{\mu g_L - N_L} \right\} \right] + O(\varepsilon \ln \varepsilon). \tag{3.77}$$

As $t \to t_-$, (3.76) gives, analogously to (3.14),

$$N_0 \sim N_- - Q_1 (t_- - t)^{1/2} + Q_2 (t_- - t) + \cdots, \tag{3.78}$$

and in the transition region at $t = t_-$, we get

$$\begin{aligned} N &= N_- + \frac{\varepsilon^{1/3} \omega}{g''_-} f, \\ t &= t_- + r(\varepsilon) + \frac{\varepsilon^{2/3}}{\omega} \tilde{t}, \end{aligned} \tag{3.79}$$

where

$$\omega = [-g''_- Q_1]^{2/3} \tag{3.80}$$

(note $g''_- < 0$ and $Q_1 > 0$).

This leads directly to (3.22), but with k, l, ω replacing κ, λ, Ω; k and l are defined in the appendix as κ and λ, but with ω, g''_-, g'''_- replacing Ω, g''_+, g'''_+. Hence

$$f \sim \frac{-\mathrm{Ai}'(-\tilde{t})}{\mathrm{Ai}(-\tilde{t})} + O(\varepsilon^{1/3}), \tag{3.81}$$

and matching occurs automatically at leading order (and $r = O(\varepsilon \ln \varepsilon)$).

The transition layer leads to a shock layer where we write, by analogy to (3.47),

$$t = t_- + \frac{\varepsilon^{2/3} \tilde{t}_0}{\omega} + [r(\varepsilon) + \varepsilon^{2/3} s(\varepsilon)] + \varepsilon t^*, \tag{3.82}$$

and $r + \varepsilon^{2/3} s = O(\varepsilon \ln \varepsilon)$. Now, notice that to obtain the $O(\varepsilon^{2/3})$ shift in (3.77), we need to know v_L, and thus E_2 and E_3 in (3.71). Similarly, we find that, putting

$$N \sim u + \varepsilon^{2/3} v + O(\varepsilon) \tag{3.83}$$

in the recovery shock, then

$$u \sim N_- + \frac{1}{g_-'' t^*} + \frac{e_1 - e_2 \ln(-t^*)}{t^{*2}} + \cdots,$$
$$v \sim -e_3[t^* + e_2 g_-'' \ln(-t^*) - e_4 \ldots] \tag{3.84}$$

as $t^* \to -\infty$, and we will need e_2 and e_3. Since the equation for f in the recovery transition region is of the same form as in the first transition region, e_2 and e_3 are found in the same way, and thus

$$e_2 = \frac{\omega + l}{3\omega g_-''}, \quad e_3 = \frac{\omega^2 \tilde{t}_0}{3 g_-''}. \tag{3.85}$$

Finally, as $t^* \to \infty$ in the recovery shock,

$$N \sim N_U - \varepsilon^{2/3} v_U + O(\varepsilon, \mathrm{TST}), \tag{3.86}$$

where

$$N_U + g_U = N_- + g_-,$$
$$v_U = \frac{e_3[\frac{3}{2} + g_-'' e_2]}{1 + g_U'}. \tag{3.87}$$

At this point, we reenter the first slow phase, and if the motion is periodic, with period $P(\varepsilon)$, then we should regain the slow phase solution (3.13) with t replaced by t_+, where

$$t_+ = t - P(\varepsilon); \tag{3.88}$$

thus

$$\int_{N_+}^{N} \left\{ \frac{1 + g'}{N - \mu g} \right\} dN \sim -t_+ = P(\varepsilon) - t, \tag{3.89}$$

and we match this directly to the recovery shock as $t_+ \to 0$. We have $N \sim N_U - \varepsilon^{2/3} v_U$, $t = t_- + \varepsilon^{2/3} \tilde{t}_0/\omega + \varepsilon t^* + O(\varepsilon \ln \varepsilon)$, and matching of the two expressions requires, using (3.77), that

$$P(\varepsilon) = \int_{N_+}^{N_U} \left(\frac{1 + g'}{N - \mu g} \right) dN + \int_{N_L}^{N_-} \left(\frac{1 + g'}{\mu g - N} \right) dN$$
$$+ \varepsilon^{2/3} \left[\tilde{t}_0 \left(\frac{1}{\omega} + \frac{1}{\Omega} \right) + v_L \left(\frac{1 + g_L'}{\mu g_L - N_L} \right) - v_U \left(\frac{1 + g_U'}{N_U - \mu g_U} \right) \right]$$
$$+ O(\varepsilon \ln \varepsilon), \tag{3.90}$$

and this completes our analysis of the periodic solutions.

4. Discussion. The model we have sought to understand is (2.6):

$$\dot{N} = g(N_1) - g(N) + \varepsilon[\mu g(N_1) - N]. \tag{4.1}$$

If written in terms of the slow time $t = \varepsilon t^*$, this is

$$\varepsilon N' = g(N_\varepsilon) - g(N) + \varepsilon[\mu g(N_\varepsilon) - N]. \tag{4.2}$$

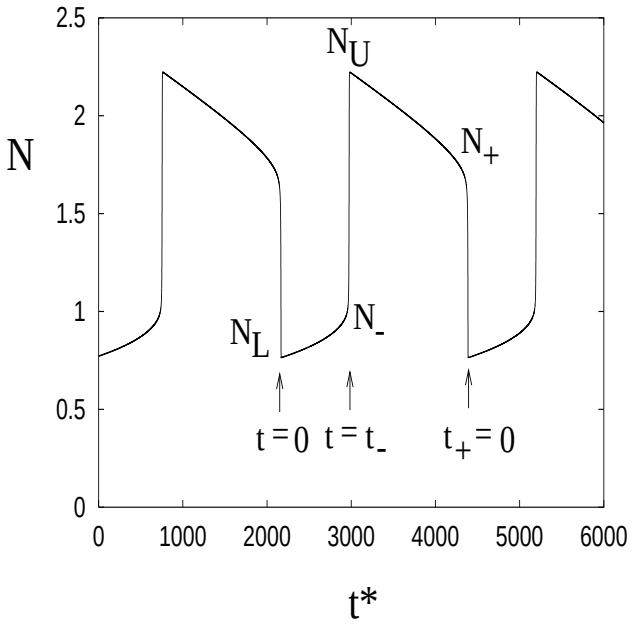


FIG. 4.1. *Numerical solution for $N(t^*)$ when $b = 3.9$, $\mu = 1.2$, $n = 3$, and $\varepsilon = 0.0001$. For a choice of origin for t at the minimum, the subsequent value t_- at the next transition is shown, and also shown is the phase resetting origin for $t_+ = t + P$.*

The analysis applies generally for unimodal functions satisfying $g'(N_\pm) = -1$, and oscillations occur for $\mu \in (\mu_-, \mu_+)$, where

$$\mu_\pm = \frac{N_\pm}{g(N_\pm)}. \tag{4.3}$$

As $\varepsilon \to 0$, we predict periodic solutions having periods (in t^*) of $P(\varepsilon)/\varepsilon$, where P is given by (3.90). The maximum and minimum values are approximately

$$N_{\max} = N_U - \varepsilon^{2/3} v_U \tag{4.4}$$

and

$$N_{\min} = N_L - \varepsilon^{2/3} v_L, \tag{4.5}$$

respectively. Figure 4.1 shows an example of the solution at very low ε, while Table 4.1 and Figures 4.2–4.5 show how these predictions compare with numerical solutions, for the particular choice of $g = bN/(1+N^n)$. It can be seen that the agreement improves, as expected, as ε becomes small.

In terms of the original dimensional quantities of the model, we see that the maximum and minimum values of N depend asymptotically entirely on the form of the function $g(N)$. The dimensional period is given to leading order by $P_0\tau/\varepsilon$, where

$$P_0 = \int_{N_+}^{N_U} \left(\frac{1+g'}{N - \mu g} \right) dN + \int_{N_L}^{N_-} \left(\frac{1+g'}{\mu g - N} \right) dN. \tag{4.6}$$

TABLE 4.1

Numerical and predicted values of $N_{\max}$, $N_{\min}$, *and period* P *given by* (4.4), (4.5), *and* (3.90). *Upper figures in each row are the values from numerical solutions; lower figures are analytical results. Parameter values used are* $n = 3$, $b = 3.9$, *and* $\mu = 1.2$. *A fourth order Runge–Kutta method is used to solve the equation, and results vary somewhat with step size, as can be seen in Figures* 4.4 *and* 4.5. *All these results are using a step size of* 0.01.

ε	Max	Min	P/ε	P
0.11	2.401	0.753	9.8	1.078
0.1	2.393	0.760	10.31	1.031
	3.546	0.578	16.14	1.614
0.05	2.342	0.778	15.03	0.7515
	3.060	0.645	21.91	1.095
0.02	2.300	0.776	26.17	0.5234
	2.681	0.698	34.58	0.692
0.005	2.260	0.765	69.11	0.34555
	2.410	0.736	80.47	0.402
0.001	2.237	0.762	260.14	0.26014
	2.293	0.753	277.12	0.277
0.0001	2.223	0.763	2220.9	0.2221
	2.245	0.759	2260.6	0.226
0.00002	2.220	0.763	10738.0	0.21476
	2.236	0.760	10841.6	0.217

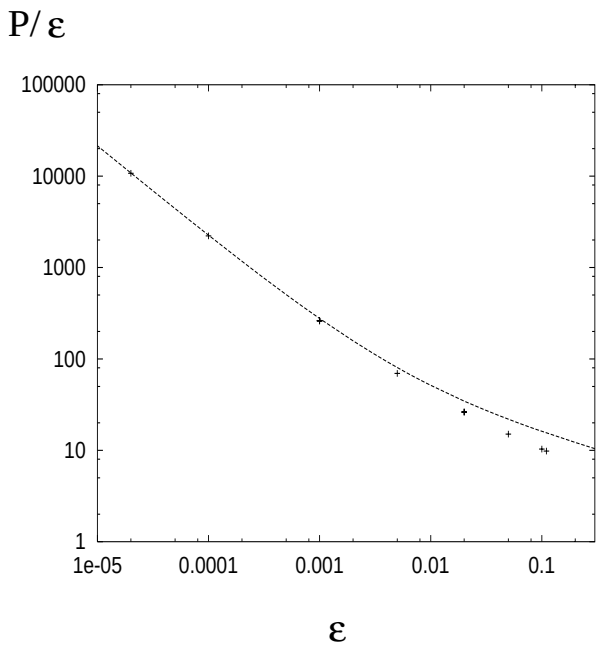


FIG. 4.2. *Variation of the actual period (in* t^**) of the numerical solution (crosses) as a function of* ε, *together with the theoretical prediction (solid curve) from* (3.90), *for* $b = 3.9$, $\mu = 1.2$, $n = 3$.

P_0 essentially depends only on the shape of $g(N)$, and thus the period is

$$P_{\text{dim}} = \frac{P_0}{\delta}, \tag{4.7}$$

that is, it is controlled by the rate of differentiation. However, oscillations do not

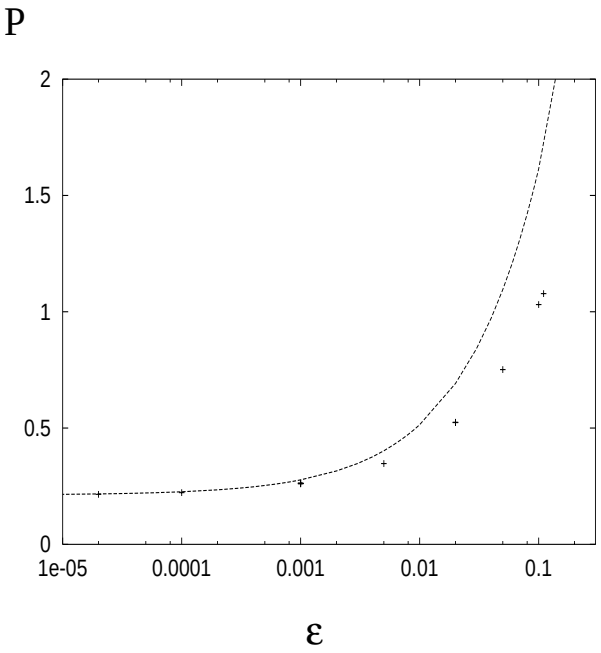


FIG. 4.3. *As for Figure* 4.2, *but plotting the period in* t, P, *versus* ε.

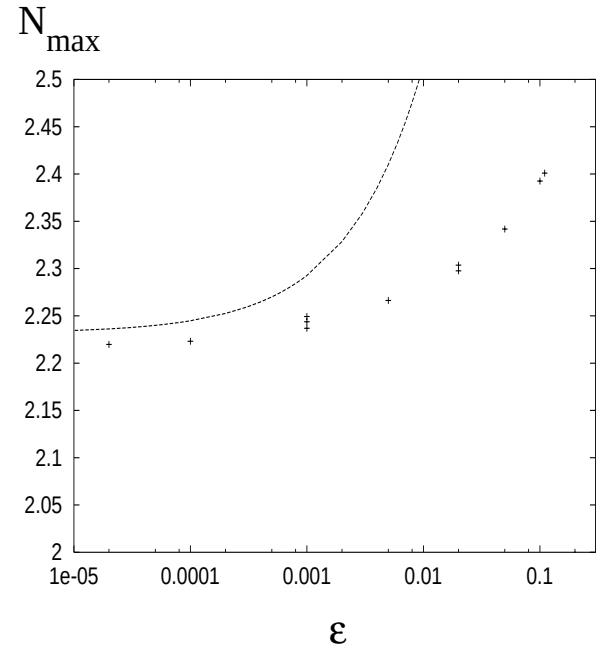


FIG. 4.4. *Numerical values of* $N_{\max}$ *(crosses) and predicted values (solid curve) from* (4.4) *as a function of* ε *for* $b = 3.9$, $\mu = 1.2$, $n = 3$. *When more than one cross is plotted, as at* $\varepsilon = 0.001$, *the different values come from the use of different step sizes in the integrator. Specifically, at* $\varepsilon = 0.001$, *decreasing step sizes* 0.01, 0.005, 0.001 *gave increasing values of* $N_{\max}$.

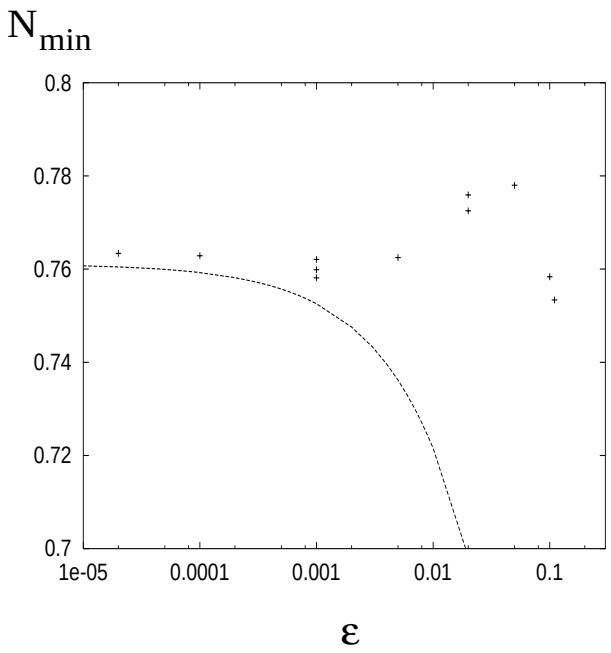


FIG. 4.5. *Computed and predicted values for* $N_{\min}$, *similar to Figure* 4.4. *Here decreasing step size at* $\varepsilon = 0.001$ *leads to decreasing* $N_{\min}$.

occur at all unless μ is a finite range of $O(1)$, and this requires that $\gamma\tau$ is increased over normal values, which can be due either to an increased proliferation delay τ or to an increased apoptotic rate γ.

It is difficult to give a useful characterization of the dimensional maximum and minimum values of N. These are simply $N_{\text{dim}}^{\max} \approx \theta N_U$ and $N_{\text{dim}}^{\min} \approx \theta N_L$. The easiest interpretation of N_U and N_L is that shown graphically in Figure 4.6. We can get a crude idea of the magnitude of the maximum and minimum values, however, if we consider the specific proliferation rate $\beta(N)$ to be adequately represented by the two quantities β_0, which is the maximum specific proliferation rate, and θ, which gives an estimate of the value of N where the proliferation rate "turns off." Our crude estimate idealizes β as being piecewise constant, with a switch off occurring at $N = \theta$, and will generally be reasonably accurate if the switch at $N \approx \theta$ is sharp. Then we have the estimates

$$\begin{aligned} N_{\text{dim}}^{\min} &\approx \frac{\theta}{1+\beta_0\tau}, \\ N_{\text{dim}}^{\max} &\approx (1+\beta_0\tau)\theta, \end{aligned} \tag{4.8}$$

and these could in principle be used to constrain the appropriate form of β in the model. The amplitude of the oscillation is, very roughly, $2\beta_0\tau\theta$.

From a mathematical perspective, the most interesting feature of the analysis is that it is completely analogous to that of a second order relaxational differential equation. In fact, Figure 4.6 indicates the similarity which can be drawn between the present model and that of the simple system

$$\begin{aligned} \varepsilon N' &= v - g(N), \\ N' + v' &= \mu g(N) - N - \varepsilon\mu v'. \end{aligned} \tag{4.9}$$

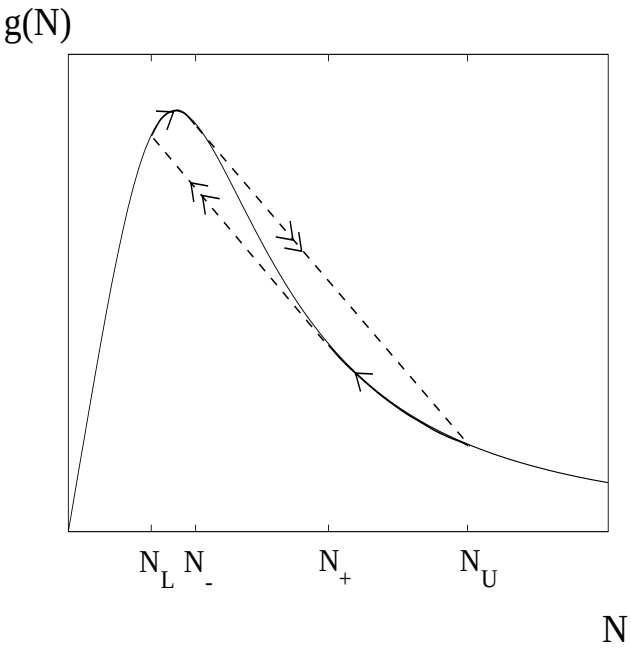


FIG. 4.6. *Phase diagram of the relaxation oscillations of both* (4.9) *and* (4.13). $g(N)$ *is plotted for* $b = 5$, $n = 3$.

In (4.9), the slow manifold is $v = g(N)$, and on this

$$N' \approx \frac{\mu g - N}{1 + g'} \tag{4.10}$$

just as for (4.2). For (4.9), there is a fast phase as $N \to N_+$ or $N \to N_-$, and in the fast phases, $N + v$ is approximately constant; since $v \to g(N)$ at either end we have the same results

$$\begin{aligned} N_+ + g(N_+) &= N_L + g(N_L), \\ N_- + g(N_-) &= N_U + g(N_U), \end{aligned} \tag{4.11}$$

as for (4.2).

The analogy can be slightly tightened by defining the functions

$$\begin{aligned} v &= g(N_\varepsilon) + \varepsilon[\mu g(N_\varepsilon) - N], \\ \hat{v} &= \frac{g(N) - g(N_\varepsilon)}{\varepsilon}. \end{aligned} \tag{4.12}$$

Evidently $\hat{v}$ is functionally dependent on v, and for slowly varying N, we have $v \approx g(N)$, $\hat{v} \approx [g(N)]'$, i.e., $\hat{v} \approx v'$; clearly this is inappropriate when N is rapidly varying. The definitions (4.12) allow us to write (4.2) in the suggestive form

$$\begin{aligned} \varepsilon N' &= v - g(N), \\ N' + \hat{v} &= \mu g(N) - N - \varepsilon\mu\hat{v}, \end{aligned} \tag{4.13}$$

and we see that the functional equation reduces precisely to the second order system (4.9) under the identification $\hat{v} = v'$. What appears to be extraordinary is that the

infinite dimensional breakdown of this approximation in the fast shock layers does not affect the analytical description in any significant way.

Apart from the mathematical novelty of solving a delay differential equation, there are some physiological ramifications of our analysis. The model for stem cell proliferation in (2.2) is a reasonable synopsis of the process, but the rate function of progress through the cycle, $\beta(N)$, is not well constrained. Nor is it possible to access this function directly, since the stem cell population itself is hidden, and oscillations are manifested in the differentiated products, which are themselves dynamically controlled by peripheral controlling mechanisms. Therefore it is useful to be able to characterize the oscillations of the resting stem cell population for a variety of different progression functions $\beta(N)$, and our analysis allows us to do this. It will also allow us in future work to analyze how oscillations in the stem cell population propagate through the maturing cell types, so that in principle we can use resulting observed cell cycles as a constraint on the stem cell dynamics.

Appendix. In (3.14), we find q_1 and q_2:

$$q_1 = \left[\frac{2(N_+ - \mu g_+)}{g''_+}\right]^{1/2}, \tag{A.1}$$

$$q_2 = \tfrac{1}{3} q_1^2 \left[\left(\frac{1+\mu}{N_+ - \mu g_+}\right) - \frac{g'''_+}{2g''_+}\right]; \tag{A.2}$$

h_+ is defined in (3.18):

$$h_+ = \frac{1+\mu}{2g''_+(N_+ - \mu g_+)}; \tag{A.3}$$

Ω is defined in (3.20):

$$\Omega = q_1^{2/3} g_+''^{2/3}; \tag{A.4}$$

κ and λ are defined in (3.23):

$$\kappa = \frac{2}{\Omega^2}(1+\mu), \tag{A.5}$$

$$\lambda = \frac{\Omega g'''_+}{g''^2_+}; \tag{A.6}$$

r_1, r_{21}, and r_{22} are defined in (3.41):

$$r_1 = \frac{1}{4g''_+}, \tag{A.7}$$

$$r_{21} = \frac{q_1}{2}\left[A_1 + \frac{q_2}{2g''_+ q_1^2} - \tfrac{1}{2} - \frac{g'''_+}{4g''^2_+} + h_+ \ln q_1\right], \tag{A.8}$$

$$r_{22} = \tfrac{1}{4} q_1 h_+; \tag{A.9}$$

s_1, s_{21}, and s_{22} are defined in (3.42):

$$s_1 = \tfrac{1}{3}\kappa - \tfrac{1}{6}\lambda, \tag{A.10}$$

$$s_{21} = \tfrac{1}{2} C_2 - \tfrac{1}{4}\Omega - \tfrac{1}{6}\lambda + \tfrac{1}{12}\kappa, \tag{A.11}$$

$$s_{22} = \tfrac{1}{8}\kappa. \tag{A.12}$$

E_2 and E_4 appear in (3.57) and (3.59):

$$E_2 = \frac{1}{3g''_+}\left(1 + \frac{g'''_+}{g''^2_+}\right), \tag{A.13}$$

$$E_4 = \tfrac{1}{2}g''_+E_2 + \frac{g'''_+}{g''^2_+} + g''_+E_1. \tag{A.14}$$

In (3.78) we find Q_1 and Q_2:

$$Q_1 = \left[\frac{2(N_- - \mu g_-)}{g''_-}\right]^{1/2}, \tag{A.15}$$

$$Q_2 = \tfrac{1}{3}Q_1^2\left[\left(\frac{1+\mu}{N_- - \mu g_-}\right) - \frac{g'''_-}{2g''_-}\right], \tag{A.16}$$

and then ω, k, and l are introduced in (3.80) and are restated below:

$$\omega = (-Q_1 g''_-)^{2/3}, \tag{A.17}$$

$$k = \frac{2}{\omega^2}(1+\mu), \tag{A.18}$$

$$l = \frac{\omega g'''_-}{g''^2_-}. \tag{A.19}$$

REFERENCES

Z. Artstein and M. Slemrod (2001), *On singularly perturbed retarded functional differential equations*, J. Differential Equations, 171, pp. 88–109.

J. Bélair, M. C. Mackey, and J. M. Mahaffy (1995), *Age-structured and two-delay models for erythropoiesis*, Math. Biosci., 128, pp. 317–346.

L. L. Bonilla and A. Liñan (1984), *Relaxation oscillations, pulses, and travelling waves in the diffusive Volterra delay-differential equation*, SIAM J. Appl. Math., 44, pp. 369–391.

S.-N. Chow and W. Z. Huang (1994), *Singular perturbation problems for a system of differential-difference equations*, J. Differential Equations, 112, pp. 257–307.

S.-N. Chow, X.-B. Lin, and J. Mallet-Paret (1989), *Transition layers for singularly perturbed delay differential equations with monotone nonlinearities*, J. Dynam. Differential Equations, 1, pp. 3–43.

S.-N. Chow and J. Mallet-Paret (1982), *Singularly perturbed delay-differential equations*, in Coupled Nonlinear Oscillators, North-Holland Math. Stud. 80, J. Chandra and A. C. Scott, eds., North-Holland, Amsterdam, pp. 7–12.

P. Fortin and M. C. Mackey (1999), *Periodic chronic myelogenous leukemia: Spectral analysis of blood cell counts and etiological implications*, Brit. J. Haematol., 104, pp. 336–345.

A. C. Fowler (1982), *An asymptotic analysis of the logistic delay equation when the delay is large*, IMA J. Appl. Math., 28, pp. 41–49.

W. S. C. Gurney, S. P. Blythe, and R. M. Nisbet (1980), *Nicholson's blowflies revisited*, Nature, 287, pp. 17–21.

J. K. Hale and W. Z. Huang (1996), *Periodic solutions of singularly perturbed delay equations*, Z. Agnew. Math. Phys., 47, pp. 57–88.

C. Haurie, D. C. Dale, and M. C. Mackey (1998), *Cyclical neutropenia and other periodic hematological disorders: A review of mechanisms and mathematical models*, Blood, 92, pp. 2629–2640.

C. Haurie, D. C. Dale, and M. C. Mackey (1999), *Occurrence of periodic oscillations in the differential blood counts of congenital, idiopathic and cyclical neutropenic patients before and during treatment with G-CSF*, Exper. Hematol., 27, pp. 401–409.

C. Haurie, R. Person, D. C. Dale, and M. C. Mackey (1999), *Hematopoietic dynamics in grey collies*, Exper. Hematol., 27, pp. 1139–1148.

C. Haurie, D. Dale, R. Rudnicki, and M. C. Mackey (2000), *Mathematical modeling of complex neutrophil dynamics in the grey collie*, J. Theor. Biol., 204, pp. 505–519.

J. Kevorkian and J. D. Cole (1981), *Perturbation Methods in Applied Mathematics*, Springer-Verlag, Berlin.

C. G. Lange and R. M. Miura (1982), *Singular perturbation analysis of boundary-value problems for differential-difference equations*, SIAM J. Appl. Math., 42, pp. 502–531.

C. G. Lange and R. M. Miura (1994), *Singular perturbation analysis of boundary-value problems for differential-difference equations. V. Small shifts with layer behavior*, SIAM J. Appl. Math., 54, pp. 249–272.

M. C. Mackey (1978), *A unified hypothesis for the origin of aplastic anemia and periodic haematopoiesis*, Blood, 51, pp. 941–956.

M. C. Mackey (1979), *Dynamic haematological disorders of stem cell origin*, in Biophysical and Biochemical Information Transfer in Recognition, J. G. Vassileva-Popova and E. V. Jensen, eds., Plenum Publishing, New York, pp. 373–409.

M. C. Mackey (1997), *Mathematical models of hematopoietic cell replication and control*, in The Art of Mathematical Modelling: Case Studies in Ecology, Physiology and Biofluids, H. G. Othmer, F. R. Adler, M. A. Lewis, and J. C. Dallon, eds., Prentice-Hall, Englewood Cliffs, NJ, pp. 149–178.

M. C. Mackey (2001), *Cell kinetic status of hematopoietic stem cells*, Cell Proliferation, 34, pp. 71–83.

J. M. Mahaffy, J. Bélair, and M. C. Mackey (1998), *Hematopoietic model with moving boundary condition and state dependent delay*, J. Theor. Biol., 190, pp. 135–146.

D. Pieroux, T. Erneux, A. Gavrielides, and V. Kovanis (2000), *Hopf bifurcation subject to a large delay in a laser system*, SIAM J. Appl. Math., 61, pp. 966–982.

M. Santillán, J. M. Mahaffy, J. Bélair, and M. C. Mackey (2000), *Regulation of platelet production: The normal response to perturbation and cyclical platelet disease*, J. Theor. Biol., 206, pp. 585–603.

J. Swinburne and M. C. Mackey (2000), *Cyclical thrombocytopenia: Characterization by spectral analysis and a review*, J. Theor. Med., 2, pp. 81–91.

SIAM J. APPL. MATH.
Vol. 63, No. 1, pp. 324–350

A FINITE ELEMENT METHOD FOR AN EIKONAL EQUATION MODEL OF MYOCARDIAL EXCITATION WAVEFRONT PROPAGATION*

KARL A. TOMLINSON†, PETER J. HUNTER†, AND ANDREW J. PULLAN†

Abstract. An efficient finite element method is developed to model the spreading of excitation in ventricular myocardium by treating the thin region of rapidly depolarizing tissue as a propagating wavefront. The model is used to investigate excitation propagation in the full canine ventricular myocardium. An eikonal-curvature equation and an eikonal-diffusion equation for excitation time are compared. A Petrov–Galerkin finite element method with cubic Hermite elements is developed to solve the eikonal-diffusion equation on a reasonably coarse mesh. The oscillatory errors seen when using the Galerkin weighted residual method with high mesh Péclet numbers are avoided by supplementing the Galerkin weights with C^0 functions based on derivatives of the interpolation functions. The ratio of the Galerkin and supplementary weights is a function of the Péclet number such that, for one-dimensional propagation, the error in the solution is within a small constant factor of the optimal error achievable in the trial space. An additional no-inflow boundary term is developed to prevent spurious excitation from initiating on the boundary. The need for discretization in time is avoided by using a continuation method to gradually introduce the nonlinear term of the governing equation. A simulation is performed in an anisotropic model of the complete canine ventricular myocardium, with 2355 degrees of freedom for the dependent variable.

Key words. eikonal equation, myocardial excitation, wavefront propagation, Petrov–Galerkin method, Hermite interpolation, numerical continuation

AMS subject classifications. 65N30, 35J60, 35K90

PII. S0036139901389513

1. Introduction.

In developing a computational model of the electrical behavior of the ventricular myocardium, it would be unreasonable to expect to be able to model every microscopic biological process that occurs within and between each and every cell. Such detail in the model is also unnecessary: the ventricular function and the electrical fields induced in the torso are not so much affected by the activity of one ion or one ion channel or even one cell as by the collective activity of many cells. Instead of resolving the small spatial detail of the microscopic processes, the collective macroscopic effect of these processes can be modelled.

The most intense electrical activity is the depolarization of cells, which leads to the activation of the mechanisms that cause the myocardium to contract and the heart to pump. Depolarization occurs quickly and in only a narrow region of cells at a time, so this narrow region can be considered as a propagating *excitation wavefront*. An eikonal model may be used to approximate the propagation process, describing the motion of the wavefront by the time at which it excites every point in the myocardium.

A finite element method with cubic Hermite elements is developed to determine excitation times on a fairly coarse mesh for large scale simulations. Petrov–Galerkin weighted residual equations, developed in section 2, are supplemented with a no-inflow term, developed in section 3, to prevent spurious excitation on boundaries, and are solved by a continuation method with Newton's method (section 4). Section 5

*Received by the editors May 16, 2001; accepted for publication (in revised form) April 9, 2002; published electronically September 12, 2002.
http://www.siam.org/journals/siap/63-1/38951.html
†Bioengineering Institute, The University of Auckland, Private Bag 92019, Auckland, New Zealand (k.tomlinson@auckland.ac.nz, p.hunter@auckland.ac.nz, a.pullan@auckland.ac.nz).

employs the computational method to simulate excitation in the complete ventricular myocardium.

1.1. The bidomain model. As a means for collecting together the microscopic functional elements of the myocardium to model their macroscopic effects, Schmitt [21] suggested the concept of two interpenetrating domains. One domain was to represent the volume-averaged properties of the intracellular contents and their interconnections, and the other domain was to represent the volume-averaged properties of the surrounding extracellular tissue and fluid. These domains were to coexist spatially, and the behavior of current flow between them was to be based on the volume-averaged properties of the cell membrane. This approach is now generally referred to as the *bidomain model* [12]. The two domains are referred to as the *intracellular* and *extracellular* domains. Each is treated as a continuum.

A reaction-diffusion system of equations for the potential ϕ_e in the extracellular domain and the difference in potential V_m across the membrane between the domains can be derived from conservation of current under the assumptions that capacitive, inductive, and electromagnetic propagative effects within the domains are negligible and that the current in each domain obeys Ohm's law:

$$\nabla \cdot \left(G^i \nabla V_m\right) = -\nabla \cdot \left((G^i + G^e)\nabla \phi_e\right),$$

$$i_{ion} + c_m \frac{\partial V_m}{\partial t} = -\nabla \cdot (G^e \nabla \phi_e). \tag{1.1}$$

The intracellular potential ϕ_i is the sum of the extracellular potential ϕ_e and the transmembrane potential V_m. G^i and G^e are intra- and extracellular *effective conductivity tensors*. The fibrous and laminar structure of the myocardium is modelled under the assumptions that the conductivities are orthotropic and that they share the same principal axes, $\boldsymbol{a}_l$, $\boldsymbol{a}_t$, and $\boldsymbol{a}_n$, where $\boldsymbol{a}_l$ is parallel to the fibers (longitudinal), $\boldsymbol{a}_t$ is transverse to the fibers but in the plane of the sheets, and $\boldsymbol{a}_n$ is normal to the sheets. i_{ion} represents the sum of the (outward) membrane ionic currents per unit tissue volume, and c_m is the membrane capacitance per unit volume.

It is assumed that the extracellular space is in direct contact with the outside volume. Continuity of the extracellular potential ϕ_e with the potential ϕ_o in the outside volume and conservation of current between the volumes leads to the boundary conditions

$$\phi_e = \phi_o, \tag{1.2}$$

$$\boldsymbol{n} \cdot G^i \nabla(\phi_e + V_m) = 0,$$

$$\boldsymbol{n} \cdot G^e \nabla \phi_e = \boldsymbol{n} \cdot \boldsymbol{j}_o \qquad \text{on } \partial\Omega, \tag{1.3}$$

where $\boldsymbol{n}$ is the unit normal to the boundary and $\boldsymbol{j}_o$ is the current density in the outside volume [15].

If the intra- and extracellular conductivity tensors were related by a constant scalar factor (equal anisotropy), then system (1.1) could be reduced to a simple *monodomain* reaction-diffusion equation in one variable:

$$i_{ion} + c_m \frac{\partial V_m}{\partial t} = \nabla \cdot (G^m \nabla V_m). \tag{1.4}$$

G^m has the same principal axes as G^i and G^e, and the reciprocals of its eigenvalues equal the sums of the reciprocals of the intra- and extracellular principal conductivi-

ties. The boundary condition on V_{m} would be

$$\boldsymbol{n} \cdot \mathrm{G}^{\mathrm{e}} \nabla V_{\mathrm{m}} = -\boldsymbol{n} \cdot \mathrm{G}^{\mathrm{o}} \nabla \phi_{\mathrm{o}} \qquad \text{on } \partial\Omega. \tag{1.5}$$

If the anisotropic ratios are not equal, the monodomain equation (1.4) may still be used as an approximation of the bidomain system (1.1). For plane wave propagation in any of the three principal directions, both (1.1) and (1.4) predict the same propagation speeds, but the predicted speeds may differ for intermediate directions.

It is convenient to scale (1.4) so that the parameters give indications of the important spatial and temporal scales. This can be done by dividing the equation by a characteristic conductance per unit volume. During the depolarization phase of the action potential, consideration of the large difference between the activation and inactivation time constants of the dominating fast sodium current leads to the approximation of i_{ion} as a time-independent function of the transmembrane voltage [5]. That is, $i_{\mathrm{ion}} = i_{\mathrm{ion}}(V_{\mathrm{m}})$. If the transmembrane potential V_{m} is near its resting potential V_{r}, the behavior of the ionic membrane currents can be approximated by assuming the membrane has a passive conductance per unit volume defined by

$$\frac{1}{\mathrm{r_m}} := \frac{\mathrm{d}i_{\mathrm{ion}}}{\mathrm{d}V_{\mathrm{m}}}(V_{\mathrm{r}}). \tag{1.6}$$

(The symbol ":=" denotes definition.) Multiplying the terms in (1.4) by an average (space-independent) value $\bar{\mathrm{r}}_{\mathrm{m}}$ of $\mathrm{r_m}$ gives

$$\bar{\mathrm{r}}_{\mathrm{m}} i_{\mathrm{ion}} + \tau_{\mathrm{m}} \frac{\partial V_{\mathrm{m}}}{\partial t} = \nabla \cdot (\mathrm{M} \nabla V_{\mathrm{m}}), \tag{1.7}$$

where

$$\mathrm{M} := \bar{\mathrm{r}}_{\mathrm{m}} \mathrm{G}^{\mathrm{m}} \qquad \text{and} \qquad \tau_{\mathrm{m}} := \bar{\mathrm{r}}_{\mathrm{m}} \mathrm{c_m} \tag{1.8}$$

are the *coupling tensor*, which has dimensions of space squared, and the *membrane time constant*, which has dimension of time. The eigenvalues of M are squares of the space constants λ_{l}, λ_{t}, and λ_{n} in each of the principal directions. They may be expressed in terms of conductivities using

$$\frac{1}{\lambda_{\mathrm{l}}^2} = \frac{1}{\bar{\mathrm{r}}_{\mathrm{m}}} \left(\frac{1}{\mathrm{g_{il}}} + \frac{1}{\mathrm{g_{el}}} \right), \qquad \text{etc.} \tag{1.9}$$

These space and time constants are appropriate when the behavior of the tissue is largely passive such as in the early stages of the action potential. The behavior in these stages is important for propagation as it initiates the change in transmembrane potential that leads to activation of the active currents. The time and space constants relevant in the fastest stage of depolarization, however, may be different. The magnitude of the maximum slope of $i_{\mathrm{ion}}(V_{\mathrm{m}})$ is much larger than the slope at $V_{\mathrm{m}} = V_{\mathrm{r}}$ used to define $\mathrm{r_m}$ in (1.6). The appropriate multiplier for scaling the system of equations is then smaller than $\bar{\mathrm{r}}_{\mathrm{m}}$, and so the appropriate space and time constants are also smaller. The space constants λ_{l}, λ_{t}, and λ_{n} probably provide an indication of the region of influence that the excitation wavefront has, and, together with τ_{m}, they provide an upper bound on the relevant spatial and temporal scales.

Solution of the reaction-diffusion equation (1.7) is very computationally demanding due to the important spatial scales being much smaller than the dimensions of

the ventricles. As discussed in [23], the space constants for the passive behavior of canine myocardium are probably $\lambda_l \approx 0.8\,\text{mm}$ and $\lambda_n < \lambda_t \approx 0.5\,\text{mm}$. Reasonable approximation of the potential would probably require at least 5 degrees of freedom to represent changes over the distance of a space constant. This implies that at least 5^3 degrees of freedom would be required to represent a volume of about $0.8 \times 0.5 \times 0.5 = 0.2\ \text{mm}^3$. For the full canine ventricular myocardium with a volume of about $0.2 \times 10^6\ \text{mm}^3$, at least 10^8 degrees of freedom would be needed.

1.2. An eikonal approach. Given the difficulty in the numerical solution of a reaction-diffusion equation for transmembrane potential, a governing equation is sought for the motion of the excitation wavefront. It is expected that the speed of propagation can be assumed to vary more slowly and over much larger spatial scales than the transmembrane potential. This assumption is probably reasonable most of the time, but there are abrupt spatial changes in propagation speed where a wavefront collides with the boundary or another wavefront. The fine details of the wavefront shape in these small collision regions are not, however, expected to have much influence on the overall ventricular function.

The wavefront motion can be described by the *excitation time* $u(\boldsymbol{x})$, defined as the time at which the wavefront passes through the point $\boldsymbol{x}$ (or, more specifically, the time at which the transmembrane potential at that point crosses the value midway between its resting and plateau potentials). The position of the wavefront at any time t is then given by the surface along which $u(\boldsymbol{x}) = t$, and the excitation time can be described numerically on a stationary mesh. A governing equation for u is referred to as an *eikonal equation*.

Many myocardial excitation models have been based on Huygens' principle (reviewed in [19]) and are effectively approximating an eikonal equation. In such models, the heart is represented by a matrix of cells or grid points. At fixed time intervals after any cell is excited, its quiescent neighboring cells are excited. The time interval before excitation of each neighboring cell depends on the distance to the cell and the propagation speed for that direction. This method requires little computational effort but has the disadvantage that the numerical treatment of the eikonal equation is very low order and propagation can occur in only a finite number of directions. The result is that the wavefronts generated are polyhedral instead of ellipsoidal.

More accurate numerical solutions for excitation wavefront propagation have been obtained using wavefront propagation equations derived from the reaction-diffusion equation (1.7) under the assumption that the profile of the depolarization upstroke varies slowly in space.

An alternative approach for describing wavefront propagation is to use a function $\varphi(\boldsymbol{x}, t)$, defined so that, at any time t, the level set of points $\boldsymbol{x}$ such that $\varphi(\boldsymbol{x}, t) = 0$ gives the position of the wavefront at that time (see [22]). Keener [13] derived an equation for φ from (1.7) by selecting a moving coordinate system such that V_m is a function of only a spatial variable normal to the wavefront and then requiring the current conservation equation to be satisfied at the wavefront. The resulting equation,

$$\left[c_0 + \nabla \cdot \left(\frac{M\nabla u}{\sqrt{\nabla u \cdot M\nabla u}}\right)\right] \sqrt{\nabla\varphi \cdot M\nabla\varphi} = \tau_m \frac{\partial\varphi}{\partial t}, \tag{1.10}$$

is parabolic and time-dependent. φ has a physical interpretation only at its zero contour, so the selection of initial conditions is unclear. If, however, $\varphi(\boldsymbol{x}, t)$ is chosen

to be $t - u(\boldsymbol{x})$, then (1.10) reduces to a parabolic eikonal equation for excitation time:

$$c_0\sqrt{\nabla u \cdot \mathrm{M}\nabla u} - \sqrt{\nabla u \cdot \mathrm{M}\nabla u}\,\nabla \cdot \left(\frac{\mathrm{M}\nabla u}{\sqrt{\nabla u \cdot \mathrm{M}\nabla u}}\right) = \tau_{\mathrm{m}}. \tag{1.11}$$

The numerical solution to (1.10) was found using finite difference discretizations in space and time. Second-order central differences were initially used for the spatial discretization [13], but [14] later replaced these with first-order upwind differences to stabilize the numerical solution.

An elliptic eikonal equation was derived by Colli Franzone, Guerri, and Rovida [8] and Colli Franzone, Guerri, and Tentoni [9] using singular perturbation techniques. The equivalent eikonal equation for reaction-diffusion equation (1.7) is

$$c_0\sqrt{\nabla u \cdot \mathrm{M}\nabla u} - \nabla \cdot (\mathrm{M}\nabla u) = \tau_{\mathrm{m}}. \tag{1.12}$$

As the equation is elliptic, a boundary condition is required around the entire boundary. Without a model of the surrounding tissue, it is not possible to predict the current flux from the outside domain, and thus the boundary condition (1.5) for the reaction-diffusion system is not helpful. However, experimental evidence suggests that epicardial isochrones are unaffected by surrounding conducting volumes [11]. Without a surrounding volume, boundary condition (1.5) leads to the simple no-flux boundary condition

$$\boldsymbol{n} \cdot \mathrm{M}\nabla u = 0, \tag{1.13}$$

where $\boldsymbol{n}$ is the unit normal to the boundary.

In their numerical solution of the eikonal equation, Colli Franzone and Guerri [6] added a time derivative term to give a related parabolic equation in space and time. The time-dependent equation for (1.12) is

$$\frac{\partial \breve{u}}{\partial t} + c_0\sqrt{\nabla \breve{u} \cdot \mathrm{M}\nabla \breve{u}} - \nabla \cdot (\mathrm{M}\nabla \breve{u}) = \tau_{\mathrm{m}}. \tag{1.14}$$

The steady-state solution for $\breve{u}(\boldsymbol{x}, t)$ is the excitation time $u(\boldsymbol{x})$. To find this solution, spatial discretization was performed using finite element–like integrals of quantities calculated by finite differences, and a finite difference scheme was used to step through time until $\breve{u}$ approached its limiting value. The spatial discretization was later modified [7] so that traditional finite element integrals were used for most terms, but a first-order upwind finite difference was used for the first-order spatial derivatives. A purely explicit finite difference scheme in time gave a method that was similar to Jacobi successive overrelaxation. In order to avoid instability, the time step (or relaxation parameter) had to be small, and thus convergence was very slow.

When compared to the reaction-diffusion equation (1.7), the eikonal equations (1.11) and (1.12) have the advantages that the domain is reduced by one dimension (because the dependent variable is no longer a function of time) and that the important spatial scales are much larger. In order to make use of these advantages, a numerical method needs to be found that requires only a spatial discretization and will work effectively when this discretization is reasonably coarse.

Both the level set and relaxation methods discussed above fail to take advantage of the fact that excitation time depends only on spatial position. The use of either of the time-dependent equations (1.10) or (1.14) increases the size of the domain by one dimension. For this reason, the method investigated here uses numerical

continuation, with Newton's method applied directly to a spatial discretization of an eikonal equation, to converge from an initial guess to the solution. Each Newton iteration requires not much more work than that required in an iteration of an implicit time stepping scheme for either (1.10) or (1.14), yet makes a considerably better attempt to go directly to the required solution.

1.3. Interpretation and comparison of eikonal equations. Interpretations of the two suggested eikonal equations (1.11) and (1.12) for wavefront propagation can be made from each of the terms involved.

The contours of u give the positions of the wavefront at time $t = u$. The gradient of u at any point along one of these contours is therefore normal to that wavefront surface and has magnitude equal to the reciprocal of the speed of that point on the wavefront. That is,

$$\nabla u = \frac{1}{\theta}\,\boldsymbol{p}, \tag{1.15}$$

where θ is the local wavefront speed and $\boldsymbol{p}$ is the unit normal to the wavefront pointing away from depolarized tissue. A space constant ρ in the direction of propagation $\boldsymbol{p}$ may be calculated from the square root of the component of the coupling tensor in that direction:

$$\rho := \sqrt{\boldsymbol{p} \cdot \mathrm{M}\boldsymbol{p}}. \tag{1.16}$$

The first term in both governing equations (1.11) and (1.12) is a nonlinear advection term, which may be written as

$$\mathrm{c}_0 \sqrt{\nabla u \cdot \mathrm{M}\nabla u} = \mathrm{c}_0 \frac{\rho}{\theta}. \tag{1.17}$$

This term is an anisotropic generalization of the left-hand side of the standard eikonal equation $|\nabla u| = 1$ and is a function of the local speed of the wavefront surface.

The second term in the parabolic equation (1.11) may be written as

$$\sqrt{\nabla u \cdot \mathrm{M}\nabla u}\,\nabla \cdot \left(\frac{\mathrm{M}\nabla u}{\sqrt{\nabla u \cdot \mathrm{M}\nabla u}}\right) = \frac{\rho}{\theta}\,\nabla \cdot \left(\frac{1}{\rho}\mathrm{M}\boldsymbol{p}\right) = \frac{\rho}{\theta}\,\kappa, \tag{1.18}$$

where κ is an anisotropic generalization of the mean curvature. It is positive when the wavefront is convex if viewed from ahead of the wavefront. The parabolic equation is therefore called an *eikonal-curvature equation*.

Using expressions (1.17) and (1.18) in the eikonal-curvature equation (1.11) gives

$$\frac{\tau_{\mathrm{m}}}{\rho}\theta = \mathrm{c}_0 - \kappa. \tag{1.19}$$

For a given propagation direction, this equation states that the speed of the wavefront is a linear function of its anisotropic mean curvature. Propagation is faster when the wavefront is concave, and slower when it is convex. This reflects the dependency of tissue depolarization on the diffusion of charge from already depolarized tissue. If there is more depolarized tissue in close proximity to a region of quiescent tissue, then that region will be depolarized faster.

If there is no curvature, the speed of propagation is c_0 space constants per time constant. The constant c_0 is therefore the dimensionless propagation speed for a planar wavefront in homogeneous tissue.

The second term in the eikonal-curvature equation may also be expressed as

(1.20)
$$\sqrt{\nabla u \cdot \mathrm{M}\nabla u}\,\nabla \cdot \left(\frac{\mathrm{M}\nabla u}{\sqrt{\nabla u \cdot \mathrm{M}\nabla u}}\right) = \nabla \cdot (\mathrm{M}\nabla u) - \nabla\sqrt{\nabla u \cdot \mathrm{M}\nabla u} \cdot \frac{\mathrm{M}\nabla u}{\sqrt{\nabla u \cdot \mathrm{M}\nabla u}},$$

where the right-hand side is an anisotropic generalization of the Laplacian of u minus the component of this term in the direction of propagation. The eikonal-curvature equation is parabolic, as it lacks this second derivative in the direction of propagation. Propagation is effectively determined only by information at the wavefront. It is unaffected by boundaries or approaching wavefronts until a collision occurs.

The eikonal equation (1.12) is elliptic, as it contains the full generalized Laplacian. Although it is difficult to comprehend diffusion of excitation time, it is not too surprising that there is a Laplacian in the governing equation, as the propagation process depends heavily on the diffusion of charge. The elliptic eikonal equation is therefore called an *eikonal-diffusion equation*. Under this equation, propagation speed depends not only on information at the wavefront but also on the activity of the surrounding tissue. The constant c_0 is still the dimensionless speed of steady planar wavefront propagation in infinite homogeneous tissue.

It is interesting to investigate three-dimensional analytic solutions to these two governing equations for a wavefront spreading out from the origin in an infinite homogeneous domain. There exist solutions that may be written as functions of only the dimensionless distance from the origin,

(1.21)
$$r := \sqrt{\boldsymbol{x} \cdot \mathrm{M}^{-1}\boldsymbol{x}}.$$

The solutions describe ellipsoidal wavefronts having the same principal axes as the coupling tensor. Both eikonal equations predict that an initial wavefront of this shape will retain the same shape as it propagates. Under the eikonal-curvature equation, the propagation speed $\theta = \rho/\frac{\mathrm{d}u}{\mathrm{d}r}$ satisfies

(1.22)
$$\frac{\tau_{\mathrm{m}}}{\rho}\theta = \mathrm{c}_0\frac{r - \frac{2}{\mathrm{c}_0}}{r},$$

and under the eikonal-diffusion equation,

(1.23)
$$\frac{\tau_{\mathrm{m}}}{\rho}\theta = \mathrm{c}_0\frac{r^2}{r^2 + \frac{2}{\mathrm{c}_0}r + \frac{2}{{\mathrm{c}_0}^2}}.$$

For very large r, both equations predict that the ellipsoid grows at the same constant speed, but for small r the equations differ in the way they predict propagation under large curvature. The eikonal-curvature equation has a change in propagation direction at $r = \frac{2}{\mathrm{c}_0}$, suggesting that the initially depolarized region must have a radius of at least $\frac{2}{\mathrm{c}_0}$ space constants in order for the region to be able to supply enough current to surrounding tissue for propagation to proceed. If the initially depolarized region is smaller than this threshold size, then the equation predicts that the wavefront will retreat and the region will repolarize. The eikonal-diffusion equation, on the other hand, predicts a zero propagation speed only at the origin, suggesting that if enough current has been injected into the tissue to depolarize a region of tissue, then propagation will proceed however small this region may be.

As a wavefront approaches a no-flux boundary or another approaching wavefront, there is less quiescent tissue to drain current from the depolarizing tissue, and thus

the reaction-diffusion model predicts an increase in propagation speed. The eikonal-curvature equation, however, does not include any effects of the boundary or collision on wavefront propagation. Solutions to the eikonal-diffusion equation, on the other hand, much more closely approximate the variations in propagation speed due to another approaching wavefront [8]. The eikonal-diffusion equation is selected as the governing eikonal equation for the work presented here and shall be referred to simply as the eikonal equation.

1.4. Solution spaces and interpolation. In the finite element method the numerical solution $U(\boldsymbol{x})$ is represented by a linear combination of known *interpolation functions* $\psi_i(\boldsymbol{x})$:

$$U(\boldsymbol{x}) := U_i\, \psi_i(\boldsymbol{x}). \tag{1.24}$$

The finite element method selects the unknown parameters U_i in an attempt to approximate the exact solution $u(\boldsymbol{x})$. The domain is divided into a number of *elements* so that, within each element, U depends on only a subset of the parameters. For each element, a local coordinate system $\boldsymbol{\xi}$ is defined, and thus, within that element,

$$U(\boldsymbol{x}(\boldsymbol{\xi})) := U_{\nu(e,j)} \Psi_j(\boldsymbol{\xi}), \tag{1.25}$$

where $\Psi_j(\boldsymbol{\xi})$ are the element's local *basis functions*, and $\nu(e, j)$ is a known function mapping the local parameter j in element e to its corresponding global parameter. The interpolation functions $\psi_i(\boldsymbol{x}(\boldsymbol{\xi}))$ are therefore equal to corresponding basis functions $\Psi_j(\boldsymbol{\xi})$ in elements influenced by U_i and zero elsewhere.

Cubic Hermite elements are used here for discretization of the geometry and dependent variables. One-dimensional basis functions are cubic polynomials that interpolate the value and first derivative of U at the two adjacent nodes. Multidimensional basis functions are obtained from tensor products of the one-dimensional functions. Hermite elements have the advantage over cubic Lagrange elements that all nodes lie on element vertices, and thus parameters can be shared by surrounding elements, and a high-order interpolation is achieved with fewer parameters. This also provides first derivative continuity (C^1) in U.

Because the exact solution u satisfies an elliptic differential equation with predominantly smooth space-dependent coefficients and boundary conditions, it is expected to be sufficiently smooth for the first derivative continuity of the interpolation. Across any surface where the coefficients of the equation are not sufficiently smooth, similar interpolation can be used but with the elements on opposing sides of the surface using separate derivative parameters.

The time and location of excitation wavefront initiation is specified by Dirichlet boundary conditions for excitation time u on Γ_D, where Γ_D denotes the portion of the boundary in which u is known from the initiation process. These boundary conditions are enforced by specifying the values of the parameters U_j that describe U on Γ_D. The set D is defined as the list of indices j for these parameters U_j and their corresponding interpolation functions ψ_j. The set N is defined as the list of indices j for the remaining parameters, which do not influence the value of U on Γ_D and are free to be determined by the finite element method.

The *trial space* S_D^h is defined as the space of possible numerical solutions U sat-

isfying the Dirichlet boundary conditions,

(1.26a)
$$S_{\mathrm{D}}^h := \left\{ V : \exists\, v_j \in \mathbb{R} \text{ for } j \in N \text{ such that (s.t.) } V = \sum_{j \in D} U_j \psi_j + \sum_{j \in N} v_j \psi_j \right\},$$

and the space S_0^h is defined as the space of possible variations in the approximation,

(1.26b)
$$S_0^h := \left\{ V : \exists\, v_j \in \mathbb{R} \text{ for } j \in N \text{ s.t. } V = \sum_{j \in N} v_j \psi_j \right\}.$$

The exact solution u is expected to lie in the Sobolev space $H^1(\Omega)$. The spaces H^1_{D} and $H^1_{\mathrm{D}_0}$ are defined for the exact solution in a manner similar to that for S^h_{D} and S^h_0 for the approximate solution:

(1.27a)
$$H^1_{\mathrm{D}} := \left\{ v \in H^1(\Omega) : v = u \text{ on } \Gamma_{\mathrm{D}} \right\},$$

(1.27b)
$$H^1_{\mathrm{D}_0} := \left\{ v \in H^1(\Omega) : v = 0 \text{ on } \Gamma_{\mathrm{D}} \right\}.$$

Note that for either Lagrange or Hermite interpolation $S^h_0 \subset H^1_{\mathrm{D}_0}$, and, assuming $\sum_{j \in D} U_j \psi_j = u$ on Γ_{D}, $S^h_{\mathrm{D}} \subset H^1_{\mathrm{D}}$. Under these conditions we also have $u - U \in H^1_{\mathrm{D}_0}$.

2. A Petrov–Galerkin finite element method. The space constants in the coupling tensor M for myocardium are several times smaller than the dimensions of the tissue, and thus the advection term in the eikonal equation tends to dominate the diffusion term. Care must be taken in selecting a spatial discretization to prevent oscillatory errors such as those that can occur in numerical solution of the steady-state linear advection-diffusion equation (e.g., [24]). A Petrov–Galerkin finite element method that avoids this problem is developed here for eikonal equation (1.12).

The general Petrov–Galerkin finite element method for determining an approximation U for u may be formulated as finding $U \in S^h_{\mathrm{D}}$ such that

(2.1)
$$B(U, W) = \langle \tau_{\mathrm{m}}, W \rangle \qquad \forall W \in T^h,$$

where T^h is the test space,

(2.2)
$$B(v, w) := \langle \mathrm{c}_0 \sqrt{\nabla v \cdot \mathrm{M} \nabla v}, w \rangle + \langle \mathrm{M} \nabla v, \nabla w \rangle,$$

and $\langle \cdot, \cdot \rangle$ denotes the inner product over the domain Ω.

In this section, a means for estimating the quality of a test space T^h is described, and a set of weighting functions, which form a basis for T^h, is selected on the grounds of keeping the expected error in the solution to a minimum and facilitating numerical solution of the resulting weighted residual equations.

2.1. Approximate symmetrization. The performance of the Galerkin finite element method is poor when diffusion is small due to the asymmetric nature of $B(\cdot, \cdot)$. The object of selecting a Petrov–Galerkin scheme is to choose a mapping from S^h_0 to T^h so that it compensates for this asymmetry. Barrett and Morton [2] showed how an error bound can be derived for a test space T^h if $B(\cdot, \cdot)$ is bilinear. This form is bilinear for the eikonal equation if and only if propagation is in only one direction, but analysis of this simple case leads to one-dimensional weighting functions that can

be extended to higher dimensions. A summary of the key points in the error bound derivation follows.

If $T^h \subset H^1_{\mathrm{D}_0}$, then the exact solution u satisfies the same weighted residual equations (2.1) as the numerical solution U. Therefore, if $B(\cdot,\cdot)$ is bilinear, the error $u-U$ satisfies the orthogonality property

$$B(u-U,W)=0 \qquad \forall W\in T^h. \tag{2.3}$$

The convergence properties implied by this orthogonality property depend on T^h.

If $B_{\mathrm{S}}(\cdot,\cdot)$ is any symmetric continuous coercive bilinear form on $H^1_{\mathrm{D}_0}\times H^1_{\mathrm{D}_0}$, then, from the Riesz representation theorem, there exists a *representer* $R_{\mathrm{S}}\colon H^1_{\mathrm{D}_0}\to H^1_{\mathrm{D}_0}$ such that

$$B(v,w)=B_{\mathrm{S}}(v,R_{\mathrm{S}}w) \qquad \forall v,w\in H^1_{\mathrm{D}_0}. \tag{2.4}$$

Assuming $u-U\in H^1_{\mathrm{D}_0}$, this means that the orthogonality property (2.3) may be written as

$$B_{\mathrm{S}}(u-U,R_{\mathrm{S}}W)=0 \qquad \forall W\in T^h. \tag{2.5}$$

The performance of the method depends on how closely S^h_0 can be approximated by $R_{\mathrm{S}}T^h$. Define the norm $\|\cdot\|_{B_{\mathrm{S}}}$ such that

$$\|v\|^2_{B_{\mathrm{S}}}:=B_{\mathrm{S}}(v,v). \tag{2.6}$$

If $T^h\subset H^1_{\mathrm{D}_0}$ and there exists a constant $\Delta_{\mathrm{S}}\in[0,1)$ such that

$$\inf_{W\in T^h}\|V-R_{\mathrm{S}}W\|_{B_{\mathrm{S}}}\leqslant\Delta_{\mathrm{S}}\|V\|_{B_{\mathrm{S}}} \qquad \forall V\in S^h_0, \tag{2.7}$$

then it is possible to determine a bound for the error in terms of the optimal error and the constant Δ_{S}:

$$\|u-U\|_{B_{\mathrm{S}}}\leqslant\frac{1}{\sqrt{1-\Delta^2_{\mathrm{S}}}}\inf_{Z\in S^h_{\mathrm{D}}}\|u-Z\|_{B_{\mathrm{S}}}. \tag{2.8}$$

The ratio of this bound on the error to the optimal solution error is therefore described by the *error factor* $(1-\Delta^2_{\mathrm{S}})^{-\frac{1}{2}}$. This factor is 1 if the test space T^h is chosen to be equal to $T^{h*}\subset H^1_{\mathrm{D}_0}$ defined such that

$$R_{\mathrm{S}}T^{h*}=S^h_0. \tag{2.9}$$

If the representer R_{S} is known, then the constant Δ_{S} may be calculated for given S^h_0 and T^h. In the Petrov–Galerkin finite element method, S^h_0 and T^h are both of dimension N. Define the N × N matrices A, B, and C with entries

$$\begin{aligned} A_{ij}&:=B_{\mathrm{S}}(R_{\mathrm{S}}w_i,R_{\mathrm{S}}w_j),\\ B_{ij}&:=B_{\mathrm{S}}(R_{\mathrm{S}}w_i,\psi_j)=B(\psi_j,w_i),\\ \text{and}\quad C_{ij}&:=B_{\mathrm{S}}(\psi_i,\psi_j), \end{aligned} \tag{2.10}$$

where the weighting functions w_i (usually based on ψ_i) form a basis for T^h. The error factor $(1-\Delta^2_{\mathrm{S}})^{-\frac{1}{2}}$ is the reciprocal of the square root of the smallest eigenvalue λ of the generalized eigenvalue problem

$$B^{\mathrm{T}}A^{-1}B\boldsymbol{V}=\lambda C\boldsymbol{V}. \tag{2.11}$$

2.2. One-dimensional optimal weighting functions. Consider a one-dimensional problem on a domain of length L with a Dirichlet boundary condition at $x = 0$ and a Neumann boundary condition at $x = \mathrm{L}$ so that wavefront propagation is only in the direction of increasing x. The excitation time $u(x)$ is required to satisfy

$$\mathrm{c}_0\sqrt{\mathrm{M}}u' - \mathrm{M}u'' = \tau_{\mathrm{m}} \quad \text{on } (0, \mathrm{L}), \tag{2.12a}$$

$$u(0) = 0, \qquad \text{and} \qquad \mathrm{M}u'(\mathrm{L}) = 0, \tag{2.12b}$$

where c_0 and M are positive constants and $\cdot'$ denotes the derivative with respect to x.

With propagation in only one direction, the form $B(\cdot,\cdot)$ is bilinear. For constant c_0 and M, it simplifies to

$$B(v, w) := \mathrm{c}_0\sqrt{\mathrm{M}}\langle v', w\rangle + \mathrm{M}\langle v', w'\rangle. \tag{2.13}$$

The second inner product is symmetric and can be used for $B_{\mathrm{S}}(\cdot,\cdot)$:

$$B_{\mathrm{S}}(v, w) := \mathrm{M}\langle v', w'\rangle. \tag{2.14}$$

For this problem the Riesz representer R_{S} may be easily found. From its defining relation (2.4) and definitions of $B(\cdot,\cdot)$ (2.13) and $B_{\mathrm{S}}(\cdot,\cdot)$ (2.14),

$$\begin{aligned} &\mathrm{c}_0\sqrt{\mathrm{M}}\langle v', w\rangle + \mathrm{M}\langle v', w'\rangle = \mathrm{M}\langle v', (R_{\mathrm{S}}w)'\rangle \\ \Longrightarrow\ &\left\langle \mathrm{M}v' \frac{\mathrm{c}_0}{\sqrt{\mathrm{M}}} w + w' - (R_{\mathrm{S}}w)' \right\rangle = 0 \qquad\qquad \forall v, w \in H^1_{\mathrm{D}_0}. \end{aligned}$$

With the Neumann boundary condition at $x = \mathrm{L}$, v is only confined to be zero at $x = 0$, and thus

$$(R_{\mathrm{S}}w)' = \gamma w + w' \qquad \forall w \in H^1_{\mathrm{D}_0}, \tag{2.15}$$

where $\gamma = \frac{\mathrm{c}_0}{\sqrt{\mathrm{M}}}$. This and the boundary condition $(R_{\mathrm{S}}w)(0) = 0$ due to $R_{\mathrm{S}}w \in H^1_{\mathrm{D}_0}$ uniquely determine $R_{\mathrm{S}}w$ for any given w.

If each of the weighting functions w_i^* were chosen such that $R_{\mathrm{S}}w_i^* = \psi_i$, they would form a basis for the optimal test space T^{h*}.

2.3. One-dimensional approximate symmetrization. The expressions for optimal one-dimensional weighting functions w_i^* become rather complicated, particularly for irregular meshes or variable coefficients. Extension to more than one dimension and to the nonlinear eikonal equation does not seem feasible. Instead, therefore, the weighting functions are chosen to be simple combinations of the optimal functions when γ approaches 0 and ∞. For the one-dimensional problem (2.12), the weighting functions are

$$w_i := A_0 w_i^0 + A_\infty w_i^\infty, \tag{2.16}$$

where

$$w_i^0 := \psi_i \qquad \text{and} \qquad w_i^\infty := \gamma^{-1}{\psi_i}', \tag{2.17}$$

and A_0 and A_∞ are functions of the mesh Péclet number,

$$P_{\mathrm{e}} := \frac{\mathrm{c}_0}{\sqrt{\mathrm{M}}} \frac{\mathrm{d}x}{\mathrm{d}\xi}. \tag{2.18}$$

These weighting functions are local and easily evaluated. With the C^1 continuity of cubic Hermite interpolation, they all lie in $H^1_{\mathrm{D}_0}$ except the function corresponding to the derivative at $x = 0$. This will be discussed and corrected below.

2.3.1. Selection of coefficients. The proportionality coefficients A_0 and A_∞ are chosen with the intention of making the factor in the error bound (2.8) as small as possible. This contrasts with the work of Christie et al. [4], in which weighting functions were assembled to cancel truncation errors in difference equations for one-dimensional equal-length elements. The error factor depends on the closeness with which $R_S T^h$ approximates S_0^h as measured by the constant Δ_S in bound (2.7). $(R_S W)'$ is given by expression (2.15), and thus bound (2.7) is equivalent to

$$\inf_{W \in T^h} \|V' - \gamma W - W'\|_{L_2} \leqslant \Delta_S \|V'\|_{L_2} \qquad \forall V \in S_0^h. \tag{2.19}$$

Bounds for error factors in terms of P_e have been obtained for meshes of equal-length one-dimensional linear elements using eigenvalue problem (2.11) (see [17]), but extension to cubic Hermite elements is difficult. Analysis is therefore simplified by considering only the function in S_0^h that is expected to be most poorly approximated by functions in $R_S T^h$.

With cubic Hermite interpolation, each V' for $V \in S_0^h$ is piecewise quadratic with C^0 continuity. V' may have discontinuities in derivatives at element boundaries. If advection dominates (γ is large), each V' must be approximated by a $W \in T^h$. With Galerkin weights ψ_i, each W is piecewise cubic with C^1 continuity and cannot approximate discontinuities in first derivatives. If elements are equally spaced, the function $V \in S_0^h$ with the largest discontinuities in first derivatives of V' relative to $\|V\|_{L_2}$ is

$$\hat{V} := \sum_{j \in N^1} \psi_j, \tag{2.20}$$

where N^1 indexes the interpolation functions corresponding to first derivatives. $\hat{V}'$ is orthogonal to every ψ_i except those corresponding to derivatives at the boundaries. This explains the poor performance of the Galerkin method in advection-dominated problems. Of course, on the other hand, the space spanned by derivative weights ${\psi_i}'$ allows any V' for $V \in S_0^h$ to be represented exactly. If diffusion dominates (γ is small), each V' must be approximated by a W' such that $W \in T^h$. Galerkin weights achieve this exactly because $T^h = S_0^h$. With derivative weights, however, each W' is piecewise linear. These W' are therefore orthogonal to the highest frequency (piecewise-quadratic) function V' such that $V \in S_0^h$. The function that cannot be approximated is again $\hat{V}'$.

Here A_0 and A_∞ are selected so that $\hat{V}'$ is approximated as closely as possible by $\gamma\hat{W} + \hat{W}'$, where $\hat{W}$ is a simple combination of the W's that provide an exact representation when P_e approaches 0 and ∞:

$$\hat{W} := A_0 \hat{V} + A_\infty \gamma^{-1} \hat{V}'. \tag{2.21}$$

The smallest eigenvalue in eigenvalue problem (2.11) is estimated by considering only $\hat{V}$ and $\hat{W}$. This leads to an estimate of the error factor in bound (2.8),

$$\frac{1}{\sqrt{1 - \Delta_S^2}} \approx \frac{\|\hat{V}\|_{B_S} \|R_S \hat{W}\|_{B_S}}{B(\hat{V}, \hat{W})} = \frac{\|\hat{V}'\|_{L_2} \|(R_S \hat{W})'\|_{L_2}}{\langle \hat{V}', (R_S \hat{W})' \rangle},$$

where, from (2.15) and (2.21),

$$(R_S \hat{W})' = \gamma A_0 \hat{V} + (A_0 + A_\infty)\hat{V}' + \gamma^{-1} A_\infty \hat{V}''.$$

If elements are of equal length h and boundary effects are ignored, the integrals may be evaluated to give

$$\frac{1}{\sqrt{1-\Delta_{\mathrm{S}}^2}} \approx \frac{\sqrt{\left(\frac{P_{\mathrm{e}}^2}{42}+1\right)A_0^2+\left(1+\frac{60}{P_{\mathrm{e}}^2}\right)A_\infty^2}}{A_0+A_\infty}. \tag{2.22}$$

This estimate is minimized when

$$\frac{A_\infty}{A_0} = \frac{P_{\mathrm{e}}^2+42}{42(P_{\mathrm{e}}^2+60)}P_{\mathrm{e}}^2. \tag{2.23}$$

With coefficients in this optimum ratio,

$$\frac{1}{\sqrt{1-\Delta_{\mathrm{S}}^2}} \approx \sqrt{\frac{P_{\mathrm{e}}^4+102P_{\mathrm{e}}^2+2520}{P_{\mathrm{e}}^4+84P_{\mathrm{e}}^2+2520}}. \tag{2.24}$$

As expected, the estimate of the error factor approaches 1 as P_{e} approaches ∞ or 0. Its maximum value is $\sqrt{\frac{2\sqrt{70}+17}{2\sqrt{70}+14}} \approx 1.05$, which is predicted at $P_{\mathrm{e}} = \sqrt{6\sqrt{70}} \approx 7.1$.

2.3.2. Dirichlet boundaries. The weighting function corresponding to the derivative at $x = 0$, where the Dirichlet boundary condition is applied, is nonzero on that boundary. This means that the weighted residual equations (2.1) are not satisfied if the exact solution u is substituted for U, and so the error orthogonality property (2.3) does not hold. This is corrected by changing the definition of the derivative term w_i^∞ to include a multiplier $\zeta \in H^1_{\mathrm{D}_0}$ so that all weighting functions lie in $H^1_{\mathrm{D}_0}$:

$$w_i^\infty := \zeta\gamma^{-1}{\psi_i}'. \tag{2.25}$$

The multiplier is chosen to be an exponential ramp,

$$\zeta := \frac{1-\mathrm{e}^{-\gamma x}}{1-\mathrm{e}^{-\gamma h}}, \tag{2.26}$$

in the element adjacent to $x = 0$, and one elsewhere, so that for large γ the behavior of w_i^∞ near $x = 0$ is similar to that of the optimal weighting functions w_i^* in section 2.2. These weighting functions still become optimal when $\gamma \to \infty$.

2.3.3. Verification of error estimates. The estimates of the optimal ratio of A_∞ to A_0 (2.23) and of the error factor (2.24) rely on the assumption that $\hat{V}$ is the function in S_0^h that is most poorly approximated by functions in $R_{\mathrm{S}}T^h$. To investigate the validity of this assumption, error factors were calculated from the smallest eigenvalues of problem (2.11) with the full trial and test spaces for various P_{e} and numbers of elements. The weighting functions w_i in (2.16) were defined using (2.25) for w_i^∞ and (2.17) for w_i^0. $\frac{A_\infty}{A_0}$ was given by (2.23). The resulting error factors are compared with estimates from (2.24) in Figure 2.1.

In all cases investigated, the calculated error factors approached the estimate (2.24) as the number of elements became large. The eigenvector of (2.11) corresponding to the smallest eigenvalue was, in each case, dominated by components corresponding to w_j for $j \in N^1$, which were almost constant in the middle of the domain but smaller nearer the boundaries. This affirms that, without boundary effects, $\hat{V}$ is indeed the most poorly approximated function in S_0^h. Near boundaries, $\hat{V}$ can be approximated better, but the estimate (2.24) based on $\hat{V}$ and ignoring boundary effects appears to provide a good upper bound on the error factor.

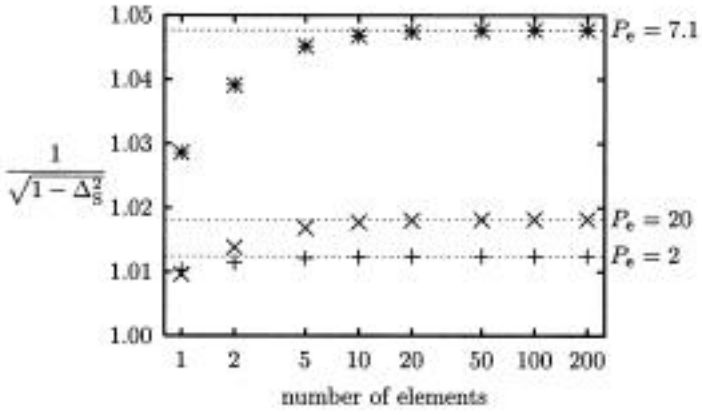


FIG. 2.1. *Comparison of calculated and estimated one-dimensional error factors. The data points are calculated values, and the lines are the estimates from* (2.24).

2.3.4. Variable lengths and coefficients. The terms in the approximately optimal weighting function (2.16) are optimal weights when P_e approaches 0 and ∞, even with unequally spaced elements and variable equation coefficients, but the ratio $\frac{A_\infty}{A_0}$ given in (2.23) is based on constant element lengths and coefficients. With variable lengths or coefficients, the best choices of A_0 and A_∞ are no longer constant. It is assumed that weighting functions (2.16) are still close to optimal if the ratio (2.23) is used, but the variation over x in the total magnitude of A_0 and A_∞ is yet to be determined. Note that with weighting functions (2.16) and propagation only in the direction of increasing x, if boundary effects are ignored, the error orthogonality property (2.3) may be written as

$$\langle u' - U', A_0\gamma \mathrm{M}V\rangle + \langle u' - U', (A_0 + A_\infty)\mathrm{M}V'\rangle + \langle u' - U', A_\infty\gamma^{-1}\mathrm{M}V''\rangle = 0 \qquad \forall V \in S_0^h.$$

If we aim for a small error in the sense of the $\|\cdot\|_{B_S}$ norm, then the left-hand side should resemble $B_S(u-U,V)$. The second term is therefore the desirable term, and its dominance is achieved by appropriate selection of $\frac{A_\infty}{A_0}$ in (2.23). The second term is equivalent to $B_S(u-U,V)$ if

$$A_0 + A_\infty := 1. \tag{2.27}$$

2.4. Extension to three dimensions. For modelling the excitation of the heart, the definitions of the terms in the weighting functions need to be extended to the three-dimensional case with wavefronts travelling in any direction. Weighting functions are still based on the simple combination (2.16) of terms selected for their performance when P_e approaches 0 and ∞, but the Riesz representer theory of section 2.1 can no longer be applied because the form $B(\cdot,\cdot)$ defined in (2.2) is no longer bilinear.

2.4.1. Selection of weight terms. When $P_e \to 0$, $B(\cdot,\cdot)$ becomes bilinear, and thus w_i^0 are defined as the optimal weights in the sense of $\|\cdot\|_{B_S}$, which are still ψ_i. When the advection term is present, its nonlinearity means that the techniques

used in section 2.2 can no longer be used to find a weight that guarantees minimum error in the sense of the $\|\cdot\|_{B_S}$ norm. However, when $P_e \to \infty$, making the residual $c_0\sqrt{\nabla u \cdot M \nabla u} - \tau_m$ orthogonal to the least squares weight $\frac{\nabla U \cdot M \nabla \psi_i}{c_0\sqrt{\nabla U \cdot M \nabla U}}$ minimizes

$$\left\| \sqrt{\nabla U \cdot M \nabla U} - \frac{\tau_m}{c_0} \right\|_{L_2} = \left\| \sqrt{\nabla U \cdot M \nabla U} - \sqrt{\nabla u \cdot M \nabla u} \right\|_{L_2}.$$

These weighting functions compare with the derivative weighting functions of Brooks and Hughes [3] in their solution of the multidimensional steady-state linear advection-diffusion equation with linear elements.

As in the one-dimensional case, a multiplier $\zeta \in H^1_{D_0}$ is included with the least squares weight to ensure that the weighting functions are zero on Dirichlet portions of the boundary. ζ is based on the one-dimensional expression (2.26) and is defined by

$$\zeta := \frac{1 - \exp\left(-\frac{P_e p_\zeta}{k_\zeta}\right)}{1 - \exp\left(-\frac{P_e}{k_\zeta}\right)}, \tag{2.28}$$

where k_ζ is a constant and p_ζ is a simple nonnegative function in $H^1_{D_0}$. In elements adjacent to Dirichlet boundaries, p_ζ is a polynomial function of $\boldsymbol{\xi}$; in other elements, $p_\zeta = 1$. This means that ζ is equal to one over most of the domain, and thus most weights are unaffected by the multiplier. Near Dirichlet boundaries, the weights have similar behavior to one-dimensional optimal weighting functions for large P_e if p_ζ increases from zero at the boundary with slope $|\nabla_{\boldsymbol{\xi}} p_\zeta| = k_\zeta$. A cubic interpolation is used for p_ζ, and k_ζ is set to 3. Away from Dirichlet boundaries, nodal values of p_ζ are set to 1, and derivatives to 0. On Dirichlet boundaries, nodal values are 1, and derivatives are set so that the slope of p_ζ at the boundary is as close to 3 as possible.

The least squares term is discontinuous at wavefront collisions, which makes it difficult to design an integration scheme such that the residuals in the resulting discrete system of nonlinear equations are continuous with respect to the nodal parameters U_j. In order to keep the integration scheme simple, the smooth term,

$$w_i^\infty := \zeta \frac{\nabla U \cdot M \nabla \psi_i}{\sqrt{(1-\alpha_\infty) {c_0}^2 \nabla U \cdot M \nabla U + \alpha_\infty {\tau_m}^2}}, \tag{2.29}$$

is used instead with the constant $\alpha_\infty \in (0,1)$. This term is close to the least squares term when advection dominates and U is close to u. At a collision, however, the denominator remains greater than zero, and so the term vanishes. The best value for α_∞ has not been thoroughly investigated, but $\alpha_\infty = \frac{1}{4}$ seems to work well.

2.4.2. Mesh Péclet number. The one-dimensional expression for P_e in (2.18) included the equation space constant $\sqrt{M}$ and an element spatial scale $\frac{dx}{d\xi}$. In more than one dimension, these quantities are not scalar, and thus the intention is to base P_e on suitable space constants in the direction of propagation.

It is convenient to define at each point in space a dimensionless *natural coordinate system* $\boldsymbol{\upsilon}$ in which the coupling tensor M transforms to the identity matrix and the advection term becomes isotropic:

$$c_0\sqrt{\nabla U \cdot M \nabla U} \equiv c_0 \left| \nabla_{\boldsymbol{\upsilon}} U \right|,$$

where $\nabla_{\boldsymbol{\upsilon}}$ denotes the gradient operator with respect to $\boldsymbol{\upsilon}$ coordinates. The one-dimensional scalar ratio of $\sqrt{\mathrm{M}}$ to $\frac{\mathrm{d}x}{\mathrm{d}\xi}$ corresponds to a multidimensional tensor $\nabla_{\boldsymbol{\upsilon}}\boldsymbol{\xi}$. A scalar quantity is selected from this using the rate of change of $\boldsymbol{\xi}$ arc length with respect to $\boldsymbol{\upsilon}$ arc length in the direction of propagation:

$$\left|\left(\frac{\nabla_{\boldsymbol{\upsilon}} U}{|\nabla_{\boldsymbol{\upsilon}} U|}\cdot\nabla_{\boldsymbol{\upsilon}}\right)\boldsymbol{\xi}\right|. \tag{2.30}$$

A smooth P_{e} is defined by

$$P_{\mathrm{e}} := \frac{\mathrm{c}_0\sqrt{(1-\alpha_\infty)\mathrm{c}_0{}^2\left|\nabla_{\boldsymbol{\upsilon}} U\right|^2+\alpha_\infty\tau_{\mathrm{m}}{}^2}}{\sqrt{(1-\alpha_\infty)\mathrm{c}_0{}^2\left|\left(\nabla_{\boldsymbol{\upsilon}} U\cdot\nabla_{\boldsymbol{\upsilon}}\right)\boldsymbol{\xi}\right|^2+\alpha_\infty\tau_{\mathrm{m}}{}^2\bar{\mu}^{\boldsymbol{\xi}}}}, \tag{2.31}$$

where

$$\bar{\mu}^{\boldsymbol{\xi}} := \frac{1}{3}\frac{\partial\xi_m}{\partial\upsilon_p}\frac{\partial\xi_m}{\partial\upsilon_p} \tag{2.32}$$

(which is an average of the diagonal elements of the coupling tensor in the $\boldsymbol{\xi}$ coordinate system).

2.4.3. Discontinuous derivatives of U. Expressions (2.23) and (2.27) for the coefficients A_0 and A_∞ are only useful if U is C^1 continuous. There are, however, places in the ventricular myocardium where u is not expected to be C^1 continuous [23]. The coefficients in (2.23) and (2.27) and w_i^∞ in (2.29) depend on first derivatives of U, so, with only C^0 continuity in U, the weighting functions (2.16) may be discontinuous. To retain continuity in the weights, A_0 is made constant and p_ζ is set to zero on interelement boundaries where C^1 continuity in U is not expected. The nodal values of p_ζ on interelement boundaries without C^1 continuity are set in the same manner as if they were on Dirichlet boundaries. In this way, $A_0 w_i^0$ retains the C^0 continuity of the interpolation functions, and $A_\infty w_i^\infty$ approaches zero at interelement boundaries where derivatives of U are not expected to be continuous.

With constant A_0, keeping the ratio of A_∞ to A_0 similar to the one-dimensional optimal ratio (2.23) would mean that for large P_{e} the weights would be heavily dependent on the direction of propagation, making the weighted residual equations very nonlinear. Instead, A_0 and A_∞ are defined by

$$A_0 := 1 \qquad \text{and} \qquad A_\infty := \frac{P_{\mathrm{e}}^2}{\mathrm{k}_{\mathrm{lim}}P_{\mathrm{e}}+50}, \tag{2.33}$$

where the constant $\mathrm{k}_{\mathrm{lim}}$ determines the maximum magnitude of the derivative term. It is chosen to be 2 (discussed below). The weighting functions are therefore given by the sum of Galerkin and supplementary weighting functions,

$$w_i = \psi_i + \hat{w}_i, \tag{2.34}$$

where the supplementary weighting functions are defined by

$$\hat{w}_i := \zeta\frac{P_{\mathrm{e}}}{2P_{\mathrm{e}}+50}\frac{\mathrm{c}_0\nabla_{\boldsymbol{\upsilon}} U\cdot\nabla_{\boldsymbol{\upsilon}}\psi_i}{\sqrt{(1-\alpha_\infty)\mathrm{c}_0{}^2\left|\left(\nabla_{\boldsymbol{\upsilon}} U\cdot\nabla_{\boldsymbol{\upsilon}}\right)\boldsymbol{\xi}\right|^2+\alpha_\infty\tau_{\mathrm{m}}{}^2\bar{\mu}^{\boldsymbol{\xi}}}}. \tag{2.35}$$

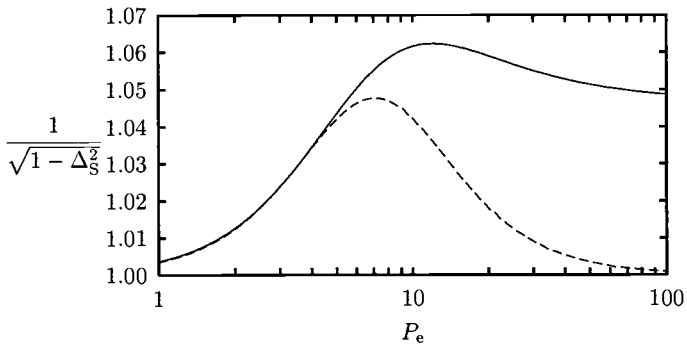


FIG. 2.2. *Comparison of one-dimensional error factor estimates from* (2.22) *with* $\frac{A_\infty}{A_0}$ *determined by the derivative-limited expressions* (2.33) *(solid line) and the optimal expression* (2.23) *(dashed line).*

For large P_e, the magnitude of the supplementary weighting functions, which are dependent on U, is similar to that of the Galerkin weighting functions, which are independent of U. This reduces the effects of nonlinearity in the weighted residual equations, facilitating their solution.

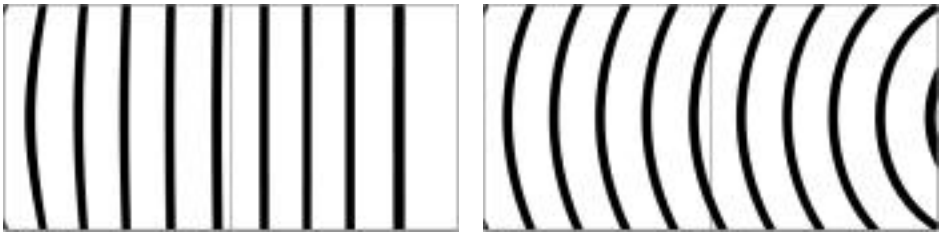
To estimate the error introduced by not using the optimal ratio of A_∞ to A_0 (2.23), one-dimensional error factor estimates for ratios from (2.33) and (2.23) are compared in Figure 2.2. These error factors are calculated from expression (2.22), which is based on constant equation coefficients and assumes a large number of equal-length one-dimensional elements. The maximum predicted error factor with (2.33) is less than two percent greater than the maximum with the optimal ratio (2.23). The constant k_{lim} in (2.33) was set to 2 to keep the nonlinear term in the weight as small as possible while not significantly increasing the expected error factor.

3. No-inflow boundary condition. For large Péclet numbers, the no-flux boundary condition derived from the diffusion of charge is not necessarily enough to sufficiently constrain the solution.

3.1. Inflow boundaries. The lack of stability under large Péclet numbers of the Petrov–Galerkin method developed in the previous section is demonstrated in the situation shown in Figure 3.1. The tissue is stimulated in such a way that the wavefront is initially concave (when viewed from inactive tissue). Note that for $P_e = 10$ the curvature of the wavefront reduces as it propagates across the tissue, but for $P_e = 100$ the curvature increases.

The nature of the solution for $P_e = 100$ is in some ways quite reasonable. The residual in the eikonal equation (1.12) is very small. An inwardly propagating circular wavefront becomes a smaller circle, so an initially concave wavefront becomes more concave. The problem with the solution is that the no-flux boundary condition (1.13) is not satisfied.

The no-flux boundary condition is not very well satisfied on the boundary at the right-hand end of the tissue in the solution for $P_e = 10$ either. Such boundaries where the wavefront extinguishes shall be referred to as *outflow* boundaries. The boundary condition at these boundaries only affects a small boundary layer of tissue, so failure to satisfy the boundary condition does not introduce much error into the



(a) $P_e = 10$ (b) $P_e = 100$

FIG. 3.1. *Excitation contours calculated by the Petrov–Galerkin method for a slice of tissue stimulated unevenly at the left-hand edge. Stimulus times are specified by a quadratic function so that the corners are stimulated first and the center of the edge last. The tissue is represented by two unit square cubic Hermite elements. Equation parameters are selected for unit plane wave speed in any direction. Contours are at intervals of* 0.2.

solution. The boundaries where the wavefronts enter the domain shall be called *inflow* boundaries. The boundary at the left-hand end of the tissue is an inflow boundary because tissue is stimulated on this boundary. In the $P_e = 10$ solution, the no-flux boundary condition on the other boundaries is satisfied very well.

In the $P_e = 100$ solution, the no-flux boundary condition on the boundaries at the top and bottom of the domain is not satisfied. The boundary condition (1.13) is derived from prevention of diffusion of charge across the boundary. For large Péclet numbers, diffusion effects are small, and so the emphasis on satisfying the boundary condition is small. Because the discretization does not allow U to exactly represent u, the numerical method selects a solution that closely satisfies the eikonal equation but almost ignores the boundary condition. As the effects of diffusion are small, the propagation speed should be almost unaffected by curvature and should be almost equal to the unit plane wave speed. This is reflected in the solution through the magnitude of the gradient of activation time which is close to one over the entire domain. Note, however, that the average propagation speed along the top and bottom edges of the domain is about 1.1. This is due to the fact that the method does not recognize that tissue needs to be excited by other excited tissue. It is assumed that the propagation direction is normal to the wavefront, but, because the boundary condition is not strongly enforced, the wavefront normal is not parallel to the boundary. The wavefront is propagating from outside the boundary into the domain, and the boundary is an inflow boundary. Tissue is being excited by nonexistent tissue outside the boundary.

Without a mechanism to prevent wavefronts from entering the domain through unwanted inflow boundaries, excitation can initiate at arbitrary points on the boundary and totally corrupt the numerical solution. This problem occurs when the diffusion term becomes insignificant, and thus the nature of propagation without diffusion is now investigated to determine a prevention mechanism.

3.2. Propagation without diffusion. For large Péclet numbers the numerical scheme behaves as if it is solving the eikonal equation without a diffusion term and without the associated no-flux boundary condition. Without these, the solution to the eikonal equation (1.12) is not unique. To reflect the fact that tissue must be excited by neighboring tissue, the governing equation should instead be

$$\sup_{\boldsymbol{a}\in A(\boldsymbol{x})} \left\{ \lim_{\alpha \searrow 0} \frac{u(\boldsymbol{x}) - u(\boldsymbol{x} - \alpha \boldsymbol{a})}{\alpha} \right\} = \tau_m(\boldsymbol{x}) \qquad \forall \boldsymbol{x} \in \Omega - \Gamma_D, \tag{3.1}$$

where, for m dimensions,

$$A(\boldsymbol{x}) := \left\{\boldsymbol{a} \in \mathbb{R}^m : \boldsymbol{a} \cdot \mathrm{M}^{-1}\boldsymbol{a} = {\mathrm{c}_0}^2;\ \exists\, \alpha \in \mathbb{R} \text{ s.t. } \alpha > 0,\ \boldsymbol{x} - \alpha\boldsymbol{a} \in \Omega\right\}. \tag{3.2}$$

Restricting the vectors $\boldsymbol{a}$ to the set $A(\boldsymbol{x})$ determines the directions in which propagation can occur at the point $\boldsymbol{x}$ and the propagation speeds for these directions.

In regions where the solution is smooth enough, this governing equation is equivalent to

$$\sup_{\boldsymbol{a} \in A} \{\boldsymbol{a} \cdot \nabla u\} = \tau_{\mathrm{m}}. \tag{3.3}$$

(If A is replaced with A_{Ω^0} defined below, this equation is a special case of that used by Falcone, Giorgi, and Loreti [10] in their analysis of front propagation problems.) For a point not on the boundary, the definition of A simplifies to

$$A_{\Omega^0} := \left\{\boldsymbol{a} \in \mathbb{R}^m : \boldsymbol{a} \cdot \mathrm{M}^{-1}\boldsymbol{a} = {\mathrm{c}_0}^2\right\}, \tag{3.4}$$

and the supremum in (3.3) occurs when

$$\boldsymbol{a} = \mathrm{c}_0 \frac{\mathrm{M}\nabla u}{\sqrt{\nabla u \cdot \mathrm{M}\nabla u}}.$$

Away from the boundaries, therefore, (3.3) is equivalent to the eikonal equation (1.12) without a diffusion term.

Without diffusion, there is no Neumann boundary condition, but the notation Γ_{N} will be used for $\partial\Omega - \Gamma_{\mathrm{D}}$, the portion of the boundary in which no Dirichlet boundary condition is applied. For a point on this portion of the boundary, A is equivalent to

$$A_{\partial\Omega} := \left\{\boldsymbol{a} \in \mathbb{R}^m : \boldsymbol{a} \cdot \mathrm{M}^{-1}\boldsymbol{a} = {\mathrm{c}_0}^2,\ \boldsymbol{n} \cdot \boldsymbol{a} \geqslant 0\right\}, \tag{3.5}$$

where $\boldsymbol{n}$ is the unit outward-pointing normal to the boundary.

In order to investigate the nature of the solution to (3.3) near boundaries, consider two points, $\boldsymbol{x}_{\partial\Omega} \in \Gamma_{\mathrm{N}}$ and $\boldsymbol{x}_{\Omega^0} \in \Omega - \partial\Omega$, such that $\boldsymbol{x}_{\Omega^0}$ is an infinitesimal distance from $\boldsymbol{x}_{\partial\Omega}$. As discussed above, the solution at $\boldsymbol{x}_{\Omega^0}$ satisfies

$$\mathrm{c}_0 \frac{\mathrm{M}\nabla u}{\sqrt{\nabla u \cdot \mathrm{M}\nabla u}} \cdot \nabla u = \tau_{\mathrm{m}}.$$

If the solution is smooth enough in the vicinity of the points, one would expect that $\nabla u(\boldsymbol{x}_{\partial\Omega})$ is equal to $\nabla u(\boldsymbol{x}_{\Omega^0})$ and should satisfy the same equation. This is only consistent with (3.3) if

$$\mathrm{c}_0 \frac{\mathrm{M}\nabla u}{\sqrt{\nabla u \cdot \mathrm{M}\nabla u}} \in A_{\partial\Omega},$$

and thus the direction of propagation on Γ_{N} is restricted by

$$\boldsymbol{n} \cdot \mathrm{M}\nabla u \geqslant 0. \tag{3.6}$$

If some diffusion is included, it can be assumed that u is smooth enough that the governing equation becomes

$$\sup_{\boldsymbol{a} \in A} \{\boldsymbol{a} \cdot \nabla u\} - \nabla \cdot (\mathrm{M}\nabla u) = \tau_{\mathrm{m}}. \tag{3.7}$$

The limit of the solution to this equation as the diffusion term vanishes satisfies (3.1).

The no-flux boundary condition on Γ_{N} ensures that $\mathrm{M}\nabla u$ is either parallel to the boundary or zero. The supremum in (3.7) therefore occurs when $\boldsymbol{a} = \mathrm{c}_0 \frac{\mathrm{M}\nabla u}{\sqrt{\nabla u \cdot \mathrm{M}\nabla u}}$ and (3.7) is equivalent to the eikonal equation (1.12). The limit of the solution to the eikonal equation (1.12) and its no-flux boundary condition (1.13) as the diffusion term vanishes satisfies the diffusionless governing equation (3.1).

3.3. A no-inflow boundary term. Although the exact solution of the eikonal equation (1.12) approaches the solution of the diffusionless propagation equation (3.1), the same is not necessarily true for the numerical solution. Unfortunately, with the Petrov–Galerkin method, when diffusion effects become small, they are swamped by discretization errors. The method behaves as if it were solving an eikonal equation without a diffusion term and without the no-flux boundary condition. Without these, the solution is not unique, and so the scheme becomes unstable. To prevent this, the numerical treatment of the advection term needs to more closely represent the corresponding term in (3.1).

With a finite difference method this is easily done by using an upwind difference scheme [18]. Such schemes can select the grid points used in the difference expressions for the advection term so that the excitation time of each grid point is calculated as the expected time for a wavefront to arrive from neighboring grid points with lower excitation times. As there are only grid points in the domain, the wavefront can only arrive from points in the domain, and there are no unwanted inflow boundaries. None of the so-called upwind finite element methods for steady-state problems provide the same restrictions on the solution. Finite element methods only evaluate the advection term at sample points in the domain, and thus the boundaries have no influence on propagation.

The approach used here to stabilize the Petrov–Galerkin solution of the eikonal equation is to add to the weighted residual equations a boundary integral term that encourages the solution to satisfy the boundary inequality (3.6). If this is satisfied, the supremum in (3.3) occurs when $\boldsymbol{a} = \mathrm{c}_0 \frac{\mathrm{M}\nabla u}{\sqrt{\nabla u \cdot \mathrm{M}\nabla u}}$, and thus the residual in the eikonal equation (1.12) is equivalent to the residual in (3.3).

The satisfaction of boundary inequality (3.6) is encouraged by including a penalty term when it is not satisfied. This penalty term is constructed from minimization of an integral over Γ_{N} of the square of a residual,

$$\int_{\Gamma_{\mathrm{N}}} A_{\mathrm{b}}\, r_{\mathrm{b}}^2 \,\mathrm{d}\Gamma, \tag{3.8}$$

where r_{b} is a residual that is zero if and only if (3.6) is satisfied, and A_{b} is a coefficient independent of U. The minimum occurs when the derivatives with respect to each parameter U_i,

$$\int_{\Gamma_{\mathrm{N}}} A_{\mathrm{b}}\, r_{\mathrm{b}}\, \frac{\partial r_{\mathrm{b}}}{\partial U_i} \,\mathrm{d}\Gamma, \tag{3.9}$$

are zero. These integrals are added to the left-hand side of the Petrov–Galerkin discrete equations (2.1) to encourage the numerical solution U to satisfy the boundary inequality (3.6).

The natural coordinate system $\boldsymbol{v}$ of section 2.4.2 may be used to express inequality (3.6) as

$$\boldsymbol{n}^{\boldsymbol{v}} \cdot \nabla_{\boldsymbol{v}} u \geqslant 0, \tag{3.10}$$

where $\boldsymbol{n}^{\boldsymbol{v}}$ is the unit outward-pointing normal to the boundary in this coordinate system. When r_b was defined as the direct residual in this inequality, the penalty term was found to place too much emphasis on satisfying this inequality at the expense of satisfying the eikonal equation. The expression for r_b was therefore selected to be more closely associated with the advection term. Consider the residual

$$r_\mathrm{b} := |\nabla_{\boldsymbol{v}} U| - \sqrt{|\nabla_{\boldsymbol{v}} U|^2 - \min(\boldsymbol{n}^{\boldsymbol{v}} \cdot \nabla_{\boldsymbol{v}} U, 0)^2}. \tag{3.11}$$

If (3.10) is satisfied, this expression is zero. If (3.10) is not satisfied, the expression is essentially the difference between the advection term and what it would be if it were calculated from only the components of $\nabla_{\boldsymbol{v}} U$ in the surface of the boundary. With this residual, the expression $r_\mathrm{b} \frac{\partial r_\mathrm{b}}{\partial U_i}$ has a discontinuity when $|\nabla_{\boldsymbol{v}} U| = \boldsymbol{n}^{\boldsymbol{v}} \cdot \nabla_{\boldsymbol{v}} U$, which corresponds to propagation into the domain normal to the boundary. It is not likely that this will occur, but, to ensure that the discrete equations are smooth enough for solution by Newton's method, the modified residual

$$r_\mathrm{b} := \sqrt{|\nabla_{\boldsymbol{v}} U|^2 + \alpha_\mathrm{b} \tfrac{{\tau_\mathrm{m}}^2}{{\mathrm{c}_0}^2}} - \sqrt{|\nabla_{\boldsymbol{v}} U|^2 - \min(\boldsymbol{n}^{\boldsymbol{v}} \cdot \nabla_{\boldsymbol{v}} U, 0)^2 + \alpha_\mathrm{b} \tfrac{{\tau_\mathrm{m}}^2}{{\mathrm{c}_0}^2}} \tag{3.12}$$

is used. As with α_∞ in section 2.4, a value of $\frac{1}{4}$ is used for α_b.

The boundary integrands are of similar magnitude to the products of the advection term and the weights in (2.34) and (2.35) if they are multiplied by

$$P_\mathrm{e}^\mathrm{b} \left(1 + \frac{P_\mathrm{e}^\mathrm{b}}{2P_\mathrm{e}^\mathrm{b} + 50} \right).$$

To retain the symmetric and positive semidefinite nature of the boundary terms (3.9), an expression that is independent of U is used for the Péclet number:

$$P_\mathrm{e}^\mathrm{b} := \mathrm{c}_0 \left| \boldsymbol{n}^{\boldsymbol{v}} \cdot \frac{\partial \boldsymbol{v}}{\partial \xi_n} \right|, \tag{3.13}$$

where ξ_n is the local element coordinate that does not vary over the boundary. The expression is based on the spatial properties in the direction normal to the boundary instead of in the direction of propagation used in (2.31).

Even with the integrands dimensionally consistent, there is still a difference between the dimensions of the boundary and domain integrals in the order of one spatial dimension. An appropriate multiplier needs to be found for the boundary term to balance the emphasis on satisfaction of the eikonal equation and of the boundary inequality. This should reflect the depth of the region of influence that the boundary terms should have. The parameters U_j that are included in the boundary terms have a significant direct influence on the solution over about half an element. If the boundary terms are given a multiplier that resembles half the width of the element, then the equations involving these parameters should put even emphasis on satisfaction of the domain equation and the boundary inequality. The multiplier is chosen to be

$$\frac{1}{2} \left| \boldsymbol{n} \cdot \frac{\partial \boldsymbol{x}}{\partial \xi_n} \right|,$$

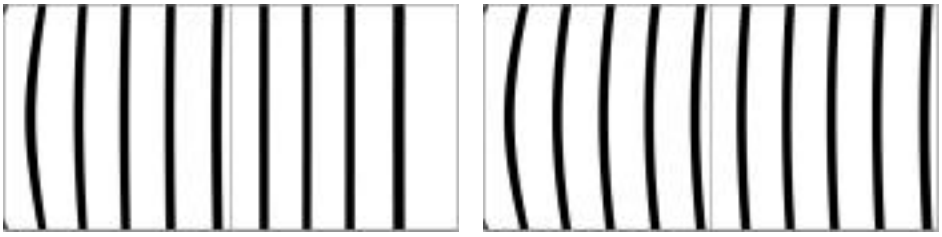


FIG. 3.2. *Excitation contours calculated by the Petrov–Galerkin method with the additional boundary term* (3.9). *The slice of tissue is described in Figure* 3.1.

so that the width of the element is estimated from information at the boundary. The coefficient A_b in the boundary terms (3.9) is therefore

$$A_b := \frac{P_e^b}{2}\left(1 + \frac{P_e^b}{2P_e^b + 50}\right)\left|\boldsymbol{n} \cdot \frac{\partial \boldsymbol{x}}{\partial \xi_n}\right|. \tag{3.14}$$

The numerical solutions obtained by this modified scheme in solving the test problem of section 3.1 are shown in Figure 3.2. For $P_e = 10$, the solution is very similar to the solution in Figure 3.1(a) obtained without the additional boundary term. For $P_e = 100$, the solution satisfies both the eikonal equation (1.12) and boundary condition (3.6) reasonably well, given the coarse discretization. Although the wavefront is not perpendicular to the top and bottom boundaries, the propagation speeds along these boundaries are sensible, and the satisfaction of (3.6) improves with distance from the initiation point.

4. Summary of the method. The numerical method developed for the simulation of excitation propagation in ventricular myocardium uses Newton's method to solve a system of weighted residual equations that are sums of the Petrov–Galerkin weighted residuals of section 2.4 and the no-inflow weighted residual of section 3.3. Newton's method requires a sufficiently good initial guess on which it can iteratively improve. If the diffusion term dominates, the equation is close to linear, and thus almost any initial guess leads to rapid convergence. An initial guess of $U_j = 0\ \forall j \in N$ is sufficient. If the advection term dominates, however, the significant nonlinearities may prevent the method from converging if the initial guess is not good enough. Approximate solutions to equations with more significant diffusion are used as initial guesses for equations with more advection in a numerical continuation method [1] on the continuum of equations,

$$\alpha_c c_0 \sqrt{\nabla u \cdot \mathrm{M} \nabla u} - \nabla \cdot (\mathrm{M} \nabla u) = \alpha_c \tau_m. \tag{4.1}$$

Here α_c is the continuation variable, which is increased from 0 to 1 to transform a diffusion equation into the desired eikonal equation.

From Petrov–Galerkin weighted residual equations (2.1), no-inflow weighted residual equations (3.9), and governing equation continuum (4.1), the weighted residual

equations are

$$\int_{\Omega} \left(\alpha_{\mathrm{c}} (\mathrm{c}_0 \sqrt{\nabla U \cdot \mathrm{M} \nabla U} - \tau_{\mathrm{m}})(\psi_i + \hat{w}_i) + \nabla U \cdot \mathrm{M} \nabla \psi_i - \nabla \cdot (\mathrm{M} \nabla U)\, \hat{w}_i \right) \mathrm{d}\Omega + \int_{\Gamma_{\mathrm{N}}} \left(\boldsymbol{n} \cdot \mathrm{M} \nabla U \hat{w}_i + A_{\mathrm{b}}\, r_{\mathrm{b}}\, \frac{\partial r_{\mathrm{b}}}{\partial U_i} \right) \mathrm{d}\Gamma = 0 \qquad \forall i \in N. \tag{4.2}$$

The boundary inequality residual definition (3.12) is used for r_{b}.

When the value of the continuation variable is less than one, the influence of the diffusion term is increased, and thus the supplementary weights $\hat{w}_i$ and the boundary integral coefficient A_{b} are calculated using the apparent Péclet number. The supplementary weighting functions $\hat{w}_i$ are defined by (2.35), with the Péclet number defined by (cf. (2.31))

$$P_{\mathrm{e}} := \frac{\alpha_{\mathrm{c}} \mathrm{c}_0 \sqrt{(1-\alpha_\infty) {\mathrm{c}_0}^2 \nabla_{\boldsymbol{\xi}} U \cdot \mathrm{M}^{\boldsymbol{\xi}} \nabla_{\boldsymbol{\xi}} U + \alpha_\infty {\tau_{\mathrm{m}}}^2}}{\sqrt{(1-\alpha_\infty) {\mathrm{c}_0}^2 \nabla_{\boldsymbol{\xi}} U \cdot \mathrm{M}^{\boldsymbol{\xi}} \mathrm{M}^{\boldsymbol{\xi}} \nabla_{\boldsymbol{\xi}} U + \alpha_\infty {\tau_{\mathrm{m}}}^2 \bar{\mu}^{\boldsymbol{\xi}}}}. \tag{4.3}$$

The boundary integral coefficient A_{b} is defined by (3.14) with the Péclet number defined by (cf. (3.13))

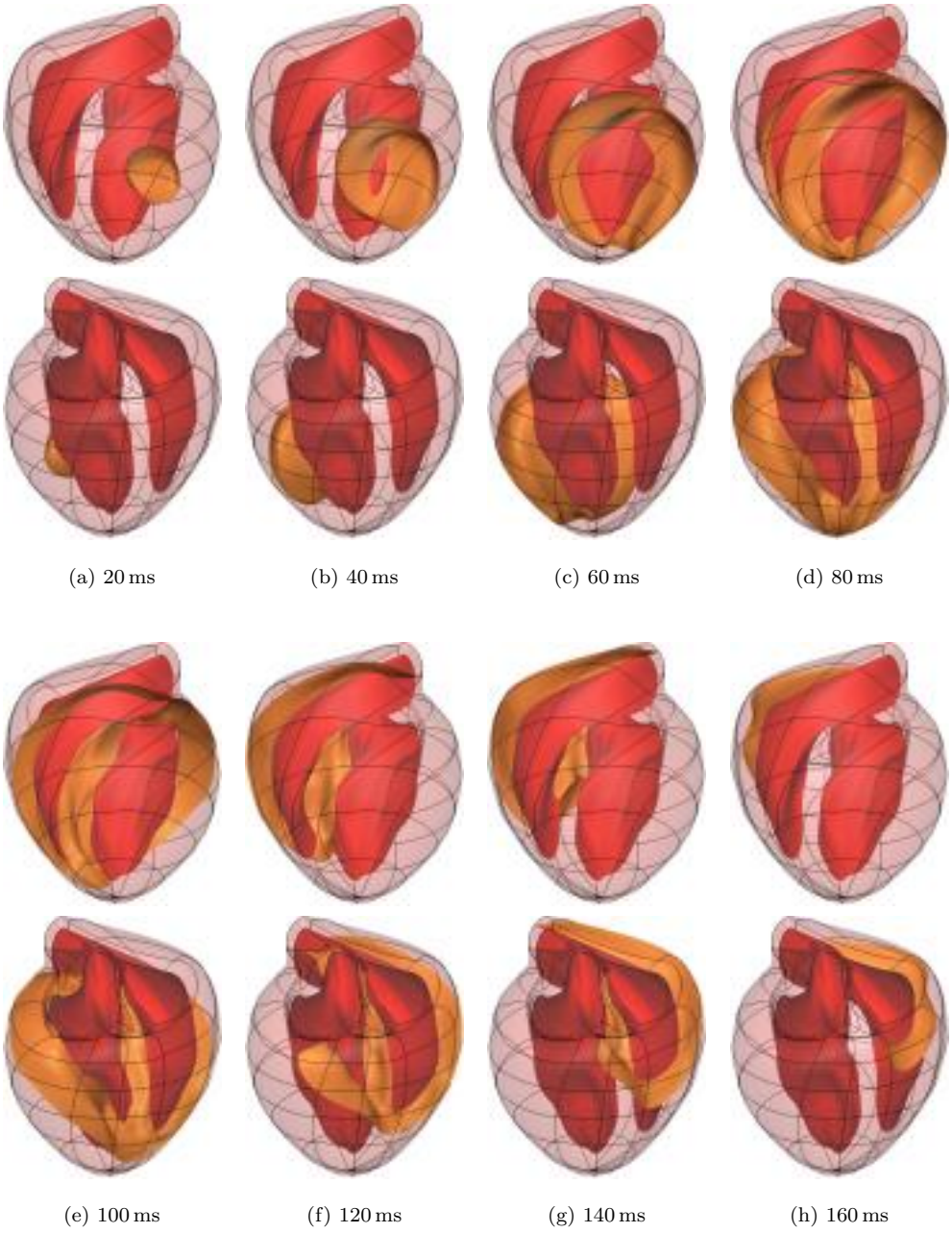
$$P_{\mathrm{e}}^{\mathrm{b}} := \alpha_{\mathrm{c}} \mathrm{c}_0 \left| \boldsymbol{n}^{\boldsymbol{v}} \cdot \frac{\partial \boldsymbol{v}}{\partial \xi_n} \right|. \tag{4.4}$$

The integrals in (4.2) are evaluated using Gauss–Legendre quadrature schemes. A grid of quadrature points with four points in each direction is used within each element. The system of linear equations for each Newton iteration is solved using the generalized minimum residual (GMRES) iterative solver [20] with a simple diagonal preconditioner and no restarts.

5. Simulation. Numerical simulation of excitation propagation through the full canine ventricular myocardium was performed using the method developed here to solve eikonal equation (1.12). The model of the canine ventricular geometry and the selection of material parameters for the governing equation (1.12) are discussed in [23]. Parameters used were $\lambda_{\mathrm{l}} = 0.8\,\mathrm{mm}$, $\lambda_{\mathrm{t}} = \lambda_{\mathrm{n}} = 0.5\,\mathrm{mm}$, $\tau_{\mathrm{m}} = 3\,\mathrm{ms}$, and $\mathrm{c}_0 = 2.5$. There were 2355 degrees of freedom for the dependent variable. The method was programmed primarily in extended FORTRAN 77 as part of the CMISS (an acronym for Continuum Mechanics, Image analysis, Signal processing and System identification) software package. It was executed on one 195 MHz MIPS R10000 processor of a Silicon Graphics Octane.

A point stimulus site was chosen to match the pacing site used for epicardial-sock activation time recordings by Le Grice [16], so that results could be compared with experimental measurements. This site is on the epicardial surface of the anterior aspect of the left ventricular free wall and located at a distance from the apex about one third of that from apex to base.

Snapshots of wavefront locations from the simulation are presented in Figure 5.1. Epicardial isochrones are similar to those from the experimental recordings for times from about 20 ms to 60 ms after stimulation but start to differ considerably outside this interval. Near the stimulus site, experimental recordings showed much slower propagation in the direction transverse to the fibers. The difference in simulation



(a) 20 ms (b) 40 ms (c) 60 ms (d) 80 ms

(e) 100 ms (f) 120 ms (g) 140 ms (h) 160 ms

FIG. 5.1. *Wavefront locations at* 20 *ms time intervals in a simulation of propagation from an epicardial point stimulus at time* 0. *For each sample time, two opposing views are shown.*

results is probably due both to the coarse discretization and to the inability of the eikonal model to reproduce the transient effects near a stimulus. The distance over which slow initial transverse propagation was observed experimentally is less than one quarter of the element length in this direction. For times greater than 65 ms

after stimulation, experimental recordings showed much earlier epicardial excitation, particularly in the more basal and posterior areas on the left ventricular free wall. The region of latest recorded excitation was at the pulmonary conus, which was excited about 125 ms after stimulation. In simulations, the latest excitation occurred about 180 ms after stimulation in the basal posterior region of the right ventricular free wall. The discrepancy is most likely due to the lack of Purkinje fibre representation in the computational model. If the effects of this fast conduction network are not included in the model, results cannot be expected to be realistic. More realistic simulations are presented in [23].

The solution to the discrete system of equations (4.2) was obtained after seventeen Newton iterations and required just less than four minutes of CPU time. One Newton iteration was performed for each increment of the continuation variable α_c until it reached one, then four Newton iterations were required before the relative change in the solution reduced to less than 10^{-5}. The time required for each iteration ranged from 10.7 s to 16.9 s. Of this, the time for calculation of the Jacobian was consistently 7.2 s, but the time for solution of the linear system of equations ranged from 3.0 s when diffusion was significant, to 5.3 s when α_c reached one, to 9.2 s in the final iteration. Most of the remaining 0.5 s in each iteration was spent evaluating the residual in the nonlinear equations.

In the solution of the linear system of equations for each Newton iteration, GMRES iterations were performed until the residual in the linear system was reduced by a factor of 10^{-3}. There was an increasing trend in the number of GMRES iterations required to achieve this, from 111 iterations when diffusion was significant, to 170 when α_c reached one, to 251 in the final Newton iteration. This suggests that the condition number of the Jacobian may increase as the effect of diffusion decreases.

Although convergence in the solution to the nonlinear system was achieved reasonably easily in this simulation with these material parameters, when α_c was increased to represent a reduction in the effects of diffusion, convergence could be achieved for $\alpha_c = 1.06$ but not for $\alpha_c = 1.07$. This means that, if the material parameters were changed so that the relative magnitude of the diffusion term was reduced by more than six percent, artificial diffusion would be required to obtain convergence.

Part of the reason for the inability to achieve convergence when the diffusion term is small may be related to the lack of C^1 continuity in U at certain places in the mesh. A close inspection of the wavefront near the apex in Figure 5.1(c) reveals that the front is starting to form a point as it approaches the apex. This feature of the wavefront vanishes when more diffusion is introduced into the equation.

As discussed in section 3.3, when the diffusion effects become very small, the numerical method behaves as if it were solving an eikonal equation without a diffusion term. The appropriate equation to solve in this situation is the diffusionless propagation equation (3.1). Although the discrepancy between this and the eikonal equation was dealt with on boundaries in section 3.3, it was assumed that inside the domain the residuals in the two equations were equivalent. The residuals are only equivalent, however, if first derivatives are continuous. The C^1 constraint vanishes at the apex because the element widths vanish. Without C^1 continuity, the eikonal equation admits solutions where tissue is not necessarily excited by neighboring tissue. Wavefronts can initiate and spread out from any point in space where C^1 continuity is not enforced. This lack of uniqueness in the solution makes the Jacobian for Newton's method singular and therefore convergence unlikely. If simulations are to be performed with less diffusion, the numerical treatment of the advection term needs to more closely represent its form in (3.1).

If diffusion needs to be added to the equation in order to obtain a stable solution, high-order convergence rates can no longer be expected, and thus one might ask the question of what the advantage of high-order elements might be. To address this question, the cubic Hermite Petrov–Galerkin finite element scheme is compared with a simple finite difference scheme using first-order upwind differences for the advection term. A simple Taylor series analysis of first-order upwind differences shows that the coefficient of numerical diffusion is half the coefficient of the advection term multiplied by the grid point spacing, which resembles $\frac{1}{2}c_0\lambda h$. When doubling the physiological diffusion to stabilize the cubic Hermite scheme, the coefficient of additional diffusion resembles λ^2. In order to make the additional diffusion in the first-order scheme of similar magnitude, the grid point spacing must be given by $h = \frac{2\lambda}{c_0}$, or equivalently, the grid Péclet number P_e must be equal to 2. In the cubic model, the volume average of the geometric mean of the mesh Péclet numbers for each direction is 22, and the maximum Péclet number in any direction at any point is 116. An optimally designed first-order grid would have grid spacings of 0.64 mm in the fibre direction and 0.4 mm in the other directions. The 0.2×10^6 mm^3 myocardial volume would need to be represented by about 2×10^6 grid points. This is a factor of about 10^3 greater than the 2355 degrees of freedom in the cubic Hermite mesh.

6. Conclusions. An efficient computational model has been developed for the excitation process in ventricular myocardium. The need to represent the small-scale ionic activity is eliminated by modelling the excitation process as a propagating wavefront of depolarizing tissue.

A Petrov–Galerkin method using cubic Hermite elements has been developed to enable numerical solution of an eikonal equation for excitation time on a reasonably coarse mesh. The method is a weighted residual method with weights that are linear combinations of Galerkin weights and C^0 continuous supplementary weights based on the derivatives of the interpolation functions in the direction of propagation. For one-directional propagation, the error in the solution is within a small constant factor of the optimal error achievable in the trial space. To estimate the constant factor in the error bound, it was only necessary to consider the function in the trial space with highest frequency first derivative and its corresponding weighting function. A function of the mesh Péclet number was selected for the ratio of the Galerkin and supplementary weights so that this error factor is small for all values of the Péclet number.

For high Péclet numbers, the numerical solution of the eikonal-diffusion equation behaves as if there is no diffusion term. An eikonal equation determines the speed of propagation at each point in space but provides no constraint on the direction of propagation. Without the diffusion term, there is no longer any no-flux boundary condition, and spurious excitation can initiate at any point on the boundary. A no-inflow boundary term has been designed to provide a penalty on such spurious excitation.

Using a continuation method to gradually introduce the nonlinear term of the governing equation, seventeen Newton iterations were required to obtain the solution for a simulation in the full ventricular myocardium. The method showed instabilities when the effect of diffusion was very small, but the level of diffusion required for stability was much less than the level of numerical diffusion that would be introduced in a first-order upwind finite difference scheme with the same number of degrees of freedom.

REFERENCES

[1] E. L. Allgower and K. Georg, *Numerical Continuation Methods: An Introduction*, Springer-Verlag, Berlin, 1990.

[2] J. W. Barrett and K. W. Morton, *Approximate symmetrization and Petrov–Galerkin methods for diffusion-convection problems*, Comput. Methods Appl. Mech. Engrg., 45 (1984), pp. 97–122.

[3] A. N. Brooks and T. J. R. Hughes, *Streamline upwind/Petrov–Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier–Stokes equations*, Comput. Methods Appl. Mech. Engrg., 32 (1982), pp. 199–259.

[4] I. Christie, D. F. Griffiths, A. R. Mitchell, and O. C. Zienkiewicz, *Finite element methods for second order differential equations with significant first derivatives*, Internat. J. Numer. Methods Engrg., 10 (1976), pp. 1389–1396.

[5] P. Colli Franzone and L. Guerri, *Models of the spreading of excitation in myocardial tissue*, in High-Performance Computing in Biomedical Research, T. C. Pilkington, B. Loftis, J. F. Thompson, S. L.-Y. Woo, T. C. Palmer, and T. F. Budinger, eds., CRC Press, Boca Raton, FL, 1993, pp. 359–401.

[6] P. Colli Franzone and L. Guerri, *Spreading of excitation in 3-D models of the anisotropic cardiac tissue.* I. *Validation of the eikonal model*, Math. Biosci., 113 (1993), pp. 145–209.

[7] P. Colli Franzone, L. Guerri, M. Pennacchio, and B. Taccardi, *Spread of excitation in 3-D models of the anisotropic cardiac tissue.* II. *Effects of fiber architecture and ventricular geometry*, Math. Biosci., 147 (1998), pp. 131–171.

[8] P. Colli Franzone, L. Guerri, and S. Rovida, *Wavefront propagation in an activation model of the anisotropic cardiac tissue: Asymptotic analysis and numerical simulations*, J. Math. Biol., 28 (1990), pp. 121–176.

[9] P. Colli Franzone, L. Guerri, and S. Tentoni, *Mathematical modelling of the excitation process in the myocardial tissue: Influence of the fibre rotation on the wavefront propagation and the potential field*, Math. Biosci., 101 (1990), pp. 155–235.

[10] M. Falcone, T. Giorgi, and P. Loreti, *Level sets of viscosity solutions: Some applications to fronts and rendez-vous problems*, SIAM J. Appl. Math., 54 (1994), pp. 1335–1354.

[11] L. S. Green, B. Taccardi, P. R. Ershler, and R. L. Lux, *Epicardial potential mapping. Effects of conducting media on isopotential and isochrone distributions*, Circulation, 84 (1991), pp. 2513–2521.

[12] C. S. Henriquez, *Simulating the electrical behaviour of cardiac tissue using the bidomain model*, Crit. Rev. Biomed. Engrg., 21 (1993), pp. 1–77.

[13] J. P. Keener, *An eikonal-curvature equation for action potential propagation in myocardium*, J. Math. Biol., 29 (1991), pp. 629–651.

[14] J. P. Keener and A. V. Panfilov, *Three-dimensional propagation in the heart: The effects of geometry and fiber orientation on propagation in myocardium*, in Cardiac Electrophysiology: From Cell to Bedside, 2nd ed., D. P. Zipes and J. Jalife, eds., Saunders, Philadelphia, 1995, pp. 335–347.

[15] W. Krassowska and J. C. Neu, *Effective boundary conditions for syncytial tissues*, IEEE Trans. Biomed. Engrg., 41 (1994), pp. 143–150.

[16] I. J. Le Grice, *A Finite Element Model of Myocardial Structure: Implications for Electrical Activation in the Heart*, Ph.D. thesis, The University of Auckland, Auckland, New Zealand, 1992.

[17] K. W. Morton, *Numerical Solution of Convection–Diffusion Problems*, Appl. Math. Math. Comput. 12, Chapman & Hall/CRC, Boca Raton, FL, 1996.

[18] S. Osher and J. A. Sethian, *Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton–Jacobi formulations*, J. Comput. Phys., 79 (1988), pp. 12–49.

[19] R. Plonsey and R. C. Barr, *Mathematical modeling of electrical activity of the heart*, J. Electrocardiol., 20 (1987), pp. 219–226.

[20] Y. Saad and M. H. Schultz, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.

[21] O. H. Schmitt, *Biological information processing using the concept of interpenetrating domains*, in Information Processing in the Nervous System, K. N. Leibovic, ed., Springer-Verlag, New York, 1969, pp. 325–331.

[22] J. A. Sethian, *Theory, algorithms, and applications of level set methods for propagating interfaces*, Acta Numer., 1996 (1996), pp. 309–395.

[23] K. A. Tomlinson, *Finite Element Solution of an Eikonal Equation for Excitiation Wavefront Propagation in Ventricular Myocardium*, Ph.D. thesis, The University of Auckland, Auckland, New Zealand, 2000.

[24] O. C. Zienkiewicz and R. L. Taylor, *The Finite Element Method*, 4th ed., McGraw–Hill, Maidenhead, England, 1994.

SIAM J. APPL. MATH.
Vol. 63, No. 1, pp. 351–372

INTERACTION OF THERMAL EXPLOSION AND NATURAL CONVECTION: CRITICAL CONDITIONS AND NEW OSCILLATING REGIMES*

T. DUMONT†, S. GÉNIEYS†, M. MASSOT†, AND V. A. VOLPERT†

Abstract. In this paper we investigate, numerically as well as analytically, the influence of natural convection on thermal explosion in a two-dimensional square vessel, filled with a reactant mixture, whose vertical walls are adiabatic and horizontal walls are infinitely conducting, preset at an equal temperature T_0. Natural convection enhances the heat losses at the boundaries while large temperatures tend to promote natural convection, thus yielding two competitive phenomena. The governing equations are taken to be the Navier–Stokes equations in the Oberbeck–Boussinesq approximation of low density variations coupled to the heat equation with an exponential chemical source term. This is valid because we consider a 1-step reaction with high heat release, we use the Frank-Kamenetskii transformation under high activation energy asymptotics, and we do not take into account thermo-diffusion as well as the different molar masses of the species. We solve the vorticity-stream function-temperature formulation with an alternating direction numerical method on finite difference approximations. The numerical results show the coexistence of 1-vortex and 2-vortex regimes, from which thermal explosion can occur. New regimes of thermal explosion are found when the Frank-Kamenetskii and Rayleigh parameters are close to the critical conditions for explosion and convection. Periodic-in-time solutions can exist from which thermal explosion can also occur. Linear stability analysis allows us to predict the onset of convection not too close to the explosion limit. Finally, we propose a model problem through a system of two ordinary differential equations which is able to reproduce the bifurcation behavior of the global system close to critical conditions for both explosion and convection.

Key words. thermal explosion, natural convection, stability and bifurcation analysis

AMS subject classifications. 76E30, 76R10, 80A32

PII. S0036139901389501

1. Introduction.

The theory of thermal explosion has been investigated in numerous works (see [7], [15], [17], [21], [22], [24]). Thermal explosion denotes the rapid buildup of energy in systems subject to exothermic reactions, the rate of which rises with temperature. It is of practical importance in problems of fire and explosion safety.

The theory of thermal explosion begins with the works by van 't Hoff who formulated the basic principles of chemical reactions in the end of the 19th century. The first part of the development of this theory was terminated by 1930 due to the work by Semenov with collaborators (see [21], [22], [24]). They explained the physical mechanism of thermal and chain explosion and offered a simple mathematical model in order to get the critical conditions of explosion with a space-independent temperature.

The next step in the development of the theory begins with the works by Frank-Kamenetskii [7], [8] who proposed to consider a space-dependent temperature distribution with Dirichlet boundary conditions in a motionless fluid. He considered a 1-step reaction in the framework of large activation energy asymptotics under the

*Received by the editors May 16, 2001; accepted for publication (in revised form) March 29, 2002; published electronically November 14, 2002.

http://www.siam.org/journals/siap/63-1/38950.html

†Laboratoire de Mathématiques Appliquées de Lyon, UMR 5585 – CNRS, Université Claude Bernard, Lyon I, 69622 Villeurbanne Cedex, France (corresponding author: massot@maply.univ-lyon1.fr). This research was supported by a Young Investigator award from the CNRS (Principal Investigator: M. Massot and V. A. Volpert) and by a BQR Grant from the Université Claude Bernard, Lyon 1 (Project Coordinator: M. Massot).

large heat release assumption, so that the reactant depletion can be neglected. Thermal explosion is viewed as the nonexistence imposibility of a stationary solution. The study of this highly unstationary problem is then reduced to the study of the existence of stationary solutions. This approach was developed in a large number of physical and mathematical works (see [2], [17], [24], and the references therein).

Many other models have been developed further to take into account heterogeneous media [1] or coupling with hydrodynamics [15], [18]. This paper is a contribution to the latter. The influence of free convection on thermal explosion has not received much attention yet. In [15], Kagan et al. considered a forced convection in a two-dimensional (2D) vessel with a large aspect ratio and weakly conducting walls. Surprisingly enough, it is shown that convection can promote explosion. Merzhanov and Shtessel in [18] have investigated the influence of natural convection on thermal explosion in a 2D square vessel with infinitely conducting horizontal walls and adiabatic vertical walls. If the fluid is motionless, the maximal temperature is then at the center of the vessel, and if it is large enough, free convection appears. Free convection can inhibit explosion by enhancing the heat losses at the walls. Critical conditions for both explosion and convection have been investigated numerically as well as analytically with simple models using a single variable: the averaged temperature [18].

In this paper, we study the influence of natural convection on critical conditions for thermal explosion and extend the work by Merzhanov and Shtessel [18]. First, from the modeling point of view, we discuss the assumptions underlying the use of the Navier–Stokes equations for a frozen mixture composition under the Oberbeck–Boussinesq approximation with an exponential source term in the temperature equation. Second, for the considered configuration, with an aspect ratio of one, we show the coexistence of 1-vortex and 2-vortex modes and analyze the corresponding bifurcation diagrams; explosion can occur from both of these regimes. Third, the behavior of the system in a neighborhood of the point where the critical conditions of explosion and convection coincide is studied through comprehensive numerical simulation which was impossible 30 years ago. New regimes of thermal explosion are found when the Frank-Kamenestskii parameter is close to the critical one. Periodic-in-time solutions can exist so that thermal explosion can occur either from these oscillating solutions or from a stationary solution where natural convection is present. Numerical simulations in this neighborhood face several difficulties: high parameter sensitivity, possible oscillations, very slow convergence to stationary or periodic solutions, and possible explosion. Fourth, we perform a linear stability analysis on a simplified model which allows us to obtain a good description of the onset of natural convection far from the critical conditions for thermal explosion. We thus validate the code and gain physical insight into the main difference between the present situation, where the chemistry is coupled to the hydrodynamics and the classical Rayleigh–Bénard problem. Fifth, we propose a simplified model in order to describe the complex bifurcation behavior of the system: a system of two ordinary differential equations for the mean temperature and the maximal stream function. In the spirit of Semenov's theory, this simplified system contains a phenomenological heat losses coefficient. We show how the model problem is able to describe the appearance of stable limit cycles and explain the possible explosion either from oscillating or stationary solutions.

The paper is organized as follows: the modeling assumptions are detailed in section 2. Numerical method and simulations obtained with the comprehensive model are presented in section 3. We then propose simplified models in order to describe the observed phenomena. In section 4 we propose a linear stability analysis of a simplified model in order to find the onset of convection away from the critical conditions of

thermal explosion. Finally, section 5 is devoted to the bifurcation analysis of a model problem close to the critical conditions for both explosion and convection, and section 6 provides a discussion which makes the link between sections 3, 4, and 5, thus completely describing the qualitative behavior of the global system by the model problem.

2. General modeling of the problem.

2.1. Configuration. The configuration studied here is a 2D square vessel whose vertical walls are adiabatic and horizontal walls are infinitely conducting, filled with a reactant mixture. An equal constant temperature T_0 is preset at the horizontal walls. This configuration is similar to the one considered by Merzhanov and Shtessel in [18]. At the starting moment the fluid is at rest at temperature T_0. Either the fluid can encounter thermal explosion (a rapid buildup of temperature modeled in the Frank-Kamenetskii theory by a temperature blow-up) or, after a characteristic time, there exists a slowly varying solution (modeled in the Frank-Kamenetskii theory by a stationary solution at frozen composition) where chemical heat release is balanced by the heat losses through the walls. When no gravity is present, the stationary one-dimensional (1D) concave temperature profile possesses an analytical expression, the characteristics of which are given in subsection 2.2.

In the presence of gravity, the upper half-part of the domain is heated from below by the reacting mixture and beyond a certain stability limit, like in the Rayleigh–Bénard problem [6], [11], convection appears and enhances the heat losses.

The purpose of the present study is to couple the heat equation which governs the classical explosion limits to the hydrodynamics through natural convection. Before dealing with the modeling of the coupled problem, let us come back to the classical results.

2.2. Frank-Kamenetskii theory of thermal explosion. In the Frank-Kamenetskii theory, the explosion limit is determined by studying the existence of stationary solutions to the nondimensional heat equation:

$$\partial_\tau \theta = \partial_{zz}\theta + \mathrm{F_K} \exp\left(\frac{\theta}{1+\dfrac{\mathcal{R}T_0}{E}\theta}\right), \qquad z \in [0,2], \qquad \theta(0) = \theta(2) = 0, \tag{2.1}$$

where T is the temperature and $\theta = (T - T_0)/\mathcal{R}T_0^2/E$ is the temperature scaled by the Frank-Kamenetskii temperature $\mathcal{R}T_0^2/E$; t^* is the time variable and $\tau = t^*/\tau_\kappa$ the nondimensional time associated with the diffusion time $\tau_\kappa = L^2/\kappa_0$; $z = z^*/L$, 2L is the size of the domain and z^* the dimensional space variable; $\mathrm{F_K}$ is the Frank-Kamenetskii parameter defined as the ratio of the diffusion time τ_κ and of a chemical ignition time τ_{ch}:

$$\mathrm{F_K} = \frac{\tau_\kappa}{\tau_{ch}}, \qquad \tau_{ch} = \frac{\mathcal{R}T_0^2}{E\,(T_b - T_0)}\frac{1}{B}\exp\left(\frac{E}{\mathcal{R}T_0}\right), \tag{2.2}$$

where κ_0 is the thermal diffusivity at $T = T_0$, $T_b - T_0 = Q/\rho_0 c_{p0}$ is the adiabatic temperature of reaction, Q is the heat of reaction, ρ_0 is the density of the mixture, c_{p0} is the heat capacity at constant pressure, and at $T = T_0$, E is the activation energy, $\mathcal{R}$ the universal gas constant, and B the frequency factor of the Arrhenius-type reaction rate. Reactant depletion has already been neglected due to the fact that the characteristic time of reactant depletion is much longer than the characteristic time of temperature buildup under the assumption of large heat release, $(T_b - T_0)/T_0 >> \mathcal{R}T_0/E$.

The explosion limit will be described by the only Frank-Kamenetskii parameter. Under the assumptions of large activation energy $E/\mathcal{R}T_0 >> 1$, we can perform the Frank-Kamenetskii transform so that the nonlinear reaction rate in (2.1) is taken to be $\mathrm{F_K}\exp(\theta)$ [8], [24].

Thermal explosion is defined as the nonexistence of a stationary solution of equation (2.1) [7], [8], [24]. It can be shown that there exists a critical value $\mathrm{F_{K}}_c$ such that a stationary temperature profile exists if and only if the Frank-Kamenetskii parameter is below the critical value $\mathrm{F_K} \leq \mathrm{F_{K}}_c$. In this case an analytical expression of the stationary temperature profile as well as an exact formula for the maximal temperature as a function of $\mathrm{F_K}$ are available. The value of $\mathrm{F_{K}}_c$ is approximately 0.88. Beyond this value no stationary solution exists and the system is said to encounter thermal explosion. It is worth noting that the maximal temperature increase at the center of the vessel when a stationary solution exists is of the order of magnitude of the Frank-Kamenetskii temperature, which means that it is small compared to T_0 under the large activation energy assumption.

When gravity is present, fluid motion should be taken into account. For large Rayleigh numbers, free convection will possibly increase heat losses through the walls, thus yielding two competitive phenomena. In order to proceed with the study of such situations, we have to first provide a model for both hydrodynamics, dissipative phenomena, and chemical reactions. This is the purpose of the next section.

2.3. Governing equations. The most general set of equations governing multicomponent reactive fluid mixtures is a system of mixed hyperbolic-parabolic equations describing hydrodynamics, complex dissipation phenomena such as viscous dissipation, multicomponent mass, and heat diffusion, as well as chemistry [5], [12], [13].

The model chosen for the chemistry is a 1-step exothermic reaction in the context of negligible composition effects as in the original Frank-Kamenetskii theory. Consequently, out of the various coupled complex phenomena described by the comprehensive equations, we want only to retain hydrodynamics, viscous dissipation, heat conduction, the effect of gravity, and the heat release of the single chemical reaction as a heat source term. Composition effects due to complex chemistry, multicomponent diffusion, thermal diffusion, as well as stratification of the fluid due to the effect of gravity on species of different molecular weights, are out of the scope of this study. We further assume, for the sake of simplicity, that the heat capacities at constant pressure and at constant volume of the various components of the mixture are constant. The thermal conduction coefficient is assumed constant, $\lambda = \lambda_0$, as well as the thermal diffusivity, $\kappa = \kappa_{_0} = \frac{\lambda_0}{\rho_0 c_{p0}}$, the shear viscosity, $\mu = \mu_0$, and the kinematic viscosity, $\nu = \nu_{_0} = \frac{\mu_0}{\rho_0 c_{p0}}$. We neglect the bulk viscosity. The resulting set of equation is the usual compressible Navier–Stokes equations for a single fluid coupled to the heat equation with a chemical source term.

This system of equations is then considered in the limit of small density variations, also called the Oberbeck-Boussinesq approximation:

$$\partial_\tau\theta + u\,\partial_x\theta + v\,\partial_z\theta = \partial_{xx}\theta + \partial_{zz}\theta + \mathrm{F_K}\exp(\theta), \tag{2.3}$$

$$\partial_\tau u + u\,\partial_x u + v\,\partial_z u = -\partial_x\mathcal{P} + \mathrm{Pr}_{_0}(\partial_{xx}u + \partial_{zz}u), \tag{2.4}$$

$$\partial_\tau v + u\,\partial_x v + v\,\partial_z v = -\partial_z\mathcal{P} + \mathrm{Pr}_{_0}(\partial_{xx}v + \partial_{zz}v) + \mathrm{Pr}_{_0}\,\mathrm{Ra}_{_0}\,\theta, \tag{2.5}$$

$$\partial_x u + \partial_z v = 0. \tag{2.6}$$

The nondimensional quantities involved in the problem are

$$\tau = \frac{t^* w_0}{L}, \quad U = (u,v)^t = \frac{U^*}{w_0}, \quad \mathcal{P} = \frac{\mathcal{P}^*}{p_0}, \quad \theta = \frac{T - T_0}{\mathcal{R}T_0^2/E}, \tag{2.7}$$

$$w_0 = \frac{\kappa_0}{L}, \quad p_0 = \rho_0 \frac{\mathcal{R}}{\bar{m}} T_0, \quad \mathrm{Pr}_0 = \frac{\nu_0}{\kappa_0}, \tag{2.8}$$

where t^*, $U^* = (u^*, v^*)^t$, and $\mathcal{P}^*$ are the dimensional quantities, respectively, time, velocity vector, and perturbation of the pressure, and where $\bar{m}$ is the molar mass of the mixture. Also, Pr_0 is the Prandtl number, x and z are the horizontal and the vertical coordinates, u and v are the horizontal and the vertical components of the scaled velocity, and Ra_0 is the Rayleigh number. Denoting by g the gravity acceleration and recalling that the thermal expansion coefficient α of a perfect gas is the inverse of the temperature and that the typical temperature increase is of the order of the Frank-Kamenetskii temperature $\mathcal{R}T_0^2/E$, the Rayleigh number then reads as $\mathrm{Ra}_0 = \frac{gL^3 \mathcal{R}T_0/E}{\kappa_0 \nu_0}$. Problem (2.3)–(2.6) is considered in the square domain $0 \le x \le 2$, $0 \le z \le 2$ with the boundary conditions

$$x = 0, 2: \ \partial_x \theta = 0, u = 0, \partial_x v = 0; \qquad z = 0, 2: \ \theta = 0, \partial_z u = 0, v = 0. \tag{2.9}$$

The justification of the Oberbeck–Boussineq approximation of small density variations has been investigated in many works, under different assumptions on the state law of the fluid and on the geometry of the vessel [14], [19], [20], [23]. In a recent work by two of the authors [9], it is investigated for a general divariant state law, and a unified approach is presented for both liquids and gases. We restate the fundamental assumptions in the framework of thermal explosion in the following proposition.

PROPOSITION 1. *Let us consider a fluid layer of thickness $2L$ under gravity conditions, the vertical temperature field of which $T_S(z)$ is given by a stationary solution of equation* (2.1) *under the assumption $\mathrm{F_K} \le \mathrm{F_{K}}_c$. Let $\chi_0 = \gamma g L / c_0^2$ and assume that*

$$\chi_0 \ll \mathcal{R}T_0/E \ll 1, \tag{2.10}$$

where c_0 is the velocity of sound in the mixture and γ is the ratio of the heat capacity at constant pressure over the heat capacity at constant volume. Then there exists a static solution (i.e., a solution without convection) of the compressible Navier–Stokes equations $(\rho_S, p_S, T_S, U_S = 0)$ with $\partial_z p_S = -\rho_S g$, $z \in [0, 2L]$, $p_S(2L) = p_0$, $p_S = \rho_S \mathcal{R} T_S / \bar{m}$, which satisfies $(\rho_S - \rho_0)/\rho_0 \ll 1$. Further assume that

$$w_0 / \sqrt{gL\mathcal{R}T_0/E} = O(1), \quad \mathrm{Pr}_0 = O(1), \quad \chi_0 = O(\epsilon^{j+1}), \quad j \ge 1, \tag{2.11}$$

where $\epsilon = \frac{\mathcal{R}T_0}{E}$. Then the solution of the nondimensional compressible Navier–Stokes equations can be formally represented in the form

$$\bar{u} = u + O(\epsilon), \qquad \bar{v} = v + O(\epsilon), \qquad \bar{\theta} = \theta + O(\epsilon), \tag{2.12}$$

$$\bar{p} = 1 + p_{j+1}\epsilon^{j+1} + p_{j+2}\epsilon^{j+2} + O(\epsilon^{j+3}), \qquad \bar{\rho} = 1 - \theta\epsilon + O(\epsilon^2), \tag{2.13}$$

where $\partial_{(x,z)} p_{j+1} = -\chi_0 \epsilon^{-(j+1)} e_z$, e_z being the vertical unit vector directed upwards, and where the variables θ, u, v and $\mathcal{P} = \frac{p_{j+2}\epsilon^{j+2} c_0^2}{\gamma w_0^2}$ satisfy the system (2.3)–(2.6). *The nondimensional pressure $1 + p_{j+1}\epsilon^{j+1}$ and density 1 in equations* (2.13) *correspond to the static solutions p_S and ρ_S.*

3. Numerical simulations. For the model introduced in the previous section, we first present the numerical method and devote the next subsection to the results.

3.1. Numerical method. We consider a $\omega - \psi$ formulation of the equations; the problem (2.3)–(2.6) with (2.9) becomes

$$\partial_\tau \theta + u\partial_x \theta + v\partial_z \theta = \partial_{xx}\theta + \partial_{zz}\theta + \mathrm{F_K}\exp(\theta), \tag{3.1}$$

$$\partial_\tau \omega + u\partial_x \omega + v\partial_z \omega = \mathrm{Pr_0}(\partial_{xx}\omega + \partial_{zz}\omega) + \mathrm{Pr_0}\ \mathrm{Ra_0}\ \partial_x \theta, \tag{3.2}$$

$$\partial_{xx}\psi + \partial_{zz}\psi = -\omega, \tag{3.3}$$

where ψ is the stream function, ω the vorticity,

$$u = \partial_z \psi, \quad v = -\partial_x \psi.$$

The free surface boundary conditions become

$$x = 0, 2 :\ \partial_x \theta = 0, \psi = 0, \omega = 0; \qquad z = 0, 2 :\ \theta = 0, \psi = 0,\ \omega = 0. \tag{3.4}$$

Problem (3.1)–(3.4) is discretized using finite differences with an alternating directions method. For (3.1), this yields

$$\frac{\theta_{ij}^{n+1/2} - \theta_{ij}^{n}}{\tau/2} + u_{ij}^{n}\ \frac{\theta_{i+1,j}^{n+1/2} - \theta_{i-1,j}^{n+1/2}}{2h} + v_{ij}^{n}\ \frac{\theta_{i,j+1}^{n} - \theta_{i,j-1}^{n}}{2h}$$

$$= \frac{\theta_{i+1,j}^{n+1/2} - 2\theta_{i,j}^{n+1/2} + \theta_{i-1,j}^{n+1/2}}{h^2} + \frac{\theta_{i,j+1}^{n} - 2\theta_{i,j}^{n} + \theta_{i,j-1}^{n}}{h^2} + \mathrm{F_K}\exp(\theta_{i,j}^{n})$$

and

$$\frac{\theta_{ij}^{n+1} - \theta_{ij}^{n+1/2}}{\tau/2} + u_{ij}^{n}\ \frac{\theta_{i+1,j}^{n+1/2} - \theta_{i-1,j}^{n+1/2}}{2h} + v_{ij}^{n}\ \frac{\theta_{i,j+1}^{n+1} - \theta_{i,j-1}^{n+1}}{2h}$$

$$= \frac{\theta_{i+1,j}^{n+1/2} - 2\theta_{i,j}^{n+1/2} + \theta_{i-1,j}^{n+1/2}}{h^2} + \frac{\theta_{i,j+1}^{n+1} - 2\theta_{i,j}^{n+1} + \theta_{i,j-1}^{n+1}}{h^2} + \mathrm{F_K}\exp(\theta_{i,j}^{n}),$$

where τ is the time step and h the space step. Other equations are discretized similarly.

A continuation method is used to study existence and stability of stable stationary and periodic-in-time solutions. Among the three parameters of the problem $\mathrm{Pr_0}$, $\mathrm{Ra_0}$, and $\mathrm{F_K}$, the last two are more essential. So far, for a fixed value of the Prandtl number ($\mathrm{Pr_0} = 1$) we vary $\mathrm{Ra_0}$ and $\mathrm{F_K}$.

3.2. Numerical results. This subsection can be considered at the heart of our work: it introduces all the new physical behaviors of the considered configuration. We first present the state of the art by recalling the results of Merzhanov and Shtessel [18]. This allows us to roughly identify zones in the $\mathrm{Ra_0} - \mathrm{F_K}$ parameter plane where explosion is to be found. Convection appears as a supercritical bifurcation from a static solution when $\mathrm{F_K} < \mathrm{F_{K}}c$; we present the first results validating our code and emphasize that 1-vortex and 2-vortex convection regimes can coexist for a given value of the Frank-Kamenetskii number. We then focus on the behavior of the system around the bifurcation point of codimension 2 at $\mathrm{F_{K}}c$ and $\mathrm{Ra}_c(\mathrm{F_{K}}c)$, and identify new oscillating regimes, which can be periodic or lead to explosion. We finally present some results on the unstationary behavior of the system in the last subsection.

3.2.1. The diagram of Merzhanov and Shtessel. Let us first come back to the work by Merzhanov and Shtessel and summarize their results. They have shown by direct numerical simulation the existence of four regions in the Rayleigh/Frank-Kamenetskii parameter plane. In region I, there exists a stationary solution in a static fluid while in region II, a stationary solution exists in a moving fluid. In region III, thermal explosion arises from a moving fluid, whereas in region IV thermal explosion arises from a static one. The boundary between regions I and IV is predicted by the Frank-Kamenetskii theory and given by $\mathrm{F_K}_c \approx 0.88$. The interesting scenario takes place in region II for the Frank-Kamenetskii parameter bigger than the critical value. Convection enhances the heat losses at the boundaries, thus inhibiting explosion. The limit between III and IV is difficult to locate since convection is always enhanced by the blow-up of the temperature so that it is difficult to decide if convection started just before or during thermal explosion. These regions were characterized in [18] and several simple models were provided in order to describe the behavior of the solutions in the various regions. However, the only convection regime considered in this paper is a 2-vortex regime. Besides, because of computation limitations, the detailed behavior at the point where the four regions merge is not studied. We first show, in the next subsection, that there exists also a 1-vortex solution. Moreover, 1- and 2-vortex regimes can coexist; explosion can occur from both of these regimes.

3.2.2. Two regimes: One or two vortices. Consider first the case where $\mathrm{F_K} < \mathrm{F_K}_c$. There exists a static solution. For Ra_0 sufficiently large, it loses stability and a convective regime appears. The bifurcation diagram is usual for convection problems. The maximum of the stream function $\psi_{\max}$ is roughly proportional to $\sqrt{\mathrm{Ra}_0 \bar{\theta} - \mathrm{Ra}_c}$ as represented on Figure 1 and as will be shown in subsection 5.5, where Ra_c is the critical value of the Rayleigh number and $\bar{\theta}$ is the averaged temperature in the vessel. The supercritical bifurcation can also be observed on a $(\psi_{\max}, \mathrm{Ra}_0)$-plane. At the same time the mean temperature decreases when Ra_0 increases. As can be expected, an increase of $\mathrm{F_K}$ results in an increase of the mean temperature and of the maximum of the stream function. It decreases the critical Rayleigh number (see Figure 1).

In the present configuration (2D square domain), only 1-vortex and 2-vortex solutions are observed. The solution bifurcating from the static solution has only one vortex which fills practically the whole domain being slightly moved to its upper part. For Ra_0 larger than a critical value $\mathrm{Ra}_{1,2}$, a transition from a 1-vortex solution to a 2-vortex solution is observed. Decreasing Ra_0 and using the 2-vortex regime as an initial condition, we observe a reverse transition to a 1-vortex solution but for $\mathrm{Ra}_{2,1} < \mathrm{Ra}_{1,2}$. Note that the number and position of the vortices depend on the geometry of the vessel and that, in the present situation, there exists a parameter range $\mathrm{Ra}_0 \in [\mathrm{Ra}_{2,1}, \mathrm{Ra}_{1,2}]$, where the two regimes 1-vortex and 2-vortex coexist.

Now consider the case $\mathrm{F_K} > \mathrm{F_K}_c$, which means that no static solution is to be found. For Ra_0 large enough, a 2-vortex solution exists (see Figure 2; $\mathrm{F_K} = 0.9$ and $\mathrm{F_K} = 1.0$). For $\mathrm{F_K} = 0.9$, decreasing Ra_0 leads to a transition to a 1-vortex regime and finally leads to explosion. However, for $\mathrm{F_K} = 1.0$, decreasing Ra_0 directly leads to an explosion from this 2-vortex regime. It can be shown by a proper choice of the initial condition that a 1-vortex solution can still be reached; if we increase Ra_0 from this 1-vortex solution, we jump onto the 2-vortex solution; a decrease of Ra_0 leads to an explosion from this 1-vortex regime. Again, one can observe the existence of a Ra_0 range for which the 1-vortex regime and the 2-vortex regimes coexist. It has to be noticed that the symbols used in Figure 2 for a 1-vortex or a 2-vortex regime do not

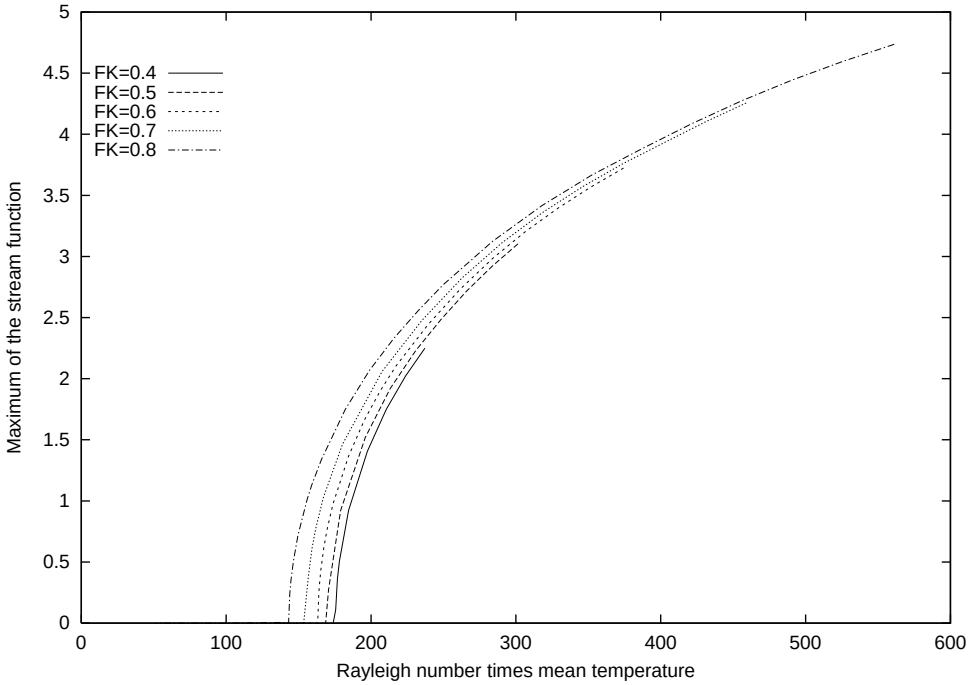


FIG. 1. *Convective regime bifurcating from the static solution for five values of the* $\mathrm{F_K}$ *parameter below the critical conditions: maximum of the stream function versus effective Rayleigh number* $\mathrm{Ra}_0\bar{\theta}$.

reproduce any real temperature or stream function field.

For Frank-Kamanetskii numbers slightly above the critical value for the existence of a stationary solution, a general scenario can be proposed. Decreasing Ra_0 leads to a decrease of the maximum of the stream function (and to an increase of the mean temperature) until a minimum is reached (see Figure 2). A further decrease of Ra_0 yields a great increase of the mean temperature and of the maximal stream function until explosion is obtained and the stationary solution disappears (see Figure 2).

3.2.3. Oscillating thermal explosion around critical conditions. From the previous subsections, one could think that all scenarios have been covered. However, a closer look at the neighborhood of the point $(\mathrm{F_{K}}_c, \mathrm{Ra}_c(\mathrm{F_{K}}_c))$ reveals a rich nonlinear behavior.

We consider three values of $\mathrm{F_K}$, two above $\mathrm{F_{K}}_c$, 0.8775 and 0.88, and one below, 0.875. Starting from values of Ra_0 for which a convective stationary solution is observed numerically, we continue this solution using Ra_0 as a bifurcation parameter.

For $\mathrm{F_K} = 0.88$, decreasing Ra_0 successively yields a decrease and then an increase of $\theta_{\max}$, appearance of a periodic-in-time mode, an increase and then a decrease of the oscillations amplitude, reappearance of a stationary solution, and explosion. For the periodic solution, the maximal temperature also oscillates and can reach the value $\theta_{\max} \simeq 2.0$, which is much larger than the critical value 1.2 of the Frank-Kamenetskii theory (see Figure 3).

For $\mathrm{F_K} = 0.8775$, one again observes the appearance of a periodic solution with an increase of the oscillations, but then there is an explosion from the oscillating solution

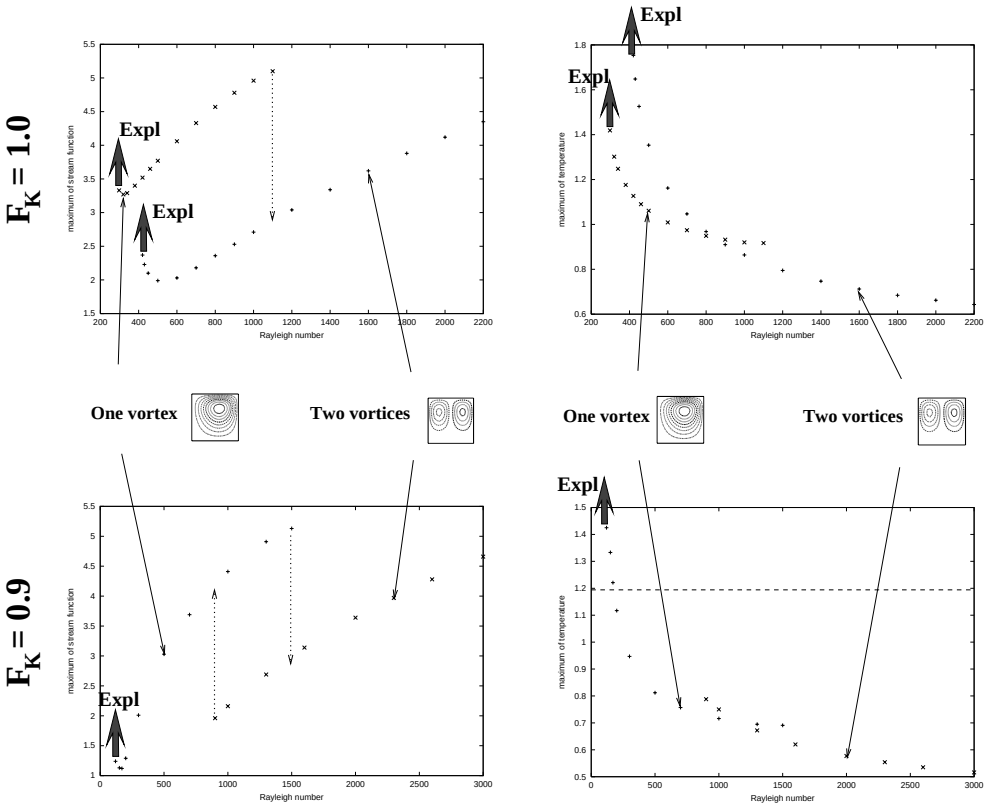


FIG. 2. *Explosion from* 1*-vortex or* 2*-vortex regimes for* $\mathrm{F_K} = 0.9, 1.0$*; on the left the maximal stream function is represented; on the right, the maximal temperature is represented; explosion is indicated by a vertical arrow.*

(see Figure 3). Using proper initial conditions permits us to follow the end of this branch: oscillations decrease, stationary solution appears again, and then explosion occurs from the stationary solution (see Figure 3). Let us mention that for this value of $\mathrm{F_K}$, the period of oscillations is very large as will be seen in the next subsection.

For $\mathrm{F_K} = 0.875$, the branch on the right reaches the nonconvective stationary solution at $\mathrm{Ra}_0 = 163$. Further decreasing the value of the Rayleigh number results in a straight line on Figure 3, since no convection is present and we plot only the maximal value of the temperature profile without convection for various values of Ra_0. We could think that for the parameter range $\mathrm{Ra}_0 \in [110, 163]$, the previous branch is the only branch. However, two striking scenarios also appear for $\mathrm{F_K} = 0.875$ for some Rayleigh numbers smaller than Ra_c. Besides the stable static solution for this value of $\mathrm{F_K}$ below the critical value, other stable solutions exist, both stationary or with periodic oscillations, and can be reached by using proper initial conditions (see Figure 3). One observes thermal explosion either from a periodic oscillating solution (increasing Ra_0) or from a stationary convective solution (decreasing Ra_0). This is all the more surprising since intuition predicts that convection should inhibit explosion by enhancing the heat losses through the boundaries. In this case, convection promotes explosion. An explanation for the presence of this branch of solutions can be suggested. In the situation without convection, for $\mathrm{F_K} < \mathrm{F_K}_c$, there exists two branches of solutions, one stable and one unstable with higher temperature; these

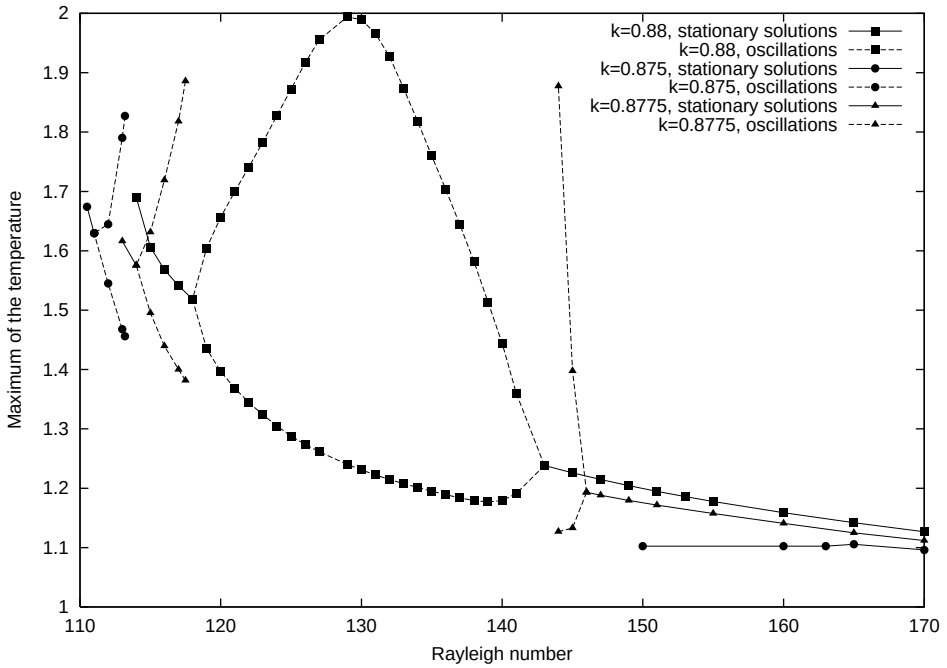


FIG. 3. *Stationary solutions, periodic-in-time solutions, and oscillating explosion. For stationary solutions, only one point is plotted associated with the maximal temperature versus* Ra_0*; for periodic solutions, we have plotted the maximum and the minimum of the maximal temperature in the vessel versus* Ra0 *(the branches of stationary unstable solutions corresponding to these periodic solutions are not represented).*

two branches join at the bifurcation point $\mathrm{F_K} = \mathrm{F_{K}}_c$. The piece of the stable branch observed for small Rayleigh numbers at $\mathrm{F_K} = 0.875$ could be the stabilization by convection of this unstable nonconvective branch for $\mathrm{F_K}$ close to $\mathrm{F_{K}}_c$.

3.2.4. Unstationary behavior. The purpose of this subsection is to describe the unstationary behavior of the oscillating solutions previously identified. We focus on value $\mathrm{F_K} = 0.8775$, the richest scenario.

We first present the growth of the oscillations when we start from a stationary convective solution at $\mathrm{Ra}_0 = 147$ and set at $t = 0$, $\mathrm{Ra}_0 = 145$. Figure 4 describes the path to the periodic oscillating solution. The very slow growth is related to the closeness to the Hopf bifurcation. More interesting is the analysis of the periodic solution; almost all the time (about 90%) the solution is close to a static one and periodically, with a very long period compared to the nondimensional time, it almost explodes, and $\psi_{\max}$ and $\bar{\theta}$ reach their maximal value.

The second point is related to the thermal explosion from a periodic oscillating solution on the other side of the branch at $\mathrm{Ra}_0 = 117.5$. In Figure 5(a) we have represented the mean temperature, the maximal temperature, and the maximal stream function, first for the periodic oscillating solution at $\mathrm{Ra}_0 = 117.5$. We then set, at $t = 0$, $\mathrm{Ra}_0 = 118$, and represent on the second figure Figure 5(b) the evolution of the previous three quantities. The system goes on its periodic trajectory for another period and then it explodes. The explosion behavior is very different from the one on the other side of the branch.

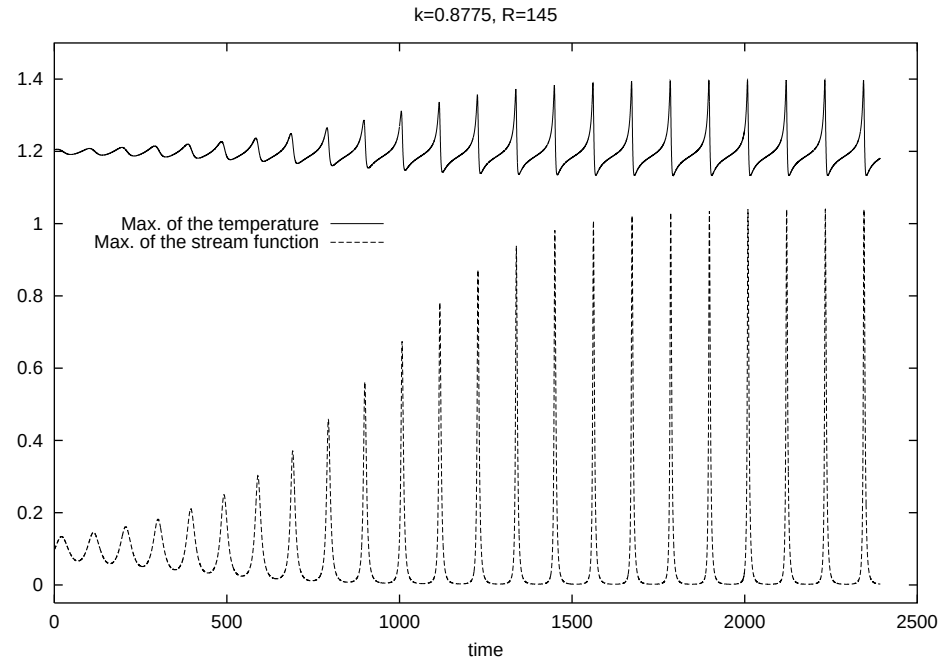
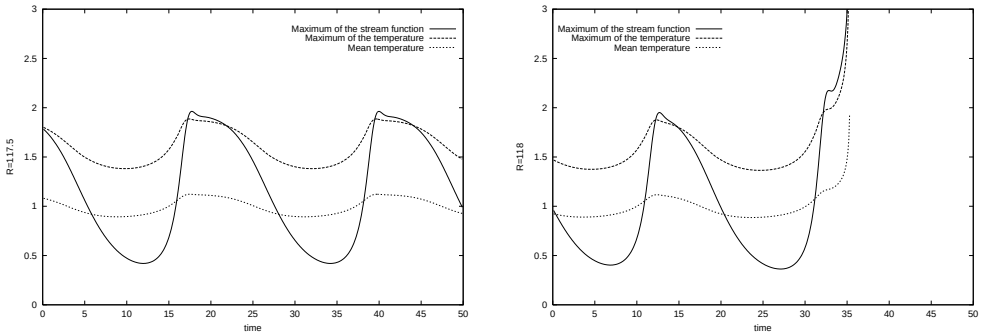


FIG. 4. *Growth and stabilization of the oscillation at* $\mathrm{Ra}_0 = 145$*: maximal temperature (solid line) and maximal stream function (dashed line) versus time.*



(a) (b)

FIG. 5. (a) *Periodic solution at* $\mathrm{Ra}_0 = 117.5$ *and* (b) *oscillating explosion at* $\mathrm{Ra}_0 = 118$ *(solid line: maximal stream function, dotted line: mean temperature, dashed line: maximal temperature).*

In conclusion, we point out that taking into account fluid motion when studying explosion reveals new phenomena—periodic-in-time solutions and explosion either from these solutions or from stationary solutions. These features happen for $\mathrm{F_K}$ and Ra_0 close to their critical value. A simplified system is proposed in section 5 in order to investigate, analyze, and reproduce these phenomena.

4. Linear stability analysis for a simplified model for $\mathbf{F_K < F_{Kc}}$. The purpose of this section is to perform a linear stability analysis on a simplified model, which allows us to obtain a good estimate of Ra_c (the critical value of Ra_0 for fixed $\mathrm{F_K} < \mathrm{F_{K}}_c$) for the onset of natural convection. It is another validation of the code, but

more specifically it allows us to gain physical insight on the main difference between the present situation, where the chemistry is coupled to the hydrodynamics, and the classical Rayleigh–Bénard problem.

The linear stability analysis of system (3.1)–(3.4) around the nonconvective solution is well known for a nonreactive fluid in the Rayleigh–Bénard configuration (see, e.g., [3], [4], [6], [10], [11]).

We start in the same manner; problem (2.3)–(2.6), (2.9) is linearized about a static solution $\theta_S(z)$, $u = v = 0$, which depends only on the vertical coordinate. Using simple algebra, pressure can be eliminated; it yields

$$\partial_\tau \theta = \Delta\theta + F'(\theta_S)\,\theta - \theta_S' v, \tag{4.1}$$

$$\partial_\tau \Delta v = \mathrm{Pr}_0\,\Delta\Delta v + \mathrm{Pr}_0\,\mathrm{Ra}_0\,\partial_{xx}\theta, \tag{4.2}$$

where $F(\theta_S)$ denotes the reaction rate. In the present section, it is assumed to be affine with respect to θ_S,

$$F(\theta_S) \simeq \mathrm{F_K}\,(1 + \theta_S), \tag{4.3}$$

which is valid for small θ_S and leads to simple computations. The boundary conditions are the same as in the classical analysis [6], [11].

We look for the solution of (4.1)–(4.3) of the form $\theta(x,z,t) = \tilde\theta(z)e^{-\lambda\tau}\cos(kx)$, $v(x,z,t) = \tilde v(z)e^{-\lambda\tau}\cos(kx)$, where $k = \pi m/2$, $m = 1, 2, \ldots$, which yields the eigenvalue problem

$$-\lambda\tilde\theta = \tilde\theta'' - k^2\tilde\theta + \mathrm{F_K}\,\tilde\theta - \theta_S'\tilde v, \tag{4.4}$$

$$-\lambda(\tilde v'' - k^2\tilde v) = \mathrm{Pr}_0\,(\tilde v'''' - 2k^2\tilde v'' + k^4\tilde v) - \mathrm{Pr}_0\,\mathrm{Ra}_0\,k^2\tilde\theta, \tag{4.5}$$

with the boundary conditions

$$z = 0, 2 : \tilde\theta = 0,\ \tilde v = \tilde v'' = 0. \tag{4.6}$$

The convective instability boundary can be found from the condition that the eigenvalue λ with the minimal real part is zero. In the present study, these eigenvalues are not computed exactly but are approximated.

With the approximation (4.3) the static solution $\theta_S(z)$ can be found explicitly. However, even in this case, problem (4.4)–(4.6) does not give a simple expression for λ. Hence $\theta_S(z)$ is approximated by the first term of its Fourier series (it can be verified that the second term is already essentially less than the first one): $\theta_S(z) \simeq \frac{2}{\pi}\sigma\sin(\frac{\pi z}{2})$ so that $\theta_S'(z) \simeq \sigma\cos(\frac{\pi z}{2})$ with $\sigma = \dfrac{8\mathrm{F_K}}{\pi^2 - 4\mathrm{F_K}}$. Note that σ is well defined if $\mathrm{F_K} < \pi^2/4$, which is satisfied for $\mathrm{F_K} < \mathrm{F}_{\mathrm{K}c}$.

We look for the solution $\tilde\theta$ and $\tilde v$ of (4.4)–(4.6) in the form of Fourier series

$$\tilde\theta = \sum_{n=1}^{+\infty} b_n \sin\left(\frac{\pi n z}{2}\right), \qquad \tilde v = \sum_{n=1}^{+\infty} c_n \sin\left(\frac{\pi n z}{2}\right)$$

and assume $\lambda = 0$. It leads to the infinite system for the coefficients c_n:

$$c_2 = -\gamma_1 c_1,$$

$$c_{n-1} + c_{n+1} = -\gamma_n c_n, \qquad n = 2, 3, \ldots.$$

TABLE 1
Values of Ra_c *given by* (4.7) *and by the numerical simulations.*

$\mathrm{F_K}$	0.4	0.5	0.6	0.7	0.8
numerical Ra_c	1025	740	547	408	280
analytical Ra_c	922	735	558	446	340
$658/\theta^{\max}$	2680	1999	1532	1178	879

Here

$$\gamma_n = \frac{\mathrm{Ra}_n}{\mathrm{Pr}_0 \mathrm{Ra}_0}, \quad \mathrm{Ra}_n = \frac{2\mathrm{Pr}_0}{\sigma k^2}\left(\left(\frac{\pi n}{2}\right)^2 + k^2\right)^2 \left(\left(\frac{\pi n}{2}\right)^2 + k^2 - \mathrm{F_K}\right).$$

The critical Rayleigh number Ra_c is given by the condition that the determinant of this infinite system of equations equals zero. To find Ra_c explicitly we have to truncate the system. Taking only two equations ($c_3 = 0$) leads to

$$\mathrm{Ra}_c^2 = \mathrm{Ra}_1 \mathrm{Ra}_2, \tag{4.7}$$

while for three equations ($c_4 = 0$),

$$\mathrm{Ra}_c^2 = \frac{\mathrm{Ra}_1 \mathrm{Ra}_2 \mathrm{Ra}_3}{\mathrm{Ra}_1 + \mathrm{Ra}_3},$$

which is close to $\mathrm{Ra}_1 \mathrm{Ra}_2$ if $\mathrm{Ra}_3 >> \mathrm{Ra}_1$.

It can be easily verified that if $\mathrm{Ra}_{n+1} \gg \mathrm{Ra}_n$, $n = 2, 4, \ldots$, then each successive approximation will be close to the previous one. In the case $k = \pi/2$ ($m = 1$), these conditions are satisfied and one can use the approximation (4.7).

In Table 1, the values of Ra_c given by (4.7) are compared with the numerical results for various values of $\mathrm{F_K}$. It is interesting to note that the agreement is pretty good far from the critical conditions for explosion; it gets worse when we get really close to $\mathrm{F}_{\mathrm{K}c}$ even if the estimate for $\mathrm{F_K} = 0.8$ is still correct up to 20%.

Note that these values of Ra_c are much smaller than $658/\theta^{\max}$ as stated in Merzhanov and Shtessel [18] and given in the third row. (658 is the critical Rayleigh number for the Rayleigh–Bénard problem in an infinite layer with free surface boundary conditions.) Actually, $658/\theta^{\max}$ would be a good approximation of Ra_c if the temperature profile were linear in the upper part of the domain, as it is in the original Rayleigh–Bénard problem. The present study shows that the nonlinear temperature profile yields quite different values of Ra_c.

5. Model problem, bifurcation analysis. In the previous section, we have considered the behavior of the system away from the explosion limit. We investigate, in the present section, a model problem to describe the observed oscillatory instability, the Hopf bifurcations, and the oscillating explosion close to the critical conditions for both convection and explosion (see subsection 3.2.3).

A similar attempt was performed in [18]; however, they wanted only to describe the nonlinear variations of the heat losses at the boundaries due to natural convection. Consequently, the authors used one differential equation on a variable representing the mean temperature, and the heat losses coefficient α was a nonlinear function of this characteristic temperature taking into account the effect of natural convection.

Here we want to describe a complex bifurcation behavior, and a single equation is not able to reproduce, for example, the observed limit cycles. We thus consider the

following model problem of two ordinary differential equations:

$$\mathrm{d}_t\bar{\theta} = F(\bar{\theta}) - \alpha\big(|\psi_{\max}|\big)\bar{\theta}, \tag{5.1}$$

$$\mathrm{d}_t\psi_{\max} = -a\psi_{\max}(\psi_{\max}^2 + \delta^2(\theta_c - \bar{\theta})), \tag{5.2}$$

where $\bar{\theta}$ and $\psi_{\max}$ correspond to the mean value of the temperature and the maximum value of the stream function, and a, δ, and θ_c are positive constants. It has to be noticed that α, a, δ, and θ_c also depend on the parameter Ra_0.

Equation (5.1) describes the heat balance between heat production due to the reaction and heat loss through the boundaries. It is almost the same as in Semenov's theory of thermal explosion with the difference that the coefficient α is not constant; $\alpha(|\psi_{\max}|)$ is now a function of the maximal stream function and depends on the Rayleigh number as a parameter.

Equation (5.2) describes a supercritical bifurcation of convective solutions. If $\bar{\theta} < \theta_c(\mathrm{Ra}_0)$, then there is only one solution $\psi_{\max} = 0$. If $\bar{\theta} > \theta_c$, then there are also two other solutions $\psi_{\max} = \pm\delta\sqrt{\bar{\theta} - \theta_c}$.

In the Rayleigh–Bénard problem, the Rayleigh number reads $\mathrm{Ra} = \dfrac{gL^3\alpha_0\Delta T}{\kappa_0\nu_0}$, where the static temperature profile is a linear function of z leading to a temperature difference of ΔT, where α_0 is the thermal expansion coefficient ($1/T_0$ in the present case of a perfect gas). In the present study, the static temperature profile is highly nonlinear as already observed in the analytical study of the critical Rayleigh number (section 4). We then consider an effective Rayleigh number based on the averaged temperature $\mathrm{Ra} = \dfrac{gL^3\bar{\theta}\mathcal{R}T_0/E}{\kappa_0\nu_0} = \mathrm{Ra}_0\bar{\theta}$, where $\mathrm{Ra}_0 = \dfrac{gL^3\mathcal{R}T_0/E}{\kappa_0\nu_0}$; besides, from subsection 3.2, we know that for a stationary convective solution, $\psi_{\max} = b\sqrt{\mathrm{Ra} - \mathrm{Ra}_c} = b\sqrt{\mathrm{Ra}_0}\sqrt{\bar{\theta} - \bar{\theta}_c}$, with $\bar{\theta}_c = \mathrm{Ra}_c/\mathrm{Ra}_0$, so that $\delta = b\sqrt{\mathrm{Ra}_0}$. We then consider $\tilde{\psi} = \psi_{\max}/\delta$ with $\tilde{a} = a\delta^2$. In what follows we omit the tildes for the equation on $\tilde{\psi}$ and the bars for the equation on $\bar{\theta}$ for the sake of simplicity and the system reads as follows:

$$\mathrm{d}_t\theta = F(\theta) - \alpha\big(|\psi|\big)\theta, \tag{5.3}$$

$$\mathrm{d}_t\psi = -a\psi(\psi^2 + \theta_c - \theta). \tag{5.4}$$

5.1. Various modeling of the heat losses α. In this paper, we consider two models for the description of the heat losses coefficient α. The first one is a very simplified model:

$$\alpha(|\psi|) = \alpha_0(1 + \mu|\psi|^2), \tag{5.5}$$

where μ is the sensitivity of the heat losses coefficient α on convection; it can depend on the parameter Ra_0. For this model, it is possible to define analytically the critical conditions in terms of μ, θ_c, and α_0 (see subsection 5.3). However, even if this model reproduces the existence of stable limit cycles and Hopf bifurcations, it occurs in such a narrow parameter range that it makes the comparison with the original partial differential equation model difficult. This is the reason why we introduce a second model; it is based on the study [18] where a formula is provided in order to approximate the heat losses coefficient in the configuration of high Rayleigh numbers Ra and for a nonreactive problem:

$$\alpha = \alpha_0\left(1 + \frac{\mu\mathrm{Ra}^n}{\mathrm{Ra} + \gamma}\right). \tag{5.6}$$

In the present situation, when Ra_0 is reaching Ra_c, where the supercritical bifurcation is taking place, the modification of α_0 should approach zero so that Ra, in (5.6), has to be replaced by $\mathrm{Ra}_0\theta - \mathrm{Ra}_c$, which can be rewritten for stationary convective solutions: $\mathrm{Ra}_0(\theta - \theta_c) = \mathrm{Ra}_0\psi^2$. Finally we consider the model:

$$\alpha(|\psi|) = \alpha_0\left(1 + \frac{\mu\,\mathrm{Ra}_0^{n-1}|\psi|^{2n}}{|\psi|^2 + \gamma/\mathrm{Ra}_0}\right); \tag{5.7}$$

we will consider essentially $n = 1$ and assume that the sensitivity of the heat losses coefficient to the convection μ depends on the parameter Ra_0.

For this second model, we will show that the parameter range for the existence of a stable limit cycle starting at the Hopf bifurcation is larger; besides, this model is able to reproduce the existence of a stable limit cycle between two Hopf bifurcations, the oscillations amplitude of which are first increasing and then decreasing without leading to explosion (sections 5.4 and 6).

5.2. Conditions for Hopf bifurcation. Let us now study the stability of a stationary solution (θ_S, ψ_S). We will show that the presence of oscillations (as observed in section 4) can be described by the model problem (5.3)–(5.4) under some assumptions. Consider the linearized system

$$\mathrm{d}_t\theta = F'(\theta_S)\theta - \alpha(|\psi_S|)\theta - \alpha'(|\psi_S|)\theta_S\psi, \tag{5.8}$$

$$\mathrm{d}_t\psi = a\psi_S\theta - 2a\psi_S^2\psi. \tag{5.9}$$

It has purely imaginary eigenvalues $\pm i\phi$ if

$$F'(\theta_S) - \alpha(|\psi_S|) - 2a\psi_S^2 = 0, \tag{5.10}$$

$$-2a\psi_S^2(F'(\theta_S) - \alpha(|\psi_S|)) + a\psi_S\alpha'(|\psi_S|)\theta_S = \phi^2. \tag{5.11}$$

Consider first equation (5.10). It cannot have solutions if $\theta_S < \theta^*$, where θ^* is the critical temperature for the Semenov's theory of thermal explosion given by the equalities

$$F(\theta^*) = \alpha^*\theta^*, \qquad F'(\theta^*) = \alpha^*.$$

Indeed, in this case

$$F'(\theta_S) - \alpha(|\psi_S|) < F'(\theta_S) - \alpha^* < 0.$$

If $\theta_S > \theta^*$, then $F'(\theta_S) > \alpha^*$ and (5.10) may be satisfied. We have

$$F'(\theta_S) - \frac{F(\theta_S)}{\theta_S} - 2a(\theta_S - \theta_c) = 0, \tag{5.12}$$

$$-4a^2\psi_S^4 + a\psi_S\alpha'(|\psi_S|)\theta_S = \phi^2. \tag{5.13}$$

Equation (5.12) and the inequality

$$\theta_S > \frac{4a\psi_S^3}{\alpha'(|\psi_S|)}$$

obtained from (5.13) determine conditions when the system (5.8), (5.9) has purely imaginary eigenvalues. These conditions can be satisfied on decreasing branches of the maximal value of the stationary stream function $\psi(\mathrm{Ra}_0)$ due to the fact that θ_c is inversely proportional to Ra_0.

5.3. Analysis of the first model. Let us first define the critical conditions as well as the conditions for Hopf bifurcation in the case of $\alpha(|\psi|) = \alpha_0(1 + \mu|\psi|^2)$. Concerning the critical conditions, θ_{LP} (LP for "limit point" or critical point) satisfies

$$(\theta_{LP} - \theta_c)^2 + (\theta_{LP} - \theta_c)(\beta + \sigma) + \beta\sigma - 1 = 0, \quad \beta = \theta_c - 1, \quad \sigma = 1/\mu - 1, \tag{5.14}$$

which always has real solutions with only one above θ_c,

$$\theta_{LP} - \theta_c = -\frac{\beta + \sigma}{2} + \sqrt{1 + \left(\frac{\beta - \sigma}{2}\right)^2}. \tag{5.15}$$

Conditions for Hopf bifurcation then read as

$$\frac{\exp(\theta_H)}{\theta_H} = \alpha_0(1 + \mu(\theta_H - \theta_c)) = 2a\frac{\theta_H\theta_c}{\theta_H - 1}, \tag{5.16}$$

$$-4a^2\psi_H^4 + 2a\psi_H^2\alpha_0\mu\theta_H = \phi^2, \qquad \theta_H > \frac{2a\psi_H^2}{\alpha_0\mu}, \tag{5.17}$$

where the H subscript corresponds to the point of Hopf bifurcation. The last inequality reduces to $\theta_H < \frac{2a\theta_c}{2a-\alpha_0\mu}$ if $2a - \alpha_0\mu > 0$; it is always true if $2a - \alpha_0\mu \leq 0$. For the Hopf bifurcation to take place, it is then necessary that

$$\left(-\frac{2a}{\alpha_0\mu} + \beta + \frac{1}{\mu}\right)^2 \geq \frac{4\beta}{\mu}. \tag{5.18}$$

The analysis of the trajectories of system (5.3), (5.4) using CONTENT [16] shows that there exist three qualitatively different situations depending on the choice of parameters: decaying oscillations where a trajectory converges to a stable focus (Case A of Figure 6), oscillations with increasing amplitude from an unstable stationary point and an oscillating explosion (Case C of Figure 6), and, in between these two, slowly decreasing oscillations from an unstable limit cycle decaying to a stable focus or slowly increasing oscillations from this unstable limit cycle yielding to explosion (Case B of Figure 6).

Decaying and growing oscillations were also observed for the complete problem (section 3). Stable periodic oscillations apparently observed for it are not found directly for the model problem by simply changing the parameters and following the trajectories. This is due to the fact that the parameter range corresponding to the existence of a stable limit cycle is extremely small even if we know from the Hopf theorem that it is not reduced to an interval of zero length.

We then conducted a more detailed analysis using limit cycle continuation in order to detect the point where an unstable limit cycle is merging with a stable one. The complete bifurcation diagram is given on Figure 6, where the corresponding phase portraits are provided at the bottom.

The main problem with this model is the very narrow parameter range where a stable limit cycle is to be found. It is all the more difficult when α_0 is becoming smaller. The idea was then to switch to another model where the parameter sensitivity is a little lower due to saturation phenomena: model 2.

5.4. Analysis of the second model. The second model is introduced because it brings three new features. The parameter range where stable limit cycles exist is larger. It is also due to the fact that the stable limit cycle reaches the loop of

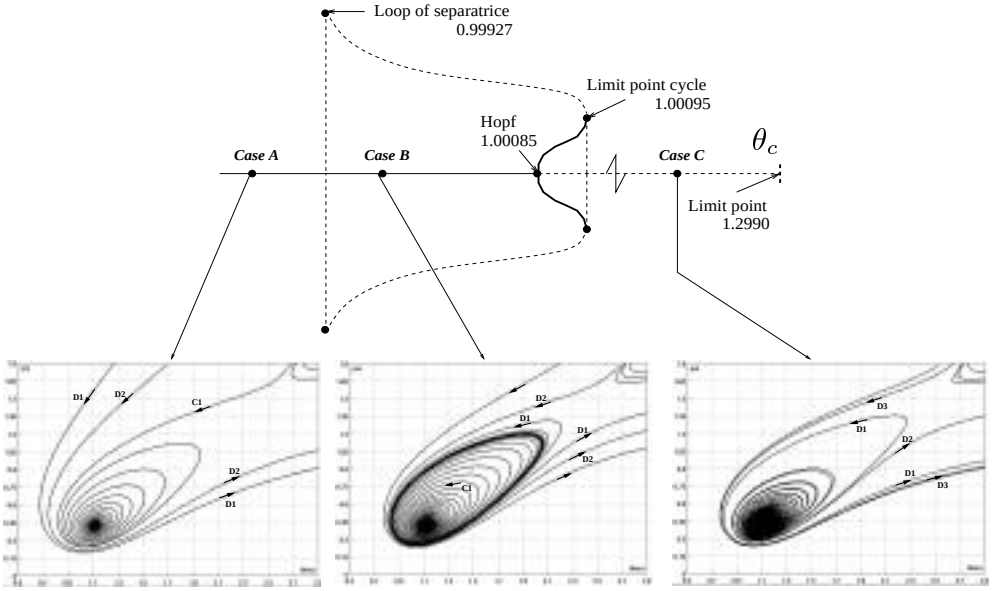


FIG. 6. *Bifurcation diagram for the first model with θ_c as the bifurcation parameter with the various phase portraits (C for converging trajectories and D for diverging trajectories).*

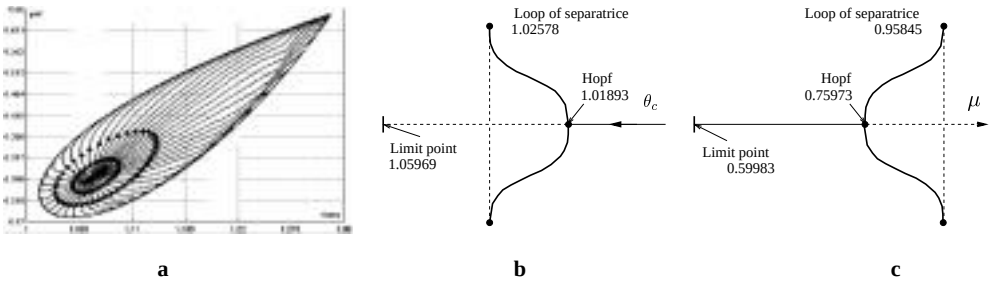


FIG. 7. (a) *Continuation of the limit cycle with θ_c as the continuation parameter in the phase plane (θ, ψ) from the Hopf bifurcation point to the loop of separatrice.* (b) *Bifurcation diagram for the second model with θ_c as the bifurcation parameter.* (c) *Bifurcation diagram for the second model with μ as the bifurcation parameter.*

separatrice associated with the other (saddle) equilibrium instead of meeting with an unstable limit cycle as in the first model (Figure 6). Finally, two Hopf bifurcations are found in the neighborhood of the initial point for both bifurcation parameters θ_c and α_0. The initial point $\alpha_0 = 2.7, \theta_c = 1.0$ is the point we start the continuation of equilibrium from.

We find the two critical points where the equilibria disappear at $\alpha_0 = 2.7, \theta_c = 1.06$ (denoted by "limit point" in Figure 7(b)) and $\alpha_0 = 2.688, \theta_c = 1.0$ (denoted by "limit point" in Figure 7(c)). Besides, two Hopf bifurcations are identified at $\alpha_0 = 2.695, \theta_c = 1$ and at $\alpha_0 = 2.7, \theta_c = 1.019$; at the initial point, the equilibrium is stable and is destabilized through the Hopf bifurcations.

On Figure 7(a), we have used CONTENT [16] in order to represent the continuation of the stable limit cycle with θ_c as the bifurcation parameter. The various lines starting from the Hopf bifurcation point and reaching the loop of separatrice

represent the continuation of one point of the limit cycle. The bifurcation diagram is presented on Figure 7(b), and we see that the stable limit cycle is present from the value $\theta_c = 1.01893$, where the Hopf bifurcation is taking place until $\theta_c = 1.02578$, where the stable limit cycle is merging with the loop of separatrice. Another possible bifurcation parameter which brings a reparametrization of $\dot{\psi} = 0$ is μ, the sensitivity of the heat losses coefficient on convection, instead of α_0. The corresponding bifurcation diagram is presented on Figure 7(c), knowing that the initial point corresponds to $\mu = 0.7$.

5.5. Comparison with direct numerical simulations. The purpose of the present section is to make the link between the direct numerical simulations of section 3, on the one side, and the results on the model problem presented in the first part of this section, on the other side. The fundamental point is to be able to use physical values of the parameters in the model problem. The first step is then to come back to the relationship between the Semenov and Frank-Kamenetskii theories and extend the relationship to the other parameters when convection is present.

The link between the Frank-Kamenetskii and the Semenov theories has already been considered in [24], thus showing that around the critical conditions for explosion, if $\bar{\theta}$ is the averaged temperature in the layer, $\overline{\exp(\theta)} \approx \exp(\bar{\theta})$. Thus, integrating (2.1) in $\bar{z}$ over $[0, 2]$ and changing the temperature scaling from the diffusion time L^2/κ_0 to the chemical time τ_{ch} introduced in subsection 2.2 yields

$$d_{\bar{\tau}}\bar{\theta} = -\Phi + \exp(\bar{\theta}), \qquad \bar{\tau} = t/\tau_{ch} = \mathrm{F_K}.\tau, \tag{5.19}$$

where Φ is the total heat flux through the boundaries, modeled by $\Phi = \alpha_0\bar{\theta}$. However, the critical value of the mean temperature given by the Frank-Kamenetskii theory and denoted $\bar{\theta}_0^*$ is .86, whereas critical conditions for (5.19) yield $\alpha_0^* = e$, $\bar{\theta}_1^* = 1$. Thus, even if $\exp(\bar{\theta})$ is a good approximation of the value in the mean of the chemical source term, it does not provide a good approximation of its derivative as a function of the mean temperature. Consequently, we consider $\eta = \bar{\theta}/\bar{\theta}_0^*$ and approximate $\exp(\bar{\theta})$ by $\bar{\theta}_0^* \exp(\bar{\theta}/\bar{\theta}_0^*)$, so that η satisfies

$$d_{\bar{\tau}}\eta = -\alpha_0\eta + \exp(\eta), \tag{5.20}$$

the critical conditions of which are defined by $\eta^* = 1$, $\alpha_0^* = e$, or $\bar{\theta} = \bar{\theta}_0^* = 0.86$. The change of variable then allows us, by changing the derivative of the chemical source term at critical conditions, to recover critical conditions compatible with the Frank-Kamenetskii theory.

We then have to check that the value of the heat flux at the boundaries is correctly predicted by our numerical model. We have computed the values of the heat flux for various values of the Frank-Kamenetskii number below the critical value, for two spatial discretizations 21 and 51 points, as well as the values predicted by the theory. Numerical results and theory match very well and the error remains below 4%. One can check from these numerical simulations of (2.1) that $\alpha_0 = \partial_{\bar{z}}\bar{\theta}|_{\bar{z}=0} * \bar{\theta}_0^*/\bar{\theta}$ approaches the value e in the neighborhood of the critical conditions.

It was proposed in subsection 3.2 that the convective instability is appearing as a supercritical bifurcation and that the maximum of the stream function is proportional to $\sqrt{\mathrm{Ra} - \mathrm{Ra}_c} = \sqrt{\mathrm{Ra}_0}\sqrt{\theta - \theta_c}$, a property used in the model problem. On Figure 8(a), we check this relationship for $\mathrm{F_K} = 0.6$, $\mathrm{F_K} = 0.875$, and $\mathrm{F_K} = 0.9$; it thus makes the link between the stream function observed with either model and the stream function given by the direct numerical simulations.

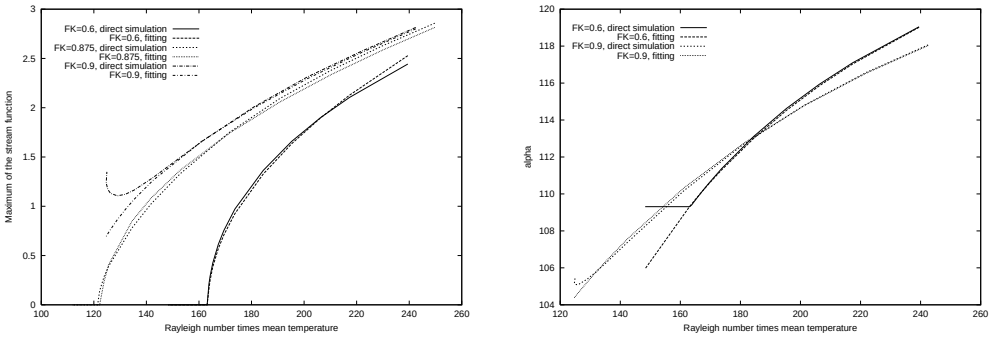


FIG. 8. *Comparison of model and simulations:* (a) $\psi_{\max}$ *at* $\mathrm{F_K} = 0.6, 0.875, 0.9$*;* (b) *heat losses coefficient at* $\mathrm{F_K} = 0.6, 0.9$.

In order to justify the choice of the various models considered for the heat losses coefficients, we have plotted α versus $\mathrm{Ra}_0\theta$ on Figure 8(b), comparing on the one side the ratios of the heat flux over the mean temperature from the simulation and on the other side the values given by the second model.

We have done that for various values of $\mathrm{F_K}$; we isolate two cases on Figure 8(b) for $\mathrm{F_K} = 0.6$, below the critical value, and $\mathrm{F_K} = 0.9$, above the critical value. For the first one, starting from the constant analytical value α_0 which corresponds to the static solution, we can fit very precisely, using the second model, the values coming from the direct numerical simulations and then prove that the choice of a constant value of μ and γ is possible. For the second one, there exists a region of the parameter Ra_0 where the maximum of the stream function has a square root-like behavior; in this region, we can still fit α very precisely as shown on Figure 8(b). However, in the region where the maximum of the stream function separates from the square root-like behavior, the fit for α is not valid anymore and the parameters μ and γ cannot remain fixed.

6. Discussion. In the present section, we discuss three key points: the definition of thermal explosion, the influence of convection on heat losses at the boundaries, and finally the ability of the model problem described in the previous section to reproduce the complete bifurcation diagram of the full partial differential equation system with Ra_0 as the bifurcation parameter at fixed $\mathrm{F_K}$.

Classically, thermal explosion is defined as the nonexistence of a stationary solution for (2.1). When gravity is present and convection interferes with heat production, our study makes it clear that this definition of thermal explosion is not complete. Actually, there is a parameter range where there exists an unstable stationary solution and where thermal explosion can still occur whatever the initial data ($F_K = 0.8775$, which is below the critical value, and for the Rayleigh range between the two Hopf bifurcations yielding oscillatory solutions; see Figure 3), whereas for some parameter range, even if the stationary solution exists and is not stable, there exists periodic oscillations and no thermal explosion is found in the attraction basin of the stable limit cycle; however, the possible domain of initial conditions in order to converge to these stable oscillations is limited. Due to the presence of this oscillatory instability, we have to give a more complete definition: thermal explosion is the blow-up of the temperature; it corresponds to three scenarios: there exists no stationary solution

and no periodic solution; there exists a stable stationary solution or a stable periodic solution with its attraction set of initial data (which can be bounded or not) and the initial solution does not belong to this set; or, finally, there exists unstable stationary solutions and the temperature blows up whatever the initial data.

The second point we want to come back to is the influence of convection on heat losses. The convective instability due to the nonlinear profile of temperature is difficult to relate to the linear one in the Rayleigh–Bénard configuration as shown in section 4, where the critical Rayleigh numbers cannot be correlated. The modeling of the heat losses in the context of thermal explosion coupled to convection was still an open problem. It then becomes clear that using the direct numerical simulations and linear stability analysis, we were able, first, to characterize the critical conditions for the convective instability to appear, and, second, to model the heat losses through an interpolation formula that was to be really precise (using μ and γ) on the square root-like branches (either for $\mathrm{F_K}$ below the critical value or for large Rayleigh numbers). There are two points still to be discussed here: first, what is the influence of the transition from one vortex to two vortices on the heat losses? and, more generally, depending on the aspect ratio, what would be the behavior of heat losses depending on the modes?; second, what is the behavior of the heat losses in the regions of parameters when we are close to thermal explosion and the stationary branch for ψ goes away from a square root-like profile? The latter point will be discussed in the next paragraph: the model problem is able to reproduce the bifurcation behavior of the full system. Finally, we would like to emphasize that we have chosen Dirichlet boundary conditions on part of the boundary, thus influencing the heat losses. It would be necessary to reproduce the same kind of study in the case of Robin boundary conditions.

The last point we want to discuss is the ability of the model problem to describe surprisingly well the bifurcation diagram of the full system of partial differential equations. We have already seen that the heat losses coefficient is well described by the second model proposed in comparison with the direct numerical simulations; the correspondence between the two systems, when the Rayleigh number is small enough, already has been shown by making the link between Semenov and Frank-Kamenetskii theories. However, the bifurcation parameters used in the direct numerical simulation were the Rayleigh number Ra_0 and the Frank-Kamenetskii parameter $\mathrm{F_K}$. The second model system, when $\mathrm{F_K}$ is given, thus setting α_0, Ra_c, and γ, still has two parameters depending on Ra_0: θ_c and μ. We have seen that there can be two causes of Hopf bifurcation: either when θ_c is increasing on a decreasing branch of ψ or when μ is decreasing. It has been shown that μ remains constant on a square root-like branch of the stream function; however, we did not inquire as to what is the behavior of μ when the stream function leaves this square root-like behavior. It can be shown for various values of $\mathrm{F_K} > \mathrm{F_K}_c$ that in the region where the stationary value of ψ as a function of Ra_0 changes its convexity, there is a strong variation of α as a function of ψ. Let us consider, for example, the case when $\mathrm{F_K} = 0.9$, where no oscillations are found. It is shown that for the same value of ψ, the value of α decreases for a smaller Ra_0. We can model this phenomenon by choosing a fixed γ and considering the evolution of μ versus Ra_0. The evolution of the heat losses coefficient sensitivity μ as a function of Ra_0 presents a strong decrease of μ in the considered region. A similar result holds for oscillating solutions.

The very striking fact is that this analysis provides a way of understanding why oscillations can start growing and then decrease onto a stable stationary solution before exploding, as is the case when $\mathrm{F_K} = 0.88$. Let us refer to the bifurcation diagram, Figure 7. It shows that an increase of θ_c (which can be associated with a

decrease of Ra_0) at given μ can initiate the oscillations through a Hopf bifurcation. If, in between the bifurcation point and the loop of separatrice, one encounters a strong decrease of μ, then the amplitude of the limit cycle can decrease until it goes through another Hopf bifurcation so that the stationary solution is stable again. Decreasing μ further leads to an explosion from the stable stationary solution, which is exactly the behavior of the full system for $\mathrm{F_K} = 0.88$. It is interesting to note finally that the model system is able to reproduce all the bifurcation diagram identified using direct numerical simulations.

7. Conclusion. In this paper we have presented a comprehensive simulation, bifurcation analysis, and modeling of the nonlinear interaction of thermal explosion and natural convection in a 2D square configuration. New stable periodic-in-time solutions and oscillating thermal explosion have been identified in the neighborhood of the critical conditions for both explosion and convection. The stable periodic solutions can be modeled by a simple system of two ordinary differential equations. We have justified the modeling of the heat losses depending on the regimes of convection and especially in the neighborhood of the oscillating solutions, where the sensitivity of heat losses to convection encounters a strong decrease. Finally, all the identified regimes and bifurcations can be qualitatively described by the proposed model problem.

REFERENCES

[1] C. Barillon, G. Makhviladze, and V. Volpert, *Heat explosion in a two-phase medium*, Proceedings of the International Conference on Multiphase Flows, Lyon, France, 1998, CD-ROM, paper 691, pp. 1–8.

[2] J. Bebernes and D. Eberly, *Mathematical Problems from Combustion Theory*, Appl. Math. Sci. 83, Springer-Verlag, New York, 1989.

[3] S. Chandrasekhar, *Hydrodynamic and Hydromagnetic Stability*, Clarendon Press, Oxford, 1961.

[4] P.G. Drazin and W.H. Reid, *Hydrodynamic Stability*, Cambridge University Press, Cambridge, UK, 1981.

[5] A. Ern and V. Giovangigli, *Multicomponent Transport Algorithms*, Lecture Notes in Phys. New Series m: Monographs 24, Springer-Verlag, Berlin, 1994.

[6] S. Fauve, *Pattern forming instabilities*, in Hydrodynamics and Nonlinear Instabilities, C. Godrèche and P. Mannevile, eds., Cambridge University Press, Cambridge, UK, 1998, pp. 387–492.

[7] D.A. Frank-Kamenetski, *Temperature distribution in a reactive vessel and stationary theory of heat explosion*, Dokl. Akad. Nauk. SSSR, 18 (1938), pp. 411–412.

[8] D.A. Frank-Kamenetskii, *Diffusion and Heat Transfer in Chemical Kinetics*, Plenum Press, New York, 1969.

[9] S. Génieys and M. Massot, *From Navier-Stokes equations to the Oberbeck-Boussinesq approximation: A unified approach*, European J. Mech. B Fluids, submitted; also available online from http://maply.univ-lyon1.fr/publis/publiv/2001/331/publi.ps.gz.

[10] G.Z. Gershuni and E.M. Zhukovitskii, *Convection Stability of Incompressible Fluids*, Halsted, New York, 1976.

[11] A.V. Getling, *Rayleigh-Bénard Convection, Structure and Dynamics*, Adv. Ser. Nonlinear Dynam. 11, World Scientific, Singapore, 1998.

[12] V. Giovangigli and M. Massot, *Asymptotic stability of equilibrium states for multicomponent reactive flows*, Math. Models Methods Appl. Sci., 8 (1998), pp. 251–297.

[13] V. Giovangigli, *Multicomponent Flow Modeling*, Modeling and Simulation in Science, Engineering, and Technology, Birkhäuser Boston, Boston, MA, 1999.

[14] D.D. Gray and A. Giorgini, *The validity of the Boussinesq approximation for liquids and gases*, Int. J. Heat Mass Transfer, 19 (1976), pp. 545–551.

[15] L. Kagan, H. Berestycki, G. Joulin, and G. Sivashinsky, *The effect of stirring on the limits of thermal explosion*, Combust. Theory Model., 1 (1997), pp. 97–111.

[16] Yu. A. Kuznetsov and V.V. Levitin, *CONTENT: A multiplatform environment for analyzing dynamical systems*. Dynamical Systems Laboratory, Centrum voor Wiskunde en Informat-

ica, Amsterdam; http://ftp.cwi.nl, 1996; also available online from http://ftp.cwi.nl.
[17] A.G. Merzhanov, V.V. Barzykin, and V.G. Abramov, *The theory of heat explosion: From N.N. Semenov to our days*, Himicheskaya Fizika, 15 (1996), pp. 3–44.
[18] A.G. Merzhanov and E.A. Shtessel, *Free convection and thermal explosion in reactive systems*, Astronautica Acta, 18 (1973), pp. 191–199.
[19] J.M. Mihaljan, *A rigorous exposition of the Boussinesq approximations applicable to a thin layer of fluid*, Astrophys. J., 136 (1962), pp. 1126–1133.
[20] K.R. Rajagopal, M. Ruzicka, and A.R. Srinivasa, *On the Oberbeck-Boussinesq approximation*, Math. Models Methods Appl. Sci, 6 (1996), pp. 1157–1167.
[21] N.N. Semenov, *To the theory of combustion processes*, Zhurnal Fizicheskoi Himii, 4 (1933), pp. 4–17.
[22] N.N. Semenov, *Thermal theory of combustion and explosions*, Uspekhi Fiz. Nauk, 23 (1940), p. 251.
[23] E.A. Spiegel and G. Veronis, *On the Boussinesq approximation for a compressible fluid*, Astrophys. J., 131 (1960), pp. 442–447.
[24] Ya. B. Zeldovich, G.I. Barenblatt, V.B. Librovich, and G. M. Makhviladze, *The Mathematical Theory of Combustion and Explosion*, Consultants Bureau, New York, 1985.

SIAM J. APPL. MATH.
Vol. 63, No. 2, pp. 373–397

A NEW RENORMALIZATION METHOD FOR THE ASYMPTOTIC SOLUTION OF WEAKLY NONLINEAR VECTOR SYSTEMS*

B. MUDAVANHU†‡ AND R. E. O'MALLEY, JR.†

Abstract. This paper considers the asymptotic integration of a special class of initial value problems involving a nonlinear regular perturbation scaled by a small parameter $\epsilon > 0$. For $t = \mathcal{O}(1/\epsilon)$, these problems were classically solved using either the method of averaging or of multiple scales to remove secular terms that arise in the natural power series procedure. Our new ansatz is straightforward and effective. Moreover, it indicates when problems might occur in providing the asymptotic solution on very long time intervals. Other closely related problems are also attacked using renormalization.

Key words. oscillations, asymptotics, renormalization

AMS subject classifications. 34C15, 34E10

PII. S0036139901394311

Background: The rise of secular terms. We shall seek the asymptotic solution $x_\epsilon(t)$ of the initial value problem for the weakly nonlinear nearly autonomous vector system

$$\dot{x} = Mx + \epsilon N(x, t, \epsilon) \tag{1}$$

on the semi-infinite time interval $t \geq 0$ as the small positive parameter ϵ tends to zero. Such problems and their generalizations describe numerous electrical, mechanical, and biological oscillations. Indeed, the asymptotic solution of related boundary value problems for partial differential equations remains of substantial interest and importance. Without further hypotheses, however, one can't predict the time interval on which the solution remains bounded. We shall assume that the matrix M has only imaginary eigenvalues, that the fundamental matrix e^{Mt} for the unperturbed problem has a period $p > 0$, and that the vector N is smooth in its three arguments and p-periodic in t. We could even assume that M is a diagonal matrix having a spectral decomposition $M = iV\Lambda V^{-1}$ with a real diagonal matrix Λ and introduce the transformation $\widetilde{x} = Vx$.

By variation of parameters,

$$z_\epsilon(t) = e^{-Mt} x_\epsilon(t) \tag{2}$$

will satisfy the transformed system

$$\dot{z} = \epsilon f(z, t, \epsilon), \tag{3}$$

analogous to (1) with $M = 0$, for the p-periodic forcing

$$f(z, t, \epsilon) \equiv e^{-Mt} N(e^{Mt} z, t, \epsilon). \tag{4}$$

*Received by the editors August 17, 2001; accepted for publication (in revised form) April 1, 2002; published electronically November 19, 2002. This research was supported in part by National Science Foundation grant DMS-0103632.
http://www.siam.org/journals/siap/63-2/39431.html
†Department of Applied Mathematics, University of Washington, Seattle, WA 98195-2420 (omalley@amath.washington.edu).
‡Current address: American International Group, Inc., Market Risk Management, 70 Pine Street, 20th Floor, New York, NY 10270 (blessing.mudavanhu@aig.com).

Indeed, we will say the system (3) is in standard form. Moreover, anticipating that (3) will have a nearly constant solution for bounded times, setting

$$x_\epsilon(t) = e^{Mt}(x(0) + \epsilon u(t, x(0), \epsilon)) \qquad \text{or} \qquad z_\epsilon(t) = x(0) + \epsilon u(t, x(0), \epsilon) \tag{5}$$

shows that the scaled correction vector u will satisfy the nearly linear initial value problem

$$\dot{u} = f(x(0) + \epsilon u, t, \epsilon) \tag{6}$$

on some interval $t \geq 0$ with $u(0, x(0), \epsilon) = 0$.

The natural starting point for obtaining an asymptotic solution $x_\epsilon(t)$ of (1) or $z_\epsilon(t)$ to (3) is to introduce the regular power series expansion

$$u(t, x(0), \epsilon) = u_0(t, x(0)) + \epsilon u_1(t, x(0)) + \epsilon^2 u_2(t, x(0)) + \cdots \tag{7}$$

for u, determining its terms u_j uniquely and successively by equating coefficients of like powers of ϵ in the differential equation (6) and the initial condition. Thus, the u_j's must satisfy the resulting sequence of linear initial value problems

$$\begin{cases} \dot{u}_0 = f(x(0), t, 0), \qquad u_0(0) = 0, \\ \dot{u}_1 = f_x(x(0), t, 0)u_0 + f_\epsilon(x(0), t, 0), \qquad u_1(0) = 0, \\ \dot{u}_2 = f_x(x(0), t, 0)u_1 + \frac{1}{2}[f_{xx}(x(0), t, 0)u_0 + 2f_{x\epsilon}(x(0), t, 0)]u_0 \\ \qquad + \frac{1}{2}f_{\epsilon\epsilon}(x(0), t, 0), \qquad u_2(0) = 0, \\ \text{etc.}, \end{cases} \tag{8}$$

and thus integrating successively immediately provides the coefficients

$$u_0(t, x(0)) = \int_0^t f(x(0), s, 0)ds, \tag{9}$$

$$u_1(t, x(0)) = \int_0^t [f_x(x(0), s, 0)u_0(s, x(0)) + f_\epsilon(x(0), s, 0)]ds, \tag{10}$$

etc., in (7). Using standard Gronwall inequality arguments (cf. Smith (1985) or Murdock (1991)), it becomes clear that the regular power series (7) provides the asymptotic solution $x_\epsilon(t)$ as $\epsilon \to 0$ on bounded t intervals.

Recall, however, that Lagrange, Laplace, Poincaré, and other developers of celestial mechanics knew that ordinary *resonance* implies that these u_j's generally contain secular terms that grow as polynomials in t of degree $j + 1$. This implies that the expansion (7) then loses its asymptotic validity on long time intervals since the terms $\epsilon^{j+1}u_j(t)$ of ϵu all attain the same asymptotic order when $t = \mathcal{O}(1/\epsilon)$. For this reason, the power series (7) was called *naive* by Chen, Goldenfeld, and Oono (1996). Many asymptotic methods have been developed to deal with this dilemma. The most important classical techniques are the Krylov–Bogoliubov *averaging* method, largely developed in Kiev in the 1930s (cf. Bogoliubov and Mitropolsky (1961)), and two-timing or the method of *multiple scales*, developed at Caltech in the 1960s (cf. Kevorkian and Cole (1996), but note independent early contributions of Kuzmak (1959), Cochran (1962), and Mahony (1962)). Our work relates closely to the *renormalization group* method of Chen, Goldenfeld, and Oono (1996) and the *invariance*

condition method of Woodruff (1993, 1995), though averaging and multiple scale concepts remain essential to its development. Readers should note that Oono (2000) and Nozaki and Oono (2001) simplify the earlier renormalization group method and that Jarrad (2001) includes a promising *variational perturbation theory.* Chen, Goldenfeld, and Oono (1996) began their paper by suggesting that "the practice of asymptotic analysis is something of an art." Like them, we seek to show that "the renormalization group approach sometimes seems to be more efficient and accurate than standard methods in extracting global information from the perturbation expansion."

Simple resonance considerations show that u_0 will grow like a multiple of t as $t \to \infty$ if and only if its known forcing $f(x(0), t, 0)$ has a nonzero *average*

$$\langle f(x(0), t, 0)\rangle \equiv \frac{1}{p}\int_0^p f(x(0), s, 0)ds, \tag{11}$$

which conveniently coincides with the leading term in its Fourier series expansion on $0 \le t \le p$. Indeed, if we split $f(x(0), t, 0)$ into its average and supplementary *fluctuating zero-average part*

$$\{f(x(0), t, 0)\} = f(x(0), t, 0) - \langle f(x(0), t, 0)\rangle, \tag{12}$$

the response u_0 analogously splits into the sum

$$u_0(t, x(0)) = ta_0(x(0)) + U_0(x(0), t) \tag{13}$$

of its corresponding secular part $a_0 t$, with the average

$$a_0(x(0)) \equiv \langle f(x(0), t, 0)\rangle \tag{14}$$

as a coefficient and with the *bounded* secular-free part

$$U_0(x(0), t) \equiv \int_0^t \{f(x(0), s, 0)\}ds. \tag{15}$$

Substituting (13) into (10), integrating by parts, and splitting f_x into its average and fluctuating parts, we next get

$$\begin{aligned} u_1(t, x(0)) &= \left(\int_0^t s f_x(x(0), s, 0)ds\right) a_0(x(0)) \\ &\qquad + \int_0^t [f_x(x(0), s, 0)U_0(x(0), s) + f_\epsilon(x(0), s, 0)]ds \\ &= \left(\frac{1}{2}t^2\langle f_x(x(0), t, 0)\rangle + t\frac{\partial U_0}{\partial x}(x(0), t)\right) a_0(x(0)) \\ &\qquad + \int_0^t \left[f_x(x(0), s, 0)U_0(x(0), s) + f_\epsilon(x(0), s, 0)\right. \\ &\qquad\qquad \left. - \frac{\partial U_0}{\partial x}(x(0), s)a_0(x(0))\right] ds. \end{aligned}$$

Thus, u_1 has the predicted polynomial form

$$\begin{aligned} u_1(t, x(0)) &= \frac{1}{2}t^2\langle f_x(x(0), t, 0)\rangle a_0(x(0)) \\ &\qquad + t\left[a_1(x(0)) + \frac{\partial U_0}{\partial x}(x(0), t)a_0(x(0))\right] + U_1(x(0), t), \end{aligned} \tag{16}$$

where the coefficients involve the average

$$a_1(x(0)) \equiv \left\langle f_x(x(0),t,0)U_0(x(0),t) + f_\epsilon(x(0),t,0) - \frac{\partial U_0}{\partial x}(x(0),t)a_0(x(0)) \right\rangle$$

and the supplementary term

$$U_1(x(0),t) \equiv \int_0^t \Big\{ f_x(x(0),s,0)U_0(x(0),s) + f_\epsilon(x(0),s,0) - \frac{\partial U_0}{\partial x}(x(0),s)a_0(x(0)) \Big\} ds \tag{17}$$

that remains bounded for all $t \geq 0$. (Corresponding higher-order first and final coefficients a_j and U_j won't be so directly linked as when $j = 0$ and 1.)

Continuing in this manner, however, we ultimately learn that a bounded asymptotic solution $z_\epsilon(t)$ results on a longer time interval from using only the bounded secular-free (or so-called *bare*) part

$$x(0) + \epsilon U_0(x(0),t) + \epsilon^2 U_1(x(0),t) + \epsilon^3(\cdots)$$

of the regular power series for z_ϵ. Further, we must generally replace the initial vector $x(0)$ by a time-varying amplitude $A_\epsilon(\tau)$ depending on the slow-time

$$\tau = \epsilon t \tag{18}$$

and found by integrating the initial value problem

$$\frac{dA_\epsilon}{d\tau} = a_0(A_\epsilon) + \epsilon a_1(A_\epsilon) + \mathcal{O}(\epsilon^2), \quad A_\epsilon(0) = x(0),$$

on (possibly unbounded) τ intervals where its solution remains bounded. (Observe that one might interpret the replacement of $x(0)$ by the slowly varying $A_\epsilon(\tau)$ as finding an envelope of solutions (cf. Ei, Fujii, and Kunihiro (2000)). Likewise, one could be motivated by Whitham's success in using slowly varying functions to asymptotically solve nonlinear partial differential equations (cf. Whitham (1974) and Debnath (1997)) or by the use of related *amplitude equations* in stability theory (cf. Coullet and Spiegel (1983), Eckhaus (1992), and Promislow (2001)). The basic ploy is to eliminate the secular terms from the naive expansion (7). Moreover, observe that replacing $x(0)$ by $A_\epsilon(\tau)$ also makes our leading-order approximation $e^{Mt}A_\epsilon(\tau)$ to $x_\epsilon(t)$ richer, although such an improvement will not be asymptotically noticeable when t is only finite. We admit that this simple renormalization result still remains largely unmotivated, but we shall now obtain it by using an effective ansatz that could be applied more generally (e.g., in asymptotically stable contexts where M is a stable matrix and $\frac{1}{p}\int_0^p f(x(0),s,0)ds$ converges as $p \to \infty$, allowing us to take an infinite p to again define the averaged equation satisfied by the limiting $A_0(\tau)$. When $M = \left(\begin{smallmatrix} -Q & 0 \\ 0 & iR \end{smallmatrix}\right)$ for an exponentially decaying matrix e^{-Qt} and a periodic e^{iRt}, we would use such a long-time average to approximate the first components of x).

The basic ansatz. We shall begin anew to solve (1) by directly introducing the multiple-scale *ansatz*

$$\begin{cases} x_\epsilon(t) = e^{Mt}z_\epsilon(t) \equiv e^{Mt}[A_\epsilon(\tau) + \epsilon U(A_\epsilon(\tau),t,\epsilon)] \\ \text{or} \\ z_\epsilon(t) = A_\epsilon(\tau) + \epsilon U(A_\epsilon(\tau),t,\epsilon) \end{cases} \tag{19}$$

corresponding to the *bare* expansion of Chen, Goldenfeld, and Oono (1996). It can be motivated for problem (1), at least for τ finite, by substituting (19) into the differential equation (3). Using the chain rule, we get

$$\frac{1}{\epsilon}\frac{dz_\epsilon}{dt} = \left(I + \epsilon\frac{\partial U}{\partial A_\epsilon}\right)\frac{dA_\epsilon}{d\tau} + \frac{\partial U}{\partial t} = f. \tag{20}$$

We now split $\frac{\partial U}{\partial A_\epsilon}$ and f into sums of their average and mean-free fluctuating parts, respectively, using the leading term and the supplementary sum of their Fourier expansions on $0 \le t \le p$, and asking that A_ϵ account for the nonzero average terms

$$\left(I + \epsilon\left\langle\frac{\partial U}{\partial A_\epsilon}\right\rangle\right)\frac{dA_\epsilon}{d\tau} = \langle f\rangle$$

in (20), while the correction ϵU to A_ϵ in (19) handles its remaining terms

$$\epsilon\left\{\frac{\partial U}{\partial A_\epsilon}\right\}\frac{dA_\epsilon}{d\tau} + \frac{\partial U}{\partial t} = \{f\}$$

with zero averages. (Readers might indeed recall that an analogous decomposition occurs in the early comparison by Morrison (1966) of averaging and two-timing.) Thus, A_ϵ should satisfy the autonomous system

$$\begin{aligned}\frac{dA_\epsilon}{d\tau} &= \left(I + \epsilon\left\langle\frac{\partial U}{\partial A_\epsilon}(A_\epsilon(\tau), t, \epsilon)\right\rangle\right)^{-1}\langle f(A_\epsilon + \epsilon U(A_\epsilon, t, \epsilon), t, \epsilon)\rangle \\ &\equiv a\ (A_\epsilon, \epsilon) \\ &= \langle f(A_\epsilon, t, 0)\rangle \\ &\quad + \epsilon\Bigg[\langle f_x(A_\epsilon, t, 0)U_0(A_\epsilon, t) + f_\epsilon(A_\epsilon, t, 0)\rangle \\ &\qquad - \left\langle\frac{\partial U_0}{\partial A_\epsilon}(A_\epsilon, t)\right\rangle\langle f(A_\epsilon, t, 0)\rangle\Bigg] + \epsilon^2(\cdots)\end{aligned} \tag{21}$$

and the initial condition $A_\epsilon(0) = x(0)$, while U must satisfy

$$\frac{\partial U}{\partial t} = \{f\} - \epsilon\left\{\frac{\partial U}{\partial A_\epsilon}\right\}a(A_\epsilon, \epsilon)$$

and the trivial initial condition $U(A_\epsilon(\tau), 0, \epsilon) = 0$. We shall call the differential equation (21) the *amplitude* or *first level RG (renormalization group) equation*, noting that an analogous evolution equation results when one uses the higher-order method of averaging. The asymptotic integration of (21) on $\tau \ge 0$ is the appropriate candidate problem to replace the integration of (1) after t becomes unbounded. In these differential equations for A_ϵ and U, the times t and τ are taken to be independent variables, as is typical in two-timing. Integrating the latter equation with respect to t shows that U must satisfy the integral equation

$$\begin{aligned}U(A_\epsilon(\tau), t, \epsilon) = &\int_0^t \{f(A_\epsilon(\tau) + \epsilon U(A_\epsilon(\tau), s, \epsilon), s, \epsilon)\}ds \\ &- \epsilon\int_0^t\left\{\frac{\partial U}{\partial A_\epsilon}(A_\epsilon(\tau), s, \epsilon)\right\}ds\, a(A_\epsilon(\tau), \epsilon).\end{aligned} \tag{22}$$

That we have obtained the compact formulae (21) and (22) to all orders in ϵ is quite helpful, though we naturally next employ power series methods to get more explicit asymptotic results for bounded τ values. Note, in particular, that $\frac{\partial U}{\partial t}$ has a zero average, so its integral U in (22) will be bounded whenever the amplitude $A_\epsilon(\tau)$ is. The resulting power series expansion

$$U(A_\epsilon(\tau), t, \epsilon) = U_0(A_\epsilon(\tau), t) + \epsilon U_1(A_\epsilon(\tau), t) + \epsilon^2(\cdots) \tag{23}$$

has coefficients successively and unambiguously given by

$$\begin{aligned} U_0(A_\epsilon(\tau), t) &= \int_0^t \{f(A_\epsilon(\tau), s, 0)\} ds, \\ U_1(A_\epsilon(\tau), t) &= \int_0^t \{f_x(A_\epsilon(\tau), s, 0) U_0(A_\epsilon(\tau), s) + f_\epsilon(A_\epsilon(\tau), s, 0)\} ds \\ &\quad - \int_0^t \left\{ \frac{\partial U_0}{\partial A_\epsilon}(A_\epsilon(\tau), s) \right\} ds\, a_0(A_\epsilon(\tau)), \end{aligned}$$

etc., corresponding to the functions previously obtained in (15) and (17) for the non-secular parts of the naive expansion (7). Note that U is p-periodic in t.

The remaining, and still formidable, task is to obtain the asymptotic solution of the initial value problem (21) for the slowly varying amplitude A_ϵ on time intervals where it will determine a bounded solution x_ϵ or z_ϵ via (19) and (22). We naturally first seek $A_\epsilon(\tau)$ as a power series

$$A_\epsilon(\tau) = A_0(\tau) + \epsilon A_1(\tau) + \epsilon^2 A_2(\tau) + \cdots, \tag{24}$$

where (21) implies that its limit A_0 must satisfy the limiting nonlinear problem

$$\frac{dA_0}{d\tau} = a_0(A_0) \equiv \frac{1}{p} \int_0^p f(A_0, s, 0) ds, \quad A_0(0) = x(0), \tag{25}$$

exactly as in classical averaging, while the next term A_1, for example, must satisfy a linearized problem

$$\frac{dA_1}{d\tau} = \frac{da_0(A_0)}{dA_0} A_1 + a_1(A_0), \quad A_1(0) = 0.$$

The uniqueness of A_0 implies the invertibility of the Jacobian matrix $\frac{\partial A_0}{\partial x(0)}$, which satisfies the homogeneous linear matrix system as long as A_0 remains defined. If A_0 ever blows up, we must naturally limit our interval of approximation. Using a variation of parameters, then,

$$A_1(\tau) = \frac{\partial A_0(\tau)}{\partial x(0)} \int_0^\tau \left(\frac{\partial A_0(s)}{\partial x(0)} \right)^{-1} a_1(A_0(s)) ds \tag{26}$$

and later terms A_j also follow successively without complication. Using the regular perturbation expansions for $A_\epsilon(\tau)$ and for $U(A_\epsilon(\tau), t, \epsilon)$ in the ansatz (19) results in an approximation for x_ϵ that agrees asymptotically with the naive expansion on intervals where t is finite, and that extends that approximation to longer intervals, at least as long as τ remains finite and $A_0(\tau)$ is defined. One possible further difficulty is instability of $A_0(\tau)$ as $\tau \to \infty$ (if $x(0)$ isn't restricted to the appropriate stable

manifold). Another is encountering τ-secular terms in the power series generated for $A_\epsilon(\tau)$. Note, indeed, that A_1 will be τ-secular if the forcing term $a_1(A_0)$ contains a nontrivial projection in the range of the fundamental matrix $\frac{\partial A_0(\tau)}{\partial x(0)}$. A bounded τ, indeed, provides the usual time limit for obtaining asymptotic solutions by the classical averaging and two-timing methods, which are quite intimately related (cf. Morrison (1966)). Instability as $\tau \to \infty$ cannot be overcome. If, however, $A_0(\tau)$ exists for all $\tau \geq 0$ and if it decays exponentially to an asymptotically stable rest point or *sink*, the resulting expansion (24) for $A_\epsilon(\tau)$ and the resulting expansion (19) for $x_\epsilon(t)$ are uniformly valid for all $t \geq 0$. Recall that Greenlee and Snow (1975) provided an early discussion of such problems, while Murdock and Wang (1996) called this the Sanchez–Palencia condition, in reference to related results for averaging. Indeed, when $a_0(A_0) \equiv 0$, we can immediately seek the asymptotic solution $x_\epsilon(t)$ on $\mathcal{O}(1/\epsilon^2)$ time intervals, as Sanders and Verhulst (1985) and Murdock and Wang (1996) show for averaging and multiple scales, respectively, by replacing the slow-time τ in (21) by the even slower-time

$$\kappa = \epsilon\tau = \epsilon^2 t. \tag{27}$$

Readers should realize that the successful ansatz (19) can be interpreted as a *near-identity* transformation for z_ϵ. Such transformations, which generalize a classical asymptotic procedure of von Ziepel, were introduced by Neu (1980). They are useful in a variety of contexts, including many where our periodicity assumption doesn't hold. In this sense, the basic method of matched asymptotic expansions (cf. Il'in (1992)) and the boundary function method of Vasil'eva, Butuzov, and Kalachev (1995) can both be considered to be extensions of our renormalization technique, as will be demonstrated below. Note further that the basic ansatz (19) is considerably more direct than traditional two-timing, since the solution is not sought as an arbitrary function of the slow-time τ, but rather as a function of t and the amplitude A_ϵ, which is obtained asymptotically as a function of τ by solving the renormalized equation (21). At any stage, we have available a finite truncation

$$U^n(A_\epsilon(\tau), t, \epsilon) \equiv \sum_{j=0}^{n} \epsilon^j U_j(A_\epsilon(\tau), t)$$

for the correction U to A_ϵ satisfying $U(A_\epsilon(\tau), t, \epsilon) = U^n(A_\epsilon(\tau), t, \epsilon) + \mathcal{O}(\epsilon^{n+1})$. Likewise, we have the truncation

$$A_\epsilon^n(\tau) = \sum_{j=0}^{n} \epsilon^j A_j(\tau)$$

such that $A_\epsilon(\tau) = A_\epsilon^n(\tau) + \mathcal{O}(\epsilon^{n+1})$. Substituting into the integral (22), this implies that

$$U(A_\epsilon(\tau), t, \epsilon) = U^n(A_\epsilon^m(\tau), t, \epsilon) + \mathcal{O}(\epsilon^{n+1}) + \mathcal{O}(\epsilon^{m+2}t) \tag{28}$$

for any positive integers m and n, at least for appropriate bounded values of τ.

Our success in using the ansatz (19) for large times suggests that we might often be able to asymptotically solve the amplitude equation (21) on long time intervals, even when τ-secular terms in the series (24) for the amplitude A_ϵ need to be eliminated, by using a secondary ansatz

$$x(0) = B_\epsilon(\kappa) + \epsilon W(B_\epsilon(\kappa), \tau, \epsilon), \tag{29}$$

analogous to (19), in (19). We can asymptotically solve the resulting second level RG equation for the amplitude $B_\epsilon(\kappa)$ to get the resulting multiscale composite expansion

$$x_\epsilon(t) = e^{Mt}[\mathcal{A}(\tau, B_\epsilon(\kappa) + \epsilon W(B_\epsilon(\kappa), \tau, \epsilon), \epsilon) + \epsilon U(\mathcal{A}(\tau, B_\epsilon(\kappa) + \epsilon W(B_\epsilon(\kappa), \tau, \epsilon), \epsilon), t, \epsilon)],$$

where we have set

$$A_\epsilon(\tau) = \mathcal{A}(\tau, x(0), \epsilon)$$

to emphasize its dependence on the initial vector $x(0)$. This expansion can be expected to be valid at least for bounded κ intervals. Moreover, we can consider the preceding expansion (19) to be an *intermediate* asymptotic expansion in the sense of Barenblatt (1996).

The critical idea behind the traditional (first level) renormalization group method of Chen, Goldenfeld, and Oono (1996) is to replace the initial value $x(0)$ in the naive expansion (7) by a slowly-varying function $A_\epsilon(\tau)$ through a near-identity transformation

$$x(0) = A_\epsilon(\tau) + \epsilon Z(A_\epsilon(\tau), t, \epsilon) \tag{30}$$

to eliminate secular (or *divergent*) terms in the naive expansion (7) by appropriate selection of the correction terms Z_j and to thereby obtain the secular-free expansion (19), where A_ϵ remains to be determined. To lowest orders, we would, for example, obtain the necessary cancellation by taking

$$Z_0(A_\epsilon, t) = -a_0(A_\epsilon)t$$

and

$$Z_1(A_\epsilon, t) = \frac{1}{2}t^2\langle f_x(A_\epsilon, t, 0)\rangle a_0(A_\epsilon) - t\left[\langle f_x(A_\epsilon, t, 0)U_0(A_\epsilon, t, 0) + f_\epsilon(A_\epsilon, t, 0)\rangle - a_0(A_\epsilon)\left\langle \frac{\partial U_0}{\partial x}(A_\epsilon, t)\right\rangle\right].$$

Upon differentiating (30) with respect to t, the *invariance condition* $\frac{dx(0)}{dt} = 0$ and the chain rule immediately imply that $A_\epsilon(\tau)$ will satisfy

$$\frac{dA_\epsilon}{d\tau} = a(A_\epsilon, \epsilon) \equiv -\left(I + \epsilon\frac{\partial Z}{\partial A_\epsilon}\right)^{-1}\frac{\partial Z}{\partial t}. \tag{31}$$

We did not take this approach above because it is more direct to immediately find A_ϵ by asymptotically integrating (21), which turns out to ultimately be independent of the secular correction Z introduced in (30). We nonetheless note how instructive it is to see how the terms of the t-secular function Z can be selected to eliminate successive secular terms in (7) and to learn how the function $a(A_\epsilon, \epsilon)$, generated by using the intermediate Z, coincides with that already defined in (21). In some sense, renormalization corresponds to a summing of secular terms. We note that Nozaki and Oono (2001) get the RG equation from an intermediate proto-RG equation and that they make a distinction between resonant and nonresonant secular terms. Indeed, when no secular terms occur in (7), A_ϵ will remain the constant $x(0)$. Next, τ-secular terms in the resulting series (24) could analogously also be eliminated, if necessary, by

replacing the initial vector $x(0)$ by a slowly varying function $B_\epsilon(\kappa)$ of the slow time $\kappa = \epsilon^2 t$ through use of another near-identity transformation (29), where $B_\epsilon(\kappa)$ must satisfy a second level RG equation

$$\frac{dB_\epsilon}{d\kappa} = b(B_\epsilon, \epsilon) \equiv -\left(I + \epsilon \frac{\partial W}{\partial B_\epsilon}\right)^{-1} \frac{\partial W}{\partial \tau} \tag{32}$$

and $B_\epsilon(0) = x(0)$ (cf. Mudavanhu and O'Malley (2001)). Assuming existence and appropriate stability of $B_0(\kappa)$, this will allow the asymptotic solution for x_ϵ to be defined beyond bounded values of κ. One may again be stopped by either blowup at finite κ, instability as $\kappa \to \infty$, or by κ-secular terms. The latter would require a higher level renormalization, and that could determine the asymptotic solution on a still longer time interval. We thus proceed in a leapfrog fashion. (Related applied work is contained in Moise and Ziane (2001) and Wirosoetisno, Shepherd, and Temam (2002).)

Two simple scalar examples. (a) Consider the simple example

$$\dot{x} = ix + \epsilon x(\alpha + x) \tag{33}$$

for some bounded complex constant α. Direct integration of this Riccati equation provides the exact solution

$$x_\epsilon(t) = e^{(i+\epsilon\alpha)t}\left[1 - \frac{\epsilon x(0)}{i+\epsilon\alpha}(e^{(i+\epsilon\alpha)t} - 1)\right]^{-1} x(0) \tag{34}$$

with a least period $\frac{2\pi}{1-i\epsilon\alpha}$ when $\operatorname{Re}\alpha = 0$. When $\alpha \neq 0$, secular terms become apparent when $e^{\epsilon\alpha t}$ is expanded in its Maclaurin series about $\epsilon = 0$. When $\operatorname{Re}\alpha < 0$, such a naive expansion in powers of ϵ is very misleading, since the actual solution decays exponentially to zero as $\tau = \epsilon t \to \infty$, while the Taylor-expanded series has unbounded secular terms. When $\operatorname{Re}\alpha > 0$, however, the solution blows up algebraically like $-\frac{i}{\epsilon}$ as $\tau \to \infty$. Thus, we can't expect an asymptotic approximation to the solution to be defined on time intervals on which τ becomes unbounded.

If we directly seek a solution to (33) of the form

$$x_\epsilon(t) = e^{it}(x(0) + \epsilon u(t, x(0), \epsilon)), \tag{35}$$

the scaled correction u must satisfy the nonlinear equation

$$\begin{aligned} \dot{u} &= f(x(0) + \epsilon u, t, 0) \\ &\equiv (\alpha + x(0)e^{it})x(0) + \epsilon(\alpha + 2x(0)e^{it})u + \epsilon^2 e^{it}u^2 \end{aligned} \tag{36}$$

and $u(0, x(0), \epsilon) = 0$. The resulting regular perturbation series is determined termwise through the successive linear initial value problems

$$\begin{aligned} \dot{u}_0 &= (\alpha + x(0)e^{it})x(0), \quad u_0(0) = 0, \\ \dot{u}_1 &= (\alpha + 2x(0)e^{it})u_0, \quad u_1(0) = 0, \\ \dot{u}_2 &= (\alpha + 2x(0)e^{it})u_1 + e^{it}u_0^2, \quad u_2(0) = 0, \end{aligned}$$

etc. Integrating termwise, we obtain the naive expansion

$$
\begin{aligned}
x_\epsilon(t) = e^{it}\Big[&x(0) + \epsilon(\alpha t + ix(0)(1-e^{it}))x(0) \\
&+ \epsilon^2 x(0)\Big[\frac{1}{2}\alpha^2 t^2 + i\alpha x(0)t(1-2e^{it}) \\
&\qquad - x(0)(1-e^{it})(\alpha + x(0)(1-e^{it}))\Big] + \mathcal{O}(\epsilon^3)\Big],
\end{aligned} \tag{37}
$$

valid asymptotically for bounded t values. The anticipated secular terms occur, however, for $\alpha \neq 0$, indicating the breakdown of the approximation (37) when $\tau = \epsilon t \to \infty$.

If we instead seek an asymptotic solution $x_\epsilon(t)$ of (33), using our ansatz

$$
x_\epsilon(t) = e^{it}(A_\epsilon(\tau) + \epsilon U(A_\epsilon(\tau), t, \epsilon)), \tag{38}
$$

the amplitude A_ϵ and the correction U will have to satisfy (21) and (22), respectively. Since $f(A_\epsilon(\tau), t, 0) = (\alpha + A_\epsilon(\tau)e^{it})A_\epsilon(\tau)$, the average

$$
\langle f(A_\epsilon(\tau), t, 0)\rangle = \alpha A_\epsilon(\tau)
$$

is the leading term of its Fourier series, and it is supplemented by

$$
\{f(A_\epsilon(\tau), t, 0)\} = A_\epsilon^2(\tau)e^{it}.
$$

This implies that both

$$
\frac{dA_\epsilon}{d\tau} = \alpha A_\epsilon + \mathcal{O}(\epsilon)
$$

and

$$
U_0(A_\epsilon(\tau), t) = iA_\epsilon^2(\tau)(1-e^{it}),
$$

and thus

$$
A_\epsilon(\tau) = e^{\alpha\tau}x(0) + \mathcal{O}(\epsilon)
$$

is defined on all $\tau \geq 0$, provided $\operatorname{Re}\alpha \leq 0$. Otherwise, the solution $x_\epsilon(t)$ will be bounded only for finite τ.

The next-order corrections to $\frac{dA_0}{d\tau}$ and U_0 involve the expression

$$
\begin{aligned}
&f_x(A_\epsilon, t, 0)U_0(A_\epsilon, t) - \frac{\partial U_0}{\partial A_\epsilon}(A_\epsilon, t)\langle f(A_\epsilon, t, 0)\rangle \\
&\qquad = (-i\alpha A_\epsilon^2 + 2iA_\epsilon^3 e^{it})(1-e^{it}).
\end{aligned}
$$

Since its average part is $-i\alpha A_\epsilon^2$, the $\mathcal{O}(\epsilon)$ improved approximations satisfy the amplitude equation

$$
\frac{dA_\epsilon}{d\tau} = \alpha A_\epsilon - \epsilon i\alpha A_\epsilon^2 + \mathcal{O}(\epsilon^2), \tag{39}
$$

and the corresponding secular-free correction to A_ϵ is given by

$$
U(A_\epsilon(\tau), t, \epsilon) = iA_\epsilon^2(1-e^{it}) - \epsilon iA_\epsilon^2(1-e^{it})(\alpha + A_\epsilon(1-e^{it})) + \mathcal{O}(\epsilon^2). \tag{40}
$$

As expected, the latter coincides with the secular-free terms of the appropriately truncated naive expansion, with the slowly varying amplitude A_ϵ replacing the initial value $x(0)$.

For $\operatorname{Re}\alpha < 0$, $A_0(\tau)$ will decay exponentially to zero, and this allows us to obtain the asymptotic solution (38) for all $\tau \geq 0$. The most interesting case occurs when $\operatorname{Re}\alpha = 0$. Then, the two-term truncation of the amplitude equation (39) is essentially the same as the original Riccati equation (33). To find the solution for $\kappa = \epsilon^2 t = \mathcal{O}(1)$ generally requires another renormalization, so we will then obtain an asymptotic representation of the solution in terms of the three scales, t, τ, and an amplitude B_ϵ that is a function of κ.

A numerical verification of any presumed approximation X can be carried out by employing the technique of Bosley (1996). We define the absolute error

$$E(t,\epsilon) = |x_{\text{exact}} - X|,$$

although we often employ a carefully computed numerical solution in place of the unknown exact solution. If $E = \mathcal{O}(\epsilon^{n+1})$ for a fixed time, the value of $\log(E)$ as a function of $\log\epsilon$ should be linear with slope $n+1$ in the limit $\epsilon \to 0$. The slope is readily determined by using a linear least squares fit. For the example, interesting results are obtained by considering the three separate cases: $\alpha = 1$, $\alpha = -1$, and $\alpha = i$. (Comparable conclusions on longer time intervals naturally follow for the example

$$\dot{x} = i\epsilon x + \epsilon^2 x(\alpha + x),$$

which we treat by immediately introducing $\tau = \epsilon t$ as a replacement for the given time t.)

Figures 1–3 below are comparisons of the exact solutions in these three cases with, respectively, regular perturbation and RG asymptotic approximations for (33), together with their numerical verifications of asymptotic validity using Bosley's technique for $t = 10$.

(b) We next consider the nonautonomous equation

$$\dot{x} = \epsilon N(x,t) \equiv \epsilon(-x^3 - x^2\cos t + \sin t), \tag{41}$$

introduced by Murdock and Wang (1996), together with a positive initial value $x(0)$. Since $x\dot{x} < 0$ for $|x|$ sufficiently large, the solution $x_\epsilon(t)$ remains bounded for all times. Setting

$$x_\epsilon(t) = x(0) + \epsilon u(t, x(0), \epsilon),$$

it follows that u must satisfy the initial value problem

$$\begin{aligned}\dot{u} = (-x^3(0) + x^2(0)\cos t + \sin t) + \epsilon(-3x^2(0) + 2x(0)\cos t)u \\ + \epsilon^2(-3x(0) + \cos t)u^2, \quad u(0) = 0,\end{aligned}$$

for which a naive expansion could be readily generated. Alternatively, the ansatz (or near-identity transformation)

$$x_\epsilon(t) = A_\epsilon(\tau) + \epsilon U(A_\epsilon(\tau), t, \epsilon) \tag{42}$$

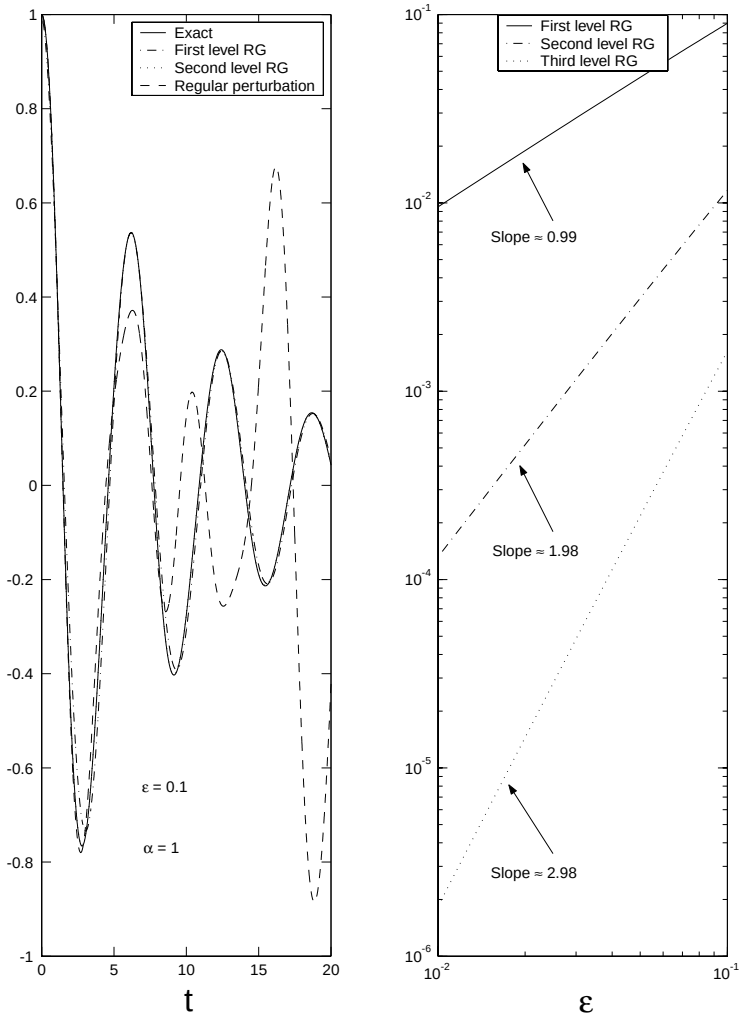


FIG. 1.

involves, to leading order, the amplitude equation

$$\frac{dA_\epsilon}{d\tau} = \langle N(A_\epsilon, t)\rangle + \mathcal{O}(\epsilon) = -A_\epsilon^3 + \mathcal{O}(\epsilon),$$

predicted by averaging, and the secular-free correction term

$$U_0(A_\epsilon, t) = \int_0^t \{N(A_\epsilon, s)\}ds = A_\epsilon^2(\tau)\sin t + 1 - \cos t.$$

Since the resulting limiting amplitude

$$A_0(\tau) = \frac{x(0)}{\sqrt{2\tau x^2(0) + 1}} \tag{43}$$

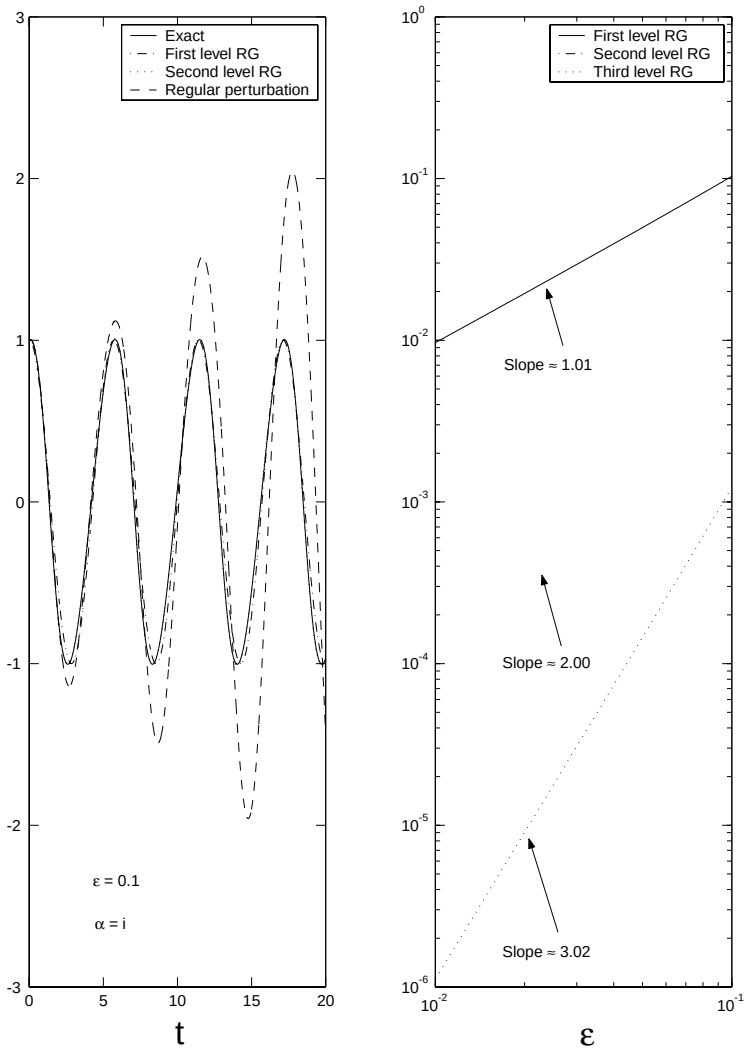


FIG. 2.

decays only algebraically as $\tau \to \infty$, we naturally seek its $\mathcal{O}(\epsilon)$ correction determined by using the average part of the expression

$$N_x(A_0,t)U_0(A_0,t) - \langle N(A_0,t)\rangle U_{0x}(A_0,t) \\ = -(3A_0^2 + A_0) - A_0^4 \sin t + (3A_0^2 + 2A_0)\cos t + A_0^3 \sin 2t - A_0 \cos 2t.$$

Since this implies the more accurate amplitude equation

$$\frac{dA_\epsilon}{d\tau} = -A_\epsilon^3 - \epsilon(3A_\epsilon^2 + A_\epsilon) + \mathcal{O}(\epsilon^2), \tag{44}$$

the regular perturbation series

$$A_\epsilon(\tau) = A_0(\tau) + \epsilon A_1(\tau) + \mathcal{O}(\epsilon^2) \tag{45}$$

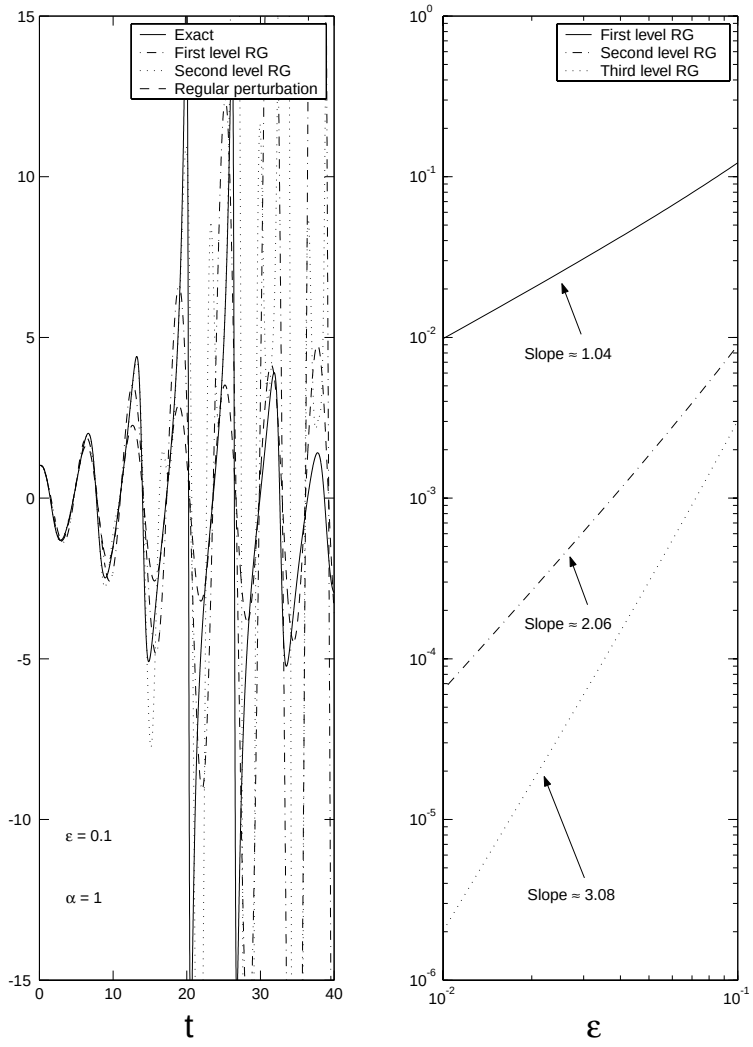


FIG. 3.

will require that the first correction term A_1 satisfy the linear initial value problem

$$\frac{dA_1}{d\tau} = -3A_0^2 A_1 - (3A_0^2 + A_0), \quad A_1(0) = 0.$$

We write its exact solution as

$$A_1(\tau) = -\frac{\sqrt{2\tau x^2(0)+1}}{4x(0)} - 1 + \left(\frac{1}{4x(0)} + 1\right) \frac{1}{(\sqrt{2\tau x^2(0)+1})^3}. \tag{46}$$

Since $|A_1|$ blows up like $\tau^{1/2}$ as $\tau \to \infty$, we shall attempt to eliminate its secular term and later ones in (45) by using a traditional renormalization. Setting

$$x(0) = B_\epsilon(\kappa) + \epsilon W(B_\epsilon(\kappa), \tau, \epsilon) \tag{47}$$

in (45) and using a power series for W, we get the power series expansion

$$A_\epsilon(\tau) = \frac{B_\epsilon + \epsilon W}{\sqrt{2\tau(B_\epsilon+\epsilon W)^2+1}} - \epsilon\left[\frac{\sqrt{2\tau(B_\epsilon+\epsilon W)^2+1}}{4(B_\epsilon+\epsilon W)} - 1 + \cdots\right] + \cdots$$
$$= \frac{B_\epsilon}{\sqrt{2\tau B_\epsilon^2+1}} + \epsilon\left[\frac{W_0}{\sqrt{2\tau B_\epsilon^2+1}} - \frac{2\tau B_\epsilon^2 W_0}{(\sqrt{2\tau B_\epsilon^2+1})^3} - \frac{\sqrt{2\tau B_\epsilon^2+1}}{4B_\epsilon} + \cdots\right] + \cdots.$$

Thus, we can cancel the troublesome τ-secular term at $\mathcal{O}(\epsilon)$ by picking

$$W_0(B_\epsilon,\tau) = \frac{(2\tau B_\epsilon^2+1)^2}{4B_\epsilon}.$$

The resulting second level RG equation (32) is

$$\frac{dB_\epsilon}{d\kappa} = -\frac{\partial W_0}{\partial \kappa} + \mathcal{O}(\epsilon) = -\frac{2}{\epsilon}\kappa B_\epsilon^2 - B_\epsilon + \mathcal{O}(\epsilon).$$

Solving the two-term approximate Riccati equation with the initial value $B_\epsilon(0) = x(0)$ determines the exponentially decaying

$$B_0(\kappa) = \frac{e^{-\kappa}x(0)}{\sqrt{\frac{x^2(0)}{\epsilon}(1-e^{-2\kappa}-2\kappa e^{-2\kappa})+1}} \tag{48}$$

(with admitted abuse of notation) and the corresponding leading-order approximation

$$x_\epsilon(t) = \frac{B_0(\kappa)}{\sqrt{\frac{2\kappa}{\epsilon}B_0^2(\kappa)+1}} + \mathcal{O}(\epsilon) \tag{49}$$

to the decaying solution, which is asymptotically valid for all $t \geq 0$. We note that the regular perturbation expansion is asymptotically correct for t finite, that the series (42) and (45) with τ-secular terms is likewise correct for τ finite, but that the twice-renormalized expansion corresponding to (49) is needed on longer time intervals. The algebraic decay of the limiting solution with $\sqrt{\kappa/\epsilon} = \sqrt{\epsilon t}$ is unexpected, but it follows from renormalization, as does the ultimate exponential decay as $\kappa \to \infty$. Analogous behavior was obtained in Mudavanhu and O'Malley (2001) in solving the second-order equation

$$\ddot{y} + y + \epsilon\dot{y}^3 + 3\epsilon^2\dot{y} = 0,$$

introduced in Morrison (1966).

Figure 4 is a comparison of the numerical solution and the first and second *level* RG asymptotic approximations for the solution of (41). The second level approximation is obtained by renormalizing the second-order amplitude equation as indicated. Figure 5 shows the numerical verifications of the RG approximations using Bosley's technique for $t = 10$.

Second-order scalar equations. Mudavanhu and O'Malley (2001) considered scalar equations of the form

$$\ddot{y} + y + \epsilon g(y,\dot{y},\epsilon) = 0 \tag{50}$$

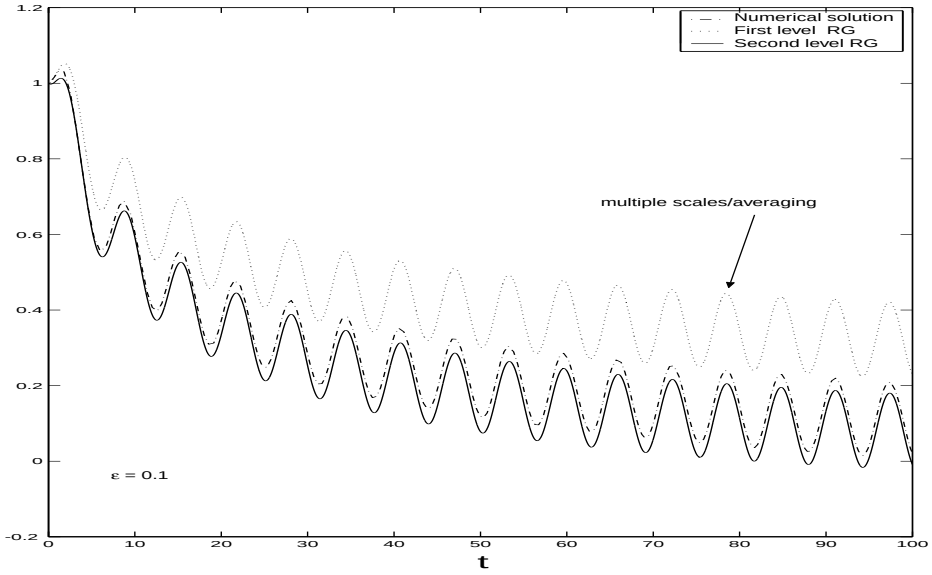


FIG. 4.

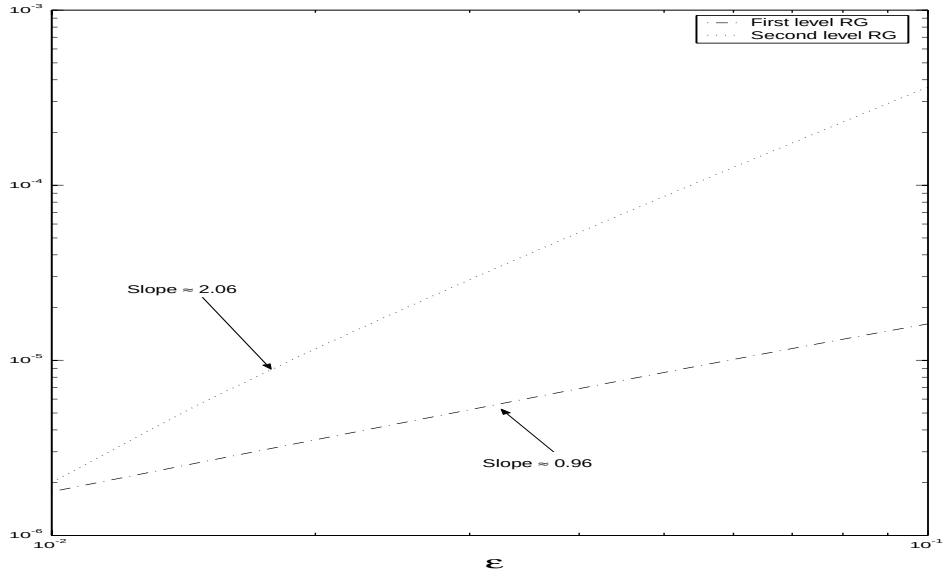


FIG. 5.

on $t \geq 0$, with $y(0)$ and $\dot{y}(0)$ prescribed. Such problems take the form (1) when one introduces

$$x = \begin{pmatrix} y \\ \dot{y} \end{pmatrix}, \qquad M = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad \text{and } N(x,\epsilon) = \begin{pmatrix} 0 \\ -g(y,\dot{y},\epsilon) \end{pmatrix}, \tag{51}$$

and thus their asymptotic solution on appropriate time intervals is determined by the preceding.

It is more traditional, however, to use the Prüfer transformation

$$y = \rho_\epsilon(t)\cos(t+\psi_\epsilon(t)), \qquad \dot{y} = -\rho_\epsilon(t)\sin(t+\psi_\epsilon(t)) \tag{52}$$

to obtain a system of differential equations for the nonnegative amplitude ρ_ϵ and the phase ψ_ϵ. As in variation of parameters,

$$\dot{y} = \frac{d\rho_\epsilon}{dt}\cos(t+\psi_\epsilon) - \rho_\epsilon \sin(t+\psi_\epsilon)\left(1+\frac{d\psi_\epsilon}{dt}\right)$$

and

$$\ddot{y} = -\frac{d\rho_\epsilon}{dt}\sin(t+\psi_\epsilon) - \rho_\epsilon \cos(t+\psi_\epsilon)\left(1+\frac{d\psi_\epsilon}{dt}\right)$$

imply a linear algebraic system for $\frac{d\rho_\epsilon}{dt}$ and $\frac{d\psi_\epsilon}{dt}$ that yields

$$\begin{cases} \dfrac{d\rho_\epsilon}{dt} = \epsilon \sin(t+\psi_\epsilon) g(\rho_\epsilon\cos(t+\psi_\epsilon), -\rho_\epsilon \sin(t+\psi_\epsilon)), \\ \dfrac{d\psi_\epsilon}{dt} = \dfrac{\epsilon}{\rho_\epsilon}\cos(t+\psi_\epsilon) g(\rho_\epsilon\cos(t+\psi_\epsilon), -\rho_\epsilon \sin(t+\psi_\epsilon)). \end{cases} \tag{53}$$

The needed initial values $\rho_\epsilon(0)$ and $\psi_\epsilon(0)$ for (53) are likewise uniquely specified since

$$y(0) = \rho_\epsilon(0)\cos\psi_\epsilon(0) \text{ and } \dot{y}(0) = -\rho_\epsilon(0)\sin\psi_\epsilon(0). \tag{54}$$

Since ψ_ϵ occurs in the combination $z \equiv t + \psi_\epsilon(t)$, we can rewrite (53) as a 2π-periodic function of z:

$$\begin{cases} \dfrac{d\rho_\epsilon}{dz} = \dfrac{\epsilon\rho_\epsilon \sin z\, g(\rho_\epsilon\cos z, -\rho_\epsilon \sin z)}{\rho_\epsilon + \epsilon\cos z\, g(\rho_\epsilon\cos z, -\rho_\epsilon\sin z)}, \\ \dfrac{d\psi_\epsilon}{dz} = \dfrac{\epsilon\cos z\, g(\rho_\epsilon\cos z, -\rho_\epsilon \sin z)}{\rho_\epsilon + \epsilon\cos z\, g(\rho_\epsilon\cos z, -\rho_\epsilon\sin z)}. \end{cases} \tag{55}$$

Our ansatz (19) suggests seeking an asymptotic solution for (55) in the form

$$\begin{cases} \rho_\epsilon(t) = R_\epsilon(\tau) + \epsilon U(R_\epsilon(\tau), \Psi_\epsilon(\tau), t, \epsilon), \\ \psi_\epsilon(t) = \Psi_\epsilon(\tau) + \epsilon V(R_\epsilon(\tau), \Psi_\epsilon(\tau), t, \epsilon). \end{cases} \tag{56}$$

The advantage obtained is that the first-order renormalized system is triangular, i.e.,

$$\begin{cases} \dfrac{dR_\epsilon}{d\tau} = \alpha(R_\epsilon,\epsilon) = \dfrac{1}{2\pi}\displaystyle\int_0^{2\pi} \sin z\, g(R_\epsilon\cos z, -R_\epsilon\sin z)dz + \mathcal{O}(\epsilon), \\ \dfrac{d\Psi_\epsilon}{d\tau} = \beta(R_\epsilon,\epsilon) = \dfrac{1}{2\pi R_\epsilon}\displaystyle\int_0^{2\pi} \cos z\, g(R_\epsilon\cos z, -R_\epsilon\sin z)dz + \mathcal{O}(\epsilon). \end{cases} \tag{57}$$

Note that $\alpha(R_\epsilon, 0)$ and $\beta(R_\epsilon, 0)$ are half of the corresponding first harmonic coefficients in the Fourier series for $g(R_\epsilon\cos z, -R_\epsilon\sin z)$ on $0 \le z \le 2\pi$. It's an easy system to solve implicitly as

$$\tau = \int_{R_\epsilon(0)}^{R_\epsilon} \frac{du}{\alpha(u,\epsilon)} \quad \text{and} \quad \Psi_\epsilon(\tau) = \psi_\epsilon(0) + \int_0^\tau \beta(R_\epsilon(p),\epsilon)dp, \tag{58}$$

although specifying where $R_\epsilon(\tau)$ is well defined involves all the anticipated complications. By the chain rule, it also follows that

(59)
$$\begin{cases} U_0(R_\epsilon(\tau),\Psi_\epsilon(\tau),t) = \displaystyle\int_0^t [\sin(s+\Psi_\epsilon(\tau))g(R_\epsilon(\tau)\cos(s+\Psi_\epsilon(\tau)) \\ \qquad\qquad - R_\epsilon(\tau)\sin(s+\Psi_\epsilon(\tau))) - \alpha(R_\epsilon(\tau),0)]ds \\ \text{and} \\ V_0(R_\epsilon(\tau),\Psi_\epsilon(\tau),t) = \dfrac{1}{R_\epsilon(\tau)}\displaystyle\int_0^t [\cos(s+\Psi_\epsilon(\tau))g(R_\epsilon(\tau)\cos(s+\Psi_\epsilon(\tau)) \\ \qquad\qquad - R_\epsilon(\tau)\sin(s+\Psi_\epsilon(\tau))) - \beta(R_\epsilon(\tau),0)]ds. \end{cases}$$

Moreover, using the ansatz (56), we get the secular-free approximations

(60)
$$\begin{cases} y = \rho\cos(t+\psi) = R_\epsilon\cos(t+\Psi_\epsilon) + \epsilon(U_0\cos(t+\Psi_\epsilon) - R_\epsilon V_0\sin(t+\Psi_\epsilon)) + \mathcal{O}(\epsilon^2) \\ \text{and} \\ \dot y = -\rho\sin(t+\psi) = -R_\epsilon\sin(t+\Psi_\epsilon) - \epsilon(U_0\sin(t+\Psi_\epsilon) + R_\epsilon V_0\cos(t+\Psi_\epsilon)) + \mathcal{O}(\epsilon^2). \end{cases}$$

Higher-order approximations follow without difficulty, even for many problems where classical methods break down.

(a) As a first concrete example, consider the Duffing–van der Pol equation

$$\ddot y + y + \epsilon y^3 + \epsilon^2(y^2-1)\dot y = 0, \tag{61}$$

introduced by Benney and Newell (1967). Seeking a solution as

$$y = \rho_\epsilon\cos(t+\psi_\epsilon) \quad \text{and} \quad \dot y = -\rho_\epsilon\sin(t+\psi_\epsilon)$$

provides the periodic forcing $g = y^3 + \epsilon(y^2-1)\dot y$ as

$$\begin{aligned} g(y,\dot y,\epsilon) = {} & \frac{1}{4}\rho_\epsilon^3[3\cos(t+\psi_\epsilon) + \cos 3(t+\psi_\epsilon)] \\ & + \epsilon\Big[-\frac{\rho_\epsilon^3}{4}(\sin(t+\psi_\epsilon) + \sin 3(t+\psi_\epsilon)) \\ & \qquad + \rho_\epsilon^2\sin 2(t+\psi_\epsilon) - \rho_\epsilon\sin(t+\psi_\epsilon)\Big]. \end{aligned}$$

Since its leading term has a trivial $\sin z$ coefficient in its Fourier series and $\frac{3}{4}\rho_\epsilon^3$ as the $\cos z$ coefficient, we will have the amplitude and phase equations

$$\frac{d\rho_\epsilon}{d\tau} = \mathcal{O}(\epsilon) \quad \text{and} \quad \frac{d\psi_\epsilon}{d\tau} = \frac{3}{8}\rho_\epsilon^2 + \mathcal{O}(\epsilon).$$

(Note that Cox and Roberts (1995) and Roberts (1997) attain such reductions efficiently by *normal form* transformations implemented using REDUCE. Mudavanhu (2000) obtains the same results and corresponding higher-order terms via a renormalization method automated using MAPLE.) Indeed, our results suggest the more efficient introduction of the slower-time $\kappa = \epsilon^2 t$. Incorporating κ and using the next

terms in (57), we obtain

$$\begin{cases} \dfrac{dR_\epsilon}{d\kappa} = \dfrac{1}{2}R_\epsilon\left(1 - \dfrac{1}{4}R_\epsilon^2\right) + \mathcal{O}(\epsilon) \\ \text{and} \\ \dfrac{d\Psi_\epsilon}{d\tau} = \dfrac{3}{8}R_\epsilon^2 - \dfrac{15\epsilon}{256}R_\epsilon^4 + \mathcal{O}(\epsilon^2). \end{cases} \tag{62}$$

Solving the limiting Bernoulli equation determines an expansion for

$$R_\epsilon(\kappa) = R_0(\kappa) + \epsilon R_1(\kappa) + \epsilon^2(\cdots) \tag{63}$$

as an exponentially decaying amplitude for all $\kappa \geq 0$, with leading term

$$R_0(\kappa) = \frac{2}{\sqrt{1 - (1 - 4/\rho_\epsilon^2(0))e^{-\kappa}}}.$$

The resulting limit cycle behavior follows with the phase

$$\Psi_\epsilon(t) = \psi_\epsilon(0) + \frac{3}{8}\int_0^t R_\epsilon^2(\epsilon^2 s)\left(1 - \frac{5\epsilon}{32}R_\epsilon^2(\epsilon^2 s)\right)ds + \mathcal{O}(\epsilon^2 t). \tag{64}$$

In the integrand, it is clearly preferable to represent $R_\epsilon(\kappa)$ as the sum of its steady-state limit plus an exponentially decaying transient. Higher-order approximations to the solution follow as in (60).

Alternatively, we can use the transformation (51) and the spectral decomposition $M = iV\Lambda V^{-1}$ for a nonsingular modal matrix $V = \left(\begin{smallmatrix} 1 & 1 \\ i & -i \end{smallmatrix}\right)$ and $\Lambda = \left(\begin{smallmatrix} -1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$. Directly applying our basic ansatz,

$$x_\epsilon(t) = Ve^{i\Lambda t}V^{-1}(A_\epsilon + \epsilon U(A_\epsilon(\tau), \epsilon)), \tag{65}$$

where $A_\epsilon = \left(\begin{smallmatrix} a_\epsilon \\ \bar{a}_\epsilon \end{smallmatrix}\right)$, for complex conjugates a_ϵ and $\bar{a}_\epsilon$, involves, to leading order, the amplitude equation

$$\frac{dA_\epsilon}{d\tau} = \langle e^{-i\Lambda t}V^{-1}N(Ve^{i\Lambda t}x, 0)\rangle + \mathcal{O}(\epsilon) = \frac{3}{2}i|a_\epsilon|^2\begin{pmatrix} a_\epsilon \\ -\bar{a}_\epsilon \end{pmatrix} + \mathcal{O}(\epsilon)$$

and the secular-free correction term

$$\begin{aligned} U_0(A_\epsilon, t) &= \int_0^t \{e^{-i\Lambda s}V^{-1}N(Ve^{i\Lambda s}x, 0)\}ds \\ &= \frac{1}{4}\begin{pmatrix} a_\epsilon^3 e^{-2it} - 3\bar{a}_\epsilon|a_\epsilon|^2 e^{2it} + \frac{1}{2}\bar{a}_\epsilon^3 e^{4it} \\ \bar{a}_\epsilon^3 e^{2it} - 3a_\epsilon|a_\epsilon|^2 e^{-2it} + \frac{1}{2}a_\epsilon^3 e^{-4it} \end{pmatrix}. \end{aligned}$$

Letting $a_\epsilon = \frac{R_\epsilon}{2}e^{-i\Psi_\epsilon}$ provides the amplitude and phase equations $\frac{dR_\epsilon}{d\tau} = \mathcal{O}(\epsilon)$, $\frac{d\Psi_\epsilon}{d\tau} = \frac{3}{8}R_\epsilon^2 + \mathcal{O}(\epsilon)$ as before and the corresponding asymptotic approximation

$$y = R_\epsilon\cos(t + \Psi_\epsilon) + \epsilon\frac{R_\epsilon}{16}\left[3\cos(t + \Psi_\epsilon) + \frac{1}{2}\cos 3(t + \Psi_\epsilon)\right] + \epsilon^2(\cdots). \tag{66}$$

Higher-order approximations follow in a straightforward fashion.

(b) We finally seek the RG equations resulting from two weakly coupled van der Pol oscillators

$$\ddot{y} + \Omega y + \epsilon(I - \mathcal{C}^2)\dot{y} = \epsilon\mathcal{B}(\alpha y + \beta\dot{y}), \tag{67}$$

where

$$y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \qquad \Omega = \begin{pmatrix} 1 & 0 \\ 0 & 1+\Delta \end{pmatrix}, \qquad \mathcal{C} = \begin{pmatrix} y_1 & 0 \\ 0 & y_2 \end{pmatrix}, \qquad \mathcal{B} = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} \tag{68}$$

and I is an identity matrix (cf. Reinhall and Storti (2000) and Low (2002)). Here α and β are coupling constants, and Δ is a detuning parameter related to the difference in the natural frequencies of the two oscillators. We now seek asymptotic solutions of the form

$$\begin{cases} y_1(t,\epsilon) = R_1(\tau,\epsilon)\cos(t + \Psi_1(\tau,\epsilon)) + \epsilon(\cdots), \\ \\ y_1(t,\epsilon) = R_2(\tau,\epsilon)\cos(t + \Psi_2(\tau,\epsilon)) + \epsilon(\cdots) \end{cases} \tag{69}$$

for the slow time $\tau = \epsilon t$, where R_j and Ψ_j, for $j = 1$ and 2, represent the amplitude and phase modulations. The functions y_j are said to be *phase locked* when the difference

$$\phi_\epsilon = \Psi_{2\epsilon} - \Psi_{1\epsilon} \tag{70}$$

is a constant. When the oscillators are running at unequal frequencies (i.e., $\Delta \neq 0$), ϕ_ϵ will grow unbounded, defining a condition known as a *phase drift*. An intermediate situation exists when ϕ_ϵ varies periodically, a condition known as *phase entrainment*.

Applying our basic ansatz, by first transforming to a four-dimensional system of the form (1), we systematically obtain the RG equations

$$\begin{aligned} 2\frac{dA_{1\epsilon}}{d\tau} &= A_{1\epsilon}(1 - |A_{1\epsilon}|^2) - \beta(A_{1\epsilon} - A_{2\epsilon}) + i\alpha(A_{1\epsilon} - A_{2\epsilon}) + \epsilon(\cdots), \\ 2\frac{dA_{2\epsilon}}{dt} &= A_{2\epsilon}(1 - |A_{2\epsilon}|^2) - \beta(A_{2\epsilon} - A_{1\epsilon}) + i\alpha(A_{2\epsilon} - A_{1\epsilon}) + i\frac{\Delta}{2}A_{1\epsilon} + \epsilon(\cdots). \end{aligned}$$

Letting $A_{j\epsilon} = R_{j\epsilon}e^{-i\Psi_{j\epsilon}}$ for $j = 1$ and 2, we get the system of three slowly varying RG equations

$$\begin{cases} 2\dfrac{dR_{1\epsilon}}{d\tau} = (1 - R_{1\epsilon}^2)R_{1\epsilon} - \beta(R_{1\epsilon} - R_{2\epsilon}\cos\phi_\epsilon) + \alpha R_{2\epsilon}\sin\phi_\epsilon + \epsilon(\cdots), \\ 2\dfrac{dR_{2\epsilon}}{d\tau} = (1 - R_{2\epsilon}^2)R_{2\epsilon} - \beta(R_{2\epsilon} - R_{1\epsilon}\cos\phi_\epsilon) \ - \alpha R_{1\epsilon}\sin\phi_\epsilon + \epsilon(\cdots), \\ 2\dfrac{d\phi_\epsilon}{d\tau} = \Delta - \beta\left(\dfrac{R_{2\epsilon}}{R_{1\epsilon}} - \dfrac{R_{1\epsilon}}{R_{2\epsilon}}\right)\sin\phi_\epsilon - \alpha\left(\dfrac{R_{2\epsilon}}{R_{1\epsilon}} + \dfrac{R_{1\epsilon}}{R_{2\epsilon}}\right)\cos\phi_\epsilon + \epsilon(\cdots). \end{cases} \tag{71}$$

Stability analyses of (67) can be carried out based on these and higher-order amplitude equations (cf. Chakraborty and Rand (1988)).

Relation to two-timing and other classical techniques. The asymptotic solution of the initial value problem (3)

$$\dot{z} = \epsilon f(z, t, \epsilon)$$

(in standard form) could be obtained using the two-time ansatz

$$z = z(t, \tau, \epsilon)$$

for the bounded slow-time $\tau = \epsilon t$. The chain rule implies that

$$\dot{z} = \frac{\partial z}{\partial t} + \epsilon \frac{\partial z}{\partial \tau},$$

and thus substituting a power series expansion

$$z(t, \tau, \epsilon) = z_0(t, \tau) + \epsilon z_1(t, \tau) + \cdots \tag{72}$$

into (3) requires that

$$\frac{\partial z_0}{\partial t} = 0 \tag{73}$$

and

$$\frac{\partial z_j}{\partial t} = g_{j-1}(t, z_0, z_1, \ldots, z_{j-1}) - \frac{\partial z_{j-1}}{\partial \tau} \tag{74}$$

for each $j \geq 1$. Here

$$f(z(t, \tau, \epsilon), t, \epsilon) \sim \sum_{j=0}^{\infty} \epsilon^j g_j(t, z_0, \ldots, z_{j-1}, z_j),$$

where

$$g_j(t, z_0, \ldots, z_{j-1}, z_j) - f_z(z_0(t, \tau), t, 0) z_j$$

is a known function of the earlier coefficients $z_0, z_1, \ldots, z_{j-1}$ and t.

We first obtain the representation

$$z_0(t, \tau) = A_0(\tau) \tag{75}$$

from integrating (73), for some unspecified $A_0(\tau)$. Taking $j = 1$, we then find that

$$\frac{\partial z_1}{\partial t} = f_0(A_0(\tau), t) - \frac{dA_0}{d\tau}. \tag{76}$$

Recall that f_0 is a periodic function of t. To get the boundedness of z_1 as $t \to \infty$ requires the right-hand side to have zero average, i.e., A_0 must be the unique solution of the initial value problem for

$$\frac{dA_0}{d\tau} = \langle f_0(A_0(\tau), t) \rangle. \tag{25}$$

This leaves $\frac{\partial z_1}{\partial t} = \{f_0(A_0(\tau), t)\}$, and so

$$z_1(t, \tau) = A_1(\tau) + U_0(A_0(\tau), t) \tag{77}$$

for an unknown A_1 and the bounded function $U_0 = \int_0^t \{f_0(A_0(\tau), s)\} ds$, already encountered. If we next consider (74) for $j = 2$, the boundedness of z_2 will require the

average of the right-hand side to be zero. This, however, shows that A_1 must satisfy an initial value problem of the form

$$\frac{dA_1}{d\tau} = \left\langle \frac{\partial f_0}{\partial z}(A_0(\tau), t) \right\rangle A_1 + \widetilde{a_1}(A_0(\tau)),$$

with a known inhomogeneity $\widetilde{a}_1$ and the trivial initial value. The unique solution follows as in (26). Continuing, in this manner, to use the Fredholm alternative to get a bounded solution at every stage, we obtain our two-time expansion to any order.

Murdock and Wang (1996) prove that this result is asymptotically valid, to all orders, for finite τ. An attempt to comprehend renormalization has thus provided an opportunity to rethink two-timing. We point out, however, that a less restrictive method of *slowly varying amplitudes* is often used in applications (cf., e.g., the final chapter of Haberman (1998), in addition to the literature already cited). It compares to our ansatz (19),

$$z_\epsilon(t) = A_\epsilon(\tau) + \epsilon U(A_\epsilon(\tau), t, \epsilon),$$

rather than the more general two-timing expansion (72). The idea is to seek an amplitude $A_\epsilon(\tau)$, varying with the slow time τ and ϵ, so that secular terms in the resulting expansion are removed by appropriately selecting successive terms in the power series expansion of the amplitude (or envelope or RG) equation

$$\frac{dA_\epsilon}{d\tau} = a(A_\epsilon, \epsilon).$$

The relationship between *asymptotic matching* or *boundary layer theory* (cf., e.g., Il'in (1992) or O'Malley (1991)) and renormalization can be illustrated by considering the singularly perturbed initial value problem

$$\left\{ \begin{array}{l} \dot{x} = xy + \epsilon a x^3, \\ \epsilon \dot{y} = -y + \epsilon b x^2 \end{array} \right. \tag{78}$$

(introduced in Kuwamura (2001)) on $t \geq 0$ when the constants a and b satisfy $a+b < 0$.

The special case $b = 0$ is of special interest because it is exactly solvable. Because

$$y_\epsilon(t) = e^{-t/\epsilon} y(0),$$

x must be the unique solution

$$x_\epsilon(t) = \frac{e^{\epsilon y(0) e^{-t/\epsilon}} x(0)}{\sqrt{1 - 2a\epsilon x^2(0) \int_0^t e^{\epsilon y(0) e^{-r/\epsilon}} dr}}$$

of the resulting Bernoulli equation. Note that the solution decays algebraically to zero when $a < 0$. It is nearly constant for $a = 0$, and it blows up when

$$\epsilon \int_0^t e^{2\epsilon y(0) e^{-r/\epsilon}} dr = \frac{1}{2ax^2(0)}$$

if $a > 0$.

If we introduce the *fast time*

$$\lambda = \frac{t}{\epsilon}$$

into the corresponding *inner problem*

$$\begin{cases} \dfrac{dx}{d\lambda} = \epsilon xy + \epsilon^2 a x^3, \\ \dfrac{dy}{d\lambda} = -y + \epsilon b x^2, \end{cases}$$

we naturally seek the *inner expansion*

$$\begin{cases} u(\lambda,\epsilon) = u_0(\lambda) + \epsilon u_1(\lambda) + \epsilon^2 u_2(\lambda) + \cdots, \\ v(\lambda,\epsilon) = v_0(\lambda) + \epsilon v_1(\lambda) + \epsilon^2 v_2(\lambda) + \cdots. \end{cases}$$

Proceeding termwise, in the naive manner, we get

$$\begin{aligned} &u_0(\lambda) = x(0), \quad v_0(\lambda) = e^{-\lambda} y(0), \\ &u_1(\lambda) = -(1-e^{-\lambda})x^2(0), \qquad v_1(\lambda) = b(1-e^{-\lambda})x^2(0), \end{aligned}$$

and then, from

$$\frac{du_2}{d\lambda} = u_0 v_1 + u_1 v_0 + a u_0^3$$

and

$$\frac{dv_2}{d\lambda} = -v_2 + 2b u_0 u_1,$$

we get

$$u_2(\lambda) = (a+b)x^3(0)\lambda + bx^3(0)(e^{-\lambda}-1) + \frac{1}{2}(e^{-\lambda}-1)^2 x(0) y^2(0)$$

and

$$v_2(\lambda) = 2bx^2(0)y(0)\left[1 - e^{-\lambda} - \lambda e^{-\lambda}\right].$$

The Tikhonov–Levinson theory applies for t finite and guarantees that the inner expansion can be written as the asymptotic sum

$$\begin{aligned} u(\lambda,\epsilon) &= X(t,\epsilon) + \epsilon\xi(\lambda,\epsilon), \\ v(\lambda,\epsilon) &= \epsilon Y(t,\epsilon) + \eta(\lambda,\epsilon), \end{aligned}$$

where $\binom{X}{\epsilon Y}$ is the *outer expansion* and $\binom{\epsilon\xi}{\eta}$ is the *inner layer correction* that decays to zero exponentially as $\lambda \to \infty$. Replacing λ by t/ϵ defines the surviving outer expansion

$$\begin{aligned} X(t,\epsilon) &= x(0) + \epsilon(x(0)y(0) + (a+b)x^3(0)t) + \epsilon^2(\cdots), \\ \epsilon Y(t,\epsilon) &= \epsilon b x^2(0) + \epsilon^2(\cdots). \end{aligned}$$

Note the secular behavior visible as $\tau = \epsilon t \to \infty$. It is not fixable using Hoppensteadt (1966), because there is no asymptotic stability in t. We can eliminate the secular term, however, by renormalizing. Setting

$$x(0) = A_\epsilon(\tau) + \epsilon P(A_\epsilon(\tau), t, \epsilon) \quad \text{and} \quad y(0) = B_\epsilon(\tau) + \epsilon Q(A_\epsilon(\tau), t, \epsilon)$$

and by picking

$$P_0(A_0,t) = -(a+b)A_0^3 t \quad \text{and} \quad Q_0(A_0,t) = 0,$$

we get a secular-free approximation. Constancy of $x(0)$ and $y(0)$, however, forces A_0 and B_0 to satisfy the limiting amplitude equations

$$\frac{dA_0}{d\tau} = (a+b)A_0^3 \quad \text{and} \quad \frac{dB_0}{d\tau} = 0.$$

This has the algebraically decaying solution

$$A_0(\tau) = \frac{x(0)}{\sqrt{1-2(a+b)x^2(0)\tau}}$$

when $a+b<0$, already observed in the special case $b=0$. Higher-order terms follow, without difficulty. One could, analogously, also directly seek the asymptotic solution as a function of the three times λ, t, and τ.

Acknowledgments. The authors are grateful to a number of colleagues for their encouragement and suggestions. Special thanks go to Bernard Deconinck, Richard Haberman, Jerry Kevorkian, Rachel Kuske, James Murdock, Yoshi Oono, Hong Qian, and Tony Roberts. R. E. O'Malley, Jr., also wishes to thank the Fields Institute for Research in the Mathematical Sciences, Toronto, and the School of Mathematics and the Institute for Mathematics and its Applications at the University of Minnesota for their hospitality and support.

REFERENCES

G.I. Barenblatt (1996), *Scaling, Self-Similarity, and Intermediate Asymptotics*, Cambridge University Press, Cambridge, UK.

D.J. Benney and A.C. Newell (1967), *Sequential time closures for interacting random waves*, J. Math. Phys., 46, pp. 363–393.

N.N. Bogoliubov and Y.A. Mitropolsky (1961), *Asymptotic Methods in the Theory of Nonlinear Oscillators*, Gordon and Breach, New York.

D.L. Bosley (1996), *A technique for the numerical verification of asymptotic expansions*, SIAM Rev., 38, pp. 128–135.

T. Chakraborty and R. Rand (1988), *The transition from phase locking to drift in a system of two weakly coupled van der Pol oscillators*, Int. J. Nonlinear Mech., 23, pp. 369–376.

L. Chen, N. Goldenfeld, and Y. Oono (1996), *Renormalization group and singular perturbations: Multiple scales, boundary layers, and reductive perturbation theory*, Phys. Rev. E, 54, pp. 376–394.

J.A. Cochran (1962), *Problems in Singular Perturbation Theory*, Ph.D. thesis, Stanford University, Stanford, CA.

P.H. Coullet and E.A. Spiegel (1983), *Amplitude equations for systems with competing instabilities*, SIAM J. Appl. Math., 43, pp. 776–821.

S.M. Cox and A.J. Roberts (1995), *Initial conditions for models of dynamical systems*, Physica D, 85, pp. 126–141.

L. Debnath (1997), *Nonlinear Partial Differential Equations*, Birkhäuser Boston, Cambridge, MA.

W. Eckhaus (1992), *On modulation equations of the Ginzburg-Landau type*, in ICIAM 91, Proceedings of the Second International Congress on Industrial and Applied Mathematics, Washington, DC, 1991, R.E. O'Malley, Jr., ed., SIAM, Philadelphia, pp. 83–98.

S. Ei, K. Fujii, and T. Kunihiro (2000), *Renormalization group method for reduction of evolution equations; Invariant manifolds and arguments*, Ann. Physics, 280, pp. 236–298.

W.M. Greenlee and R.E. Snow (1975), *Two-timing on the half-line for damped oscillation equations*, J. Math. Anal. Appl., 51, pp. 394–428.

R. Haberman (1998), *Elementary Applied Partial Differential Equations,* 3rd ed., Prentice–Hall, Upper Saddle River, NJ.

F.C. Hoppensteadt (1966), *Singular perturbations on the infinite interval*, Trans. Amer. Math. Soc., 123, pp. 521–535.

A.M. Il'in (1992), *Matching of Asymptotic Expansions of Solutions of Boundary Value Problems*, American Mathematical Society, Providence, RI.

G.A. Jarrad (2001), *Perturbations, Chaos, and Waves*, Ph.D. thesis, University of South Australia, Adelaide, Australia.

J. Kevorkian and J.D. Cole (1996), *Multiple Scales and Singular Perturbation Methods*, Springer, New York.

M. Kuwamura (2001), *A perspective of renormalization group methods*, Japan J. Indust. Appl. Math., 18, pp. 739–768.

G.E. Kuzmak (1959), *Asympotic solutions of nonlinear second order differential equations with variable coefficients*, J. Appl. Math. Mech., 23, pp. 730–744.

L.A. Low (2002), *Stability of Coupled van der Pol Oscillators and Applications to Gait Control in Simple Animals*, Ph.D. thesis, University of Washington, Seattle, WA.

J.J. Mahony (1962), *An expansion method for singular perturbation problems*, J. Austral. Math. Soc., 2, pp. 440–463.

I. Moise and M. Ziane (2001), *Renormalization group method. Applications to partial differential equations*, J. Dynam. Differential Equations, 13, pp. 275–321.

J.A. Morrison (1966), *Comparison of the modified method of averaging and the two variable expansion procedure*, SIAM Rev., 8, pp. 66–85.

B. Mudavanhu (2000), *Singular Perturbation Techniques: Multiple Scales, Averaging, Renormalization Group and Invariance Condition Methods*, manuscript.

B. Mudavanhu and R.E. O'Malley, Jr. (2001), *A renormalization group method for nonlinear oscillators*, Studies in Appl. Math., 107, pp. 63–79.

J.A. Murdock (1991), *Perturbations*, Wiley, New York.

J.A. Murdock and L.-C. Wang (1996), *Validity of the multiple scale method for very long intervals*, Z. Angew. Math. Phys., 47, pp. 760–789.

J.C. Neu (1980), *The method of near-identity transformations and its applications*, SIAM J. Appl. Math., 38, pp. 189–208.

K. Nozaki and Y. Oono (2001), *Renormalization-group theoretical reduction*, Phys. Rev. E, 63, paper 046101.

R.E. O'Malley, Jr. (1991), *Singular Perturbation Methods for Ordinary Differential Equations,* Springer-Verlag, New York.

Y. Oono (2000), *Renormalization and asymptotics*, Int. J. Modern Phys. B, 14, pp. 1327–1361.

K. Promislow (2001), *A Renormalization Method for Modulational Stability of Quasi-Steady Patterns in Dispersive Systems*, preprint.

P.G. Reinhall and D.W. Storti (2000), *Phase-locked mode stability for coupled van der Pol oscillators*, J. Vibrations and Acoustics, 122, pp. 1–7.

A.J. Roberts (1997), *Low-dimensional modeling of dynamics via computer algebra*, Comput. Phys. Comm., 100, pp. 215–230.

J.A. Sanders and F. Verhulst (1985), *Averaging Methods in Nonlinear Dynamical Systems*, Springer-Verlag, New York.

D.R. Smith (1985), *Singular-Perturbation Theory*, Cambridge University Press, Cambridge, UK.

A.B. Vasil'eva, V.F. Butuzov, and L.V. Kalachev (1995), *The Boundary Function Method for Singular Perturbation Problems*, SIAM Stud. Appl. Math. 14, SIAM, Philadelphia.

G.B. Whitham (1974), *Linear and Nonlinear Waves*, Wiley, New York.

D. Wirosoetisno, T.G. Shepherd, and R.M. Temam (2002), *Free gravity waves and balanced dynamics*, J. Atmos. Sci., to appear.

S.L. Woodruff (1993), *The use of an invariance condition in the solution of multiple scale singular perturbation problems: Ordinary differential equations*, Stud. Appl. Math., 90, pp. 225–248.

S.L. Woodruff (1995), *A uniformly valid asymptotic solution to a matrix system of ordinary differential equations and a proof of its validity*, Stud. Appl. Math., 94, pp. 393–413.

SIAM J. APPL. MATH.
Vol. 63, No. 2, pp. 398–422

A DYNAMIC PRIORITY QUEUE MODEL FOR SIMULTANEOUS SERVICE OF TWO TRAFFIC TYPES*

CHARLES KNESSL†, DOO IL CHOI‡, AND CHARLES TIER†

Abstract. We consider a priority queue with a dynamic, queue-length-threshold scheduling policy. Customers are classed into two types (type-1 and type-2), and the service order of the two classes depends on the queue length of the type-1 queue. The high priority (type-2) class (e.g., voice) is served until the low priority (e.g., data) queue exceeds the threshold L, at which time service is given to the low priority class until its queue length decreases to L. The arrivals of the two classes follow independent Poisson processes, and the service time of each customer has an exponential distribution with parameter μ. We derive the balance equations in the steady state, and explicitly obtain the joint probability generating function for the queue lengths of the two customer classes. This gives the joint queue length distribution as an integral. We then obtain detailed asymptotic results for the joint distribution. In particular, we study the tail behavior. We also discuss heavy traffic diffusion approximations for this model.

Key words. dynamic priority queue, integral representation, asymptotic approximation

AMS subject classifications. 60K25, 34E05, 34E20

PII. S0036139901390842

1. Introduction.

1.1. Background. Due to recent applications in ATM (asynchronous transfer mode) networks, there has been renewed interest in priority queues. Here we consider the following model for providing simultaneous service to two classes of customers with different service requirements. There are two classes of customers (called type-1 and type-2 customers) and a single server. The arrivals of type-1 and type-2 customers follow independent Poisson processes with rates λ_1 and λ_2, respectively. The two streams of customers are accommodated into two queues with infinite capacities. Customers in each class are served on a first-come first-served basis. The service order of the two classes is determined by the queue-length-threshold (QLT) scheduling policy. First, we place a threshold L on the queue for type-1 customers. If the number of type-1 customers in the queue is less than or equal to the threshold L, the type-2 customers are served. Otherwise, the type-1 customers are served. If one of the queues is empty, the customers in other queue are served. The service time of a customer of either type has an exponential distribution with parameter μ. This queueing system is called a dynamic (or hybrid) priority queue with QLT scheduling policy (Figure 1.1).

In many modern applications, some classes of customers may have different service requirements than others. A method for supporting the different classes is the use of priority queues. There are static and dynamic priority queues. Examples of static priority queues, including nonpreemptive and preemptive queues, are given in [1], [2], [3], [4]. In such systems, the high priority class (e.g., voice) has much more

*Received by the editors June 13, 2001; accepted for publication (in revised form) April 15, 2002; published electronically November 19, 2002. This research was supported in part by NSF grant DMS 99-71656.

http://www.siam.org/journals/siap/63-2/39084.html

†University of Illinois at Chicago, 851 South Morgan Street, Chicago, IL 60607 (knessl@uic.edu, tier@math.uic.edu).

‡Department of Applied Mathematics, Halla University, Kangwon-do, South Korea. The research of this author was supported by grant KMS 2001-08 from KOSEF (dichoi@hit.halla.ac.kr).

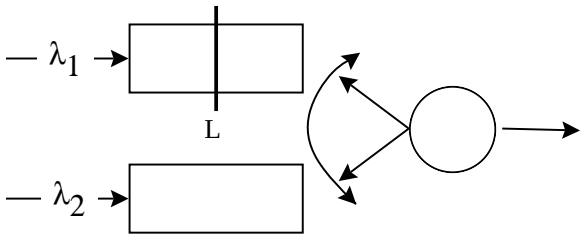


FIG. 1.1. *Model of a dynamic priority queue.*

stringent delay requirements than the low priority class (e.g., data). Therefore, the high priority class may have better performance than its delay requirements, but the low priority class may suffer from unacceptably long delays. In order to overcome this shortcoming, dynamic priority queues have been proposed and studied in [6], [7], [8]. Such models make it possible to improve the performance of the low priority class, while still meeting the service requirement (delay) of the high priority class.

This paper gives exact and asymptotic analyses of a dynamic priority queue with QLT scheduling. The results can be applied to traffic control, and in particular to satisfying the quality of service (QoS) requirements of real-time and nonreal-time traffic in ATM networks [9]. For example, in ATM networks, data is loss-sensitive but delay-insensitive, while voice is delay-sensitive but loss-insensitive. Thus, the high priority class (type-2 customers) may be considered as real-time traffic such as voice and the low priority class (type-1 customers) may be considered as nonreal-time traffic such as data. The value of L is chosen to insure that the type-2 traffic is within its delay requirements, i.e., the more stringent the delay requirement, the larger the value of L. The QLT scheduling policy was shown in [5] to be able to provide higher utilization and more flexible performance, with the proper adjustment of L, than several other scheduling schemes. Also, when $L = \infty$ in our model (or $L = 0$ in reversed priority), this is the well-known nonpreemptive static priority queue.

There has been much previous work on dynamic and static priority queues. Queueing systems with static priority are discussed in [1], [2], [3], [4]. Fratini [6] analyzed a dynamic single server priority queue, with the same scheduling policy as ours. He assumed that the queue for type-1 customers had infinite capacity and that the queue for type-2 customers had a finite capacity (K). The two classes had different general service time distributions. Using a discrete-time Markov chain embedded at service completion epochs, he identified the state transition probability matrix P and numerically analyzed the stationary probability vector x defined by

$$xP = x, \qquad xe = 1.$$

As an application for real-time and nonreal-time traffic in ATM networks, Lee and Sengupta [7] considered a dynamic priority queue. Their model is different from ours in that if the queue length in the buffer of real-time traffic is greater than a threshold, the next packet to be served is from the buffer of real-time traffic. Otherwise, the server checks the type of the packet that it has just served, and serves a packet of the other type. The real-time and nonreal-time traffic follow independent Poisson processes, and the queues have infinite capacities. By using the embedded Markov chain method at the service completion epochs, they obtained the joint probability generating function for queue lengths at these epochs. Then, they derived the marginal probability generating function for each queue length at an arbitrary time

using the obtained generating function and the PASTA (Poisson arrivals see time averages) property.

Recently, Choi and Choi [8] considered the $MMPP, M/G/1$ finite capacity queue with the same scheduling policy as ours. In this paper, the arrivals of type-1 customers follow a Poisson process, and the arrivals of type-2 customers follow a Markov-modulated Poisson process (MMPP) [10]. They applied this model to traffic control for real-time and nonreal-time traffic in ATM networks and assumed that the arrival of type-2 customers (real-time traffic) is MMPP, thus modeling the burstiness of real-time traffic. By using the embedded Markov chain method, they numerically obtained the marginal probability distribution for each queue.

Our model is in some respects simpler than those in [6], [7], [8]. However, we are able to provide more explicit analytic expressions. We also give detailed asymptotic results for the tail behavior of the joint queue length distribution. We first derive the balance equations for our queueing system. In section 2, we consider a simplified model in which the server leaves the type-2 queue only when the type-1 queue exceeds L. We consider the full model in section 3. For both models, we explicitly obtain the joint probability generating function for the queue lengths. By inverting the generating function, we obtain the joint queue length distribution as an integral. From this integral we compute asymptotic approximations for the tail behavior of the joint distribution. This leads to simple formulas that reveal the basic qualitative properties of the distribution. We also discuss heavy traffic diffusion approximations for the models. The main exact results are summarized in Theorems 2.1 and 3.1, the tail probabilities are given in Theorems 2.2–2.4, and the heavy traffic diffusion results are summarized in Theorems 2.5, 2.6, 3.2, and 3.3. The numerical accuracy of our asymptotic results is demonstrated in section 4.

1.2. Formulation of our queueing system. Let $N_1(t)$ and $N_2(t)$ be the queue length of the type-1 and type-2 customers, respectively, at time t. We also let λ_1 and λ_2 be the arrival rates of the type-1 and type-2 customers, and μ is the service rate of both classes. The process $(N_1(t), N_2(t))$ is Markov. We focus on analyzing the stationary probability distribution for the joint queue length, defined by

$$p(m,n) = \lim_{t\to\infty} Pr[N_1(t) = m, N_2(t) = n], \qquad m,\ n \geq 0.$$

We then obtain the following balance equations for the queueing system,

$$m \geq L+1,\ n \geq 1: \tag{1.1}$$
$$(\mu+\lambda_1+\lambda_2)p(m,n) = \lambda_1 p(m-1,n) + \lambda_2 p(m,n-1) + \mu p(m+1,n),$$

$$m = L,\ n \geq 1: \tag{1.2}$$
$$(\mu+\lambda_1+\lambda_2)p(L,n) = \lambda_1 p(L-1,n) + \lambda_2 p(L,n-1) + \mu p(L+1,n) + \mu p(L,n+1),$$

$$1 \leq m \leq L-1,\ n \geq 1: \tag{1.3}$$
$$(\mu+\lambda_1+\lambda_2)p(m,n) = \lambda_1 p(m-1,n) + \lambda_2 p(m,n-1) + \mu p(m,n+1),$$

$$m \geq L+1,\ n = 0: \tag{1.4}$$
$$(\mu+\lambda_1+\lambda_2)p(m,0) = \lambda_1 p(m-1,0) + \mu p(m+1,0),$$

$$m = L,\ n = 0: \tag{1.5}$$
$$(\mu + \lambda_1 + \lambda_2)p(L,0) = \lambda_1 p(L-1,0) + \mu p(L+1,0) + \mu p(L,1),$$

$$1 \le m \le L-1,\ n = 0: \tag{1.6}$$
$$(\mu + \lambda_1 + \lambda_2)p(m,0) = \lambda_1 p(m-1,0) + \mu p(m+1,0) + \mu p(m,1),$$

$$m = 0,\ n \ge 1: \tag{1.7}$$
$$(\mu + \lambda_1 + \lambda_2)p(0,n) = \lambda_2 p(0,n-1) + \mu p(0,n+1),$$

$$m = 0,\ n = 0: \tag{1.8}$$
$$(\lambda_1 + \lambda_2)p(0,0) = \mu p(0,1) + \mu p(1,0),$$

and the normalization

$$\sum_{m,n=0}^{\infty} p(m,n) = 1. \tag{1.9}$$

We assume that the following stability condition holds:

$$\rho_1 + \rho_2 < 1, \tag{1.10}$$

where $\rho_i = \lambda_i/\mu, i = 1, 2$.

2. A simplified model. We start by analyzing a simplified version of our queueing system. We assume that if $N_1(t) \le L$ and $N_2(t) = 0$, then the server becomes idle. Thus the server remains at the high priority (type-2) queue and leaves only when $N_1(t) > L$. This model is reasonable if there is a large overhead associated with switching between the two queues. We shall show that the solution to this simplified model may be used as a "building block" to analyze the more complicated system (1.1)–(1.9). In particular, the results in Theorems 3.1–3.3 in section 3 use the corresponding results in this section.

2.1. Exact solution. For the simplified model, once $N_1(t) \ge L$ for some time t, we have $N_1(t) \ge L$ for all future times. Then, $p(m,n)$ is nonzero only for $m \ge L, n \ge 0$, and we may set $p(L-1,n) = 0$. The boundary condition (1.5) is changed to

$$(\lambda_1 + \lambda_2)p(L,0) = \mu p(L+1,0) + \mu p(L,1), \tag{2.1}$$

and we need consider only (1.1), (1.2), (1.4), and (2.1).

To obtain the joint queue length distribution, we introduce the following generating functions:

$$G(z,w) = \sum_{n=0}^{\infty}\sum_{m=L}^{\infty} p(m,n) z^{m-L} w^n,$$
$$H(w) = \sum_{n=0}^{\infty} p(L,n) w^n.$$

From (1.1), (1.2), (1.4), and (2.1), we obtain the equation

$$\begin{aligned}\left[\lambda_1 z + \lambda_2 w - (\mu + \lambda_1 + \lambda_2) + \frac{\mu}{z}\right] G(z,w) \\ = \mu\left(\frac{1}{w} - 1\right) p(L,0) + \mu\left(\frac{1}{z} - \frac{1}{w}\right) H(w).\end{aligned} \tag{2.2}$$

The left-hand side may be rewritten as

$$\frac{\lambda_1}{z}(z - z_*(w))(z - \tilde{z}(w))G(z,w),$$

where

$$\begin{aligned} z_*(w) &= \frac{1}{2\lambda_1}\left[\mu + \lambda_1 + \lambda_2 - \lambda_2 w - \sqrt{(\mu + \lambda_1 + \lambda_2 - \lambda_2 w)^2 - 4\lambda_1\mu}\right], \\ \tilde{z}(w) &= \frac{1}{2\lambda_1}\left[\mu + \lambda_1 + \lambda_2 - \lambda_2 w + \sqrt{(\mu + \lambda_1 + \lambda_2 - \lambda_2 w)^2 - 4\lambda_1\mu}\right]. \end{aligned}$$

Noting that $|z_*(w)| < 1$ and requiring analyticity of $G(z,w)$ in the complex domain $\{|z| < 1, |w| < 1\}$, we set $z = z_*(w)$ and obtain $H(w)$ as

$$H(w) = \frac{z_*(w)(w-1)}{w - z_*(w)} p(L,0). \tag{2.3}$$

Substituting (2.3) into (2.2), we obtain

$$G(z,w) = \frac{p(L,0)}{\rho_1}\left(\frac{1-w}{w - z_*(w)}\right)\frac{1}{z - \tilde{z}(w)}. \tag{2.4}$$

Using the normalization condition $G(1,1) = 1$, $p(L,0)$ is given by

$$p(L,0) = 1 - \rho_1 - \rho_2.$$

By inverting the generating function $G(z,w)$, we obtain an integral representation for the joint probabilities $p(m,n)$, as given below.

THEOREM 2.1. *The joint queue length distribution for type-*1 *and type-*2 *customers for the simplified model is given by*

$$\begin{aligned} p(L+m,n) = \frac{p(L,0)}{2\pi i}\int_C \frac{1-w}{1-\rho_1 w\tilde{z}(w)}\frac{1}{[\tilde{z}(w)]^m}\frac{1}{w^{n+1}}dw, \\ m,n \geq 0, \end{aligned} \tag{2.5}$$

where $\tilde{z}(w)$ is given below (2.2)*, and*

$$p(L,0) = 1 - \rho_1 - \rho_2. \tag{2.6}$$

The contour C is a small loop about $w = 0$.

Next we investigate the marginal probabilities defined by

$$\begin{aligned} \overline{p}(m) &= \sum_{n=0}^{\infty} p(L+m,n), \\ p(n) &= \sum_{m=0}^{\infty} p(L+m,n). \end{aligned}$$

To obtain $\overline{p}(m)$, we shift the contour C in (2.5) to $C' : |w| = 1+\varepsilon, \varepsilon > 0$. We choose ε sufficiently small so that the only pole inside C' is at $w = 1$. Then we may sum (2.5) over n, yielding

$$\begin{aligned}\overline{p}(m) &= \frac{(1-\rho_1-\rho_2)}{2\pi i}\int_{C'} \frac{1}{\rho_1 w\tilde{z}(w)-1}(\tilde{z}(w))^{-m}dw \\ &= (1-\rho_1-\rho_2)\frac{1}{\rho_1\tilde{z}(1)+\rho_1\tilde{z}'(1)}(\tilde{z}(1))^{-m}.\end{aligned}$$

Recalling that $\tilde{z}(1) = \frac{1}{\rho_1}$ and $\tilde{z}'(1) = \frac{-\rho_2}{\rho_1(1-\rho_1)}$, we obtain

$$\overline{p}(m) = (1-\rho_1)\rho_1^m. \tag{2.7}$$

Noting that $\tilde{z}(w) > 1$ for w small, we obtain the other marginal as

$$p(n) = \frac{(1-\rho_1-\rho_2)}{2\pi i}\int_C \frac{1-w}{1-\rho_1 w\tilde{z}(w)}\left(\frac{\tilde{z}(w)}{\tilde{z}(w)-1}\right)\frac{1}{w^{n+1}}dw. \tag{2.8}$$

2.2. Tail behavior. We compute the probabilities of having large queue lengths by evaluating $p(L+m, n)$ asymptotically, for m and/or n large. For a fixed $\rho_1+\rho_2 < 1$ and m or n large, we shall show that the joint probability is exponentially small. However, having accurate estimates of the tail of the distribution may be important to analyzing system performance, as tail behavior may be used to estimate loss rates, overflow probabilities, etc. In addition, the asymptotic formulas we obtain are much simpler than the exact result (2.5) and thus yield qualitative insights into the structure of the joint distribution.

We estimate (2.5) by using the saddle point and related methods for the asymptotic evaluation of integrals [11]. By writing $\tilde{z}^{-m}w^{-n} = \exp[-m\log\tilde{z} - n\log w]$, the integral in (2.5) has the form $\int_C G(w)e^{mF(w;\alpha)}dw$, where $\alpha = n/m$ and $F = -\log\tilde{z}(w) - \alpha\log w$. This is the standard form for applying the saddle point method, which yields the asymptotic behavior of the integral as $m \to \infty$, with α fixed. However, we will show that different results are obtained for different ranges of $\alpha = n/m$ and also for different ranges of the parameters ρ_1 and ρ_2.

The saddle point equation is

$$\frac{d}{dw}\left[-\log\tilde{z}(w) - \frac{n}{m}\log w\right] = 0$$

or

$$\frac{n}{m} = -\frac{\tilde{z}'(w)w}{\tilde{z}(w)}. \tag{2.9}$$

We recall that $\lambda_1\tilde{z} + \mu/\tilde{z} + \lambda_2 w = \mu + \lambda_1 + \lambda_2$, with which (2.9) simplifies to

$$\sqrt{(\mu+\lambda_1+\lambda_2-\lambda_2 w)^2 - 4\lambda_1\mu} = \frac{m}{n}w\lambda_2,$$

whose solution is

$$w = w_0\left(\frac{m}{n}\right) \equiv \frac{1}{\lambda_2}\frac{(\mu+\lambda_1+\lambda_2)^2 - 4\lambda_1\mu}{\sqrt{(\mu+\lambda_1+\lambda_2)^2\dfrac{m^2}{n^2} - 4\lambda_1\mu\left(\dfrac{m^2}{n^2}-1\right)} + \mu + \lambda_1 + \lambda_2}.$$

We then define

$$z_0\left(\frac{m}{n}\right) = \tilde{z}(w_0) = \frac{1}{2\lambda_1}\left[\left(\frac{m}{n}-1\right)\lambda_2 w_0 + \mu + \lambda_1 + \lambda_2\right].$$

At the saddle point, we have

$$\begin{aligned} m\frac{d^2}{dw^2}\left[-\log\tilde{z}(w) - \frac{n}{m}\log w\right]|_{w=w_0} \\ = \frac{n}{w_0^2} + m\left(\frac{\tilde{z}'(w_0)}{\tilde{z}(w_0)}\right)^2 - m\frac{\tilde{z}''(w_0)}{\tilde{z}(w_0)} \\ = w_0^{-2}n\left[1 + \frac{n}{m} + \frac{2\mu}{\lambda_2}\frac{n^2}{m^2}\frac{1}{z_0 w_0}\right] \equiv w_0^{-2}\mathcal{D}. \end{aligned}$$

Here we have used the quadratic equation satisfied by $\tilde{z}(w)$, and its derivatives. Since $\mathcal{D} > 0$, the steepest decent directions at the saddle are $\arg(w - w_0) = \pm\pi/2$, which correspond to a steepest descent contour that is locally parallel to the imaginary axis in the complex w-plane.

In addition to the saddle point, the asymptotic expansion of (2.5) may depend on the singularities of the integrand. Setting

$$G(w) = \frac{1-\rho_1-\rho_2}{2\pi i}\left(\frac{1-w}{1-\rho_1 w\tilde{z}(w)}\right)\frac{1}{w},$$

we see that $G(w)$ has branch points at

$$w = W_\pm = \frac{1}{\lambda_2}\left[\mu + \lambda_1 + \lambda_2 \pm 2\sqrt{\lambda_1\mu}\right] = 1 + \frac{(\sqrt{\mu} \pm \sqrt{\lambda_1})^2}{\lambda_2}. \tag{2.10}$$

Furthermore, the equation $1 = \rho_1 w\tilde{z}(w)$ has the solution

$$w = w_* \equiv \frac{\mu}{\lambda_1+\lambda_2} = \frac{1}{\rho_1+\rho_2}$$

if $\mu\lambda_1 < (\lambda_1+\lambda_2)^2$, and thus our integrand has a pole at w_* for this parameter range. In Figure 2.1 we sketch the parameter range in the (ρ_1, ρ_2) plane where the queue is stable, which is the triangle $T = \{(\rho_1,\rho_2) :\ \rho_1 \geq 0,\ \rho_2 \geq 0,\ \rho_1+\rho_2 < 1\}$. We divide this into two subregions, according to whether the pole at w_* is present or absent. Hence, we set

$$\begin{aligned} R_A &= \{(\rho_1,\rho_2) :\ 0 \leq \rho_1 < 1,\ 0 \leq \rho_2 < \sqrt{\rho_1}(1-\sqrt{\rho_1})\}, \\ R_B &= \{(\rho_1,\rho_2) :\ 0 \leq \rho_1 < 1,\ \sqrt{\rho_1}(1-\sqrt{\rho_1}) < \rho_2 < 1\} \end{aligned}$$

and note that the curve separating R_A from R_B is precisely $\mu\lambda_1 = (\lambda_1+\lambda_2)^2$ or $\rho_1 = (\rho_1+\rho_2)^2$ (see Figure 2.1).

We first consider region R_A. For $m \to \infty$ and the ratio m/n fixed, we shift the contour into another circle about the origin, which passes through $w = w_0$. Since $w_0 < W_-$ and there is no pole at w_*, such a contour deformation is permissible. Furthermore, the new contour traverses the saddle point in the steepest descent direction, and then the standard estimate (see [11]) yields

$$\begin{aligned} \int_C G(w)e^{mF(w;\alpha)}dw &\sim G(w_0)i\sqrt{\frac{2\pi}{mF''(w_0;\alpha)}}e^{mF(w_0;\alpha)} \\ &= \frac{(1-\rho_1-\rho_2)(1-w_0)}{1-\rho_1 w_0 z_0}\frac{w_0^{-n}z_0^{-m}}{\sqrt{2\pi\mathcal{D}}}. \end{aligned}$$

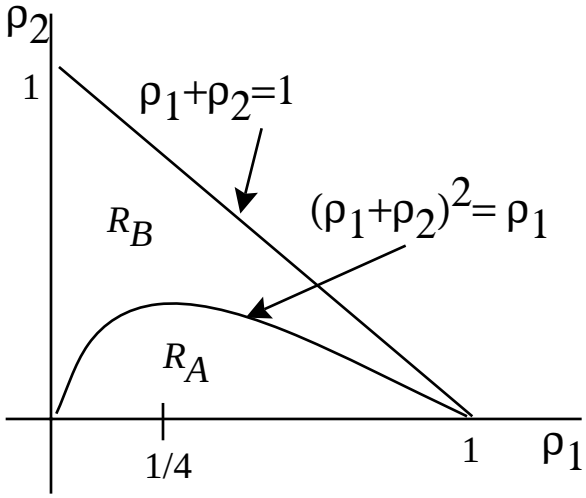


FIG. 2.1. *A sketch of the domains R_A and R_B in the (ρ_1, ρ_2) parameter space.*

This approximation is no longer valid as $m/n \to 0$, for then the saddle point w_0 approaches the branch point at W_-. It also becomes invalid as $m/n \to \infty$, as then $w_0 \to 0$, which is also a singularity of the integrand.

To construct expansions appropriate for m/n large or small, we first consider the limit $m = \mathcal{O}(1)$ as $n \to \infty$. Now the expansion of the integral in (2.5) is determined by the singularity that is closest to the origin in the w-plane, and this is the branch point at $w = W_-$ (cf. (2.10)). We approximate the integrand near W_- using

$$\tilde{z}(w) = \frac{1}{2\lambda_1}\left[2\sqrt{\mu\lambda_1} + \lambda_2\sqrt{W_+ - W_-}\sqrt{W_- - w} + \mathcal{O}(W_- - w)\right],$$

$$[\tilde{z}(w)]^{-m} = \rho_1^{m/2}\left[1 - \sqrt{\frac{\rho_2}{\rho_1}}\rho_1^{1/4}\sqrt{W_- - w}\; m + \mathcal{O}(W_- - w)\right],$$

$$\frac{1-w}{1-\rho_1 w\tilde{z}(w)} = \frac{1-W_-}{1-\sqrt{\rho_1}W_-}\left[1 + \frac{\sqrt{\rho_2}\rho_1^{1/4}W_-}{1-\sqrt{\rho_1}W_-}\sqrt{W_- - w} + \mathcal{O}(W_- - w)\right].$$

Here we have used $W_+ - W_- = 4\sqrt{\lambda_1\mu}/\lambda_2$. Using the above, we obtain the leading order term in the expansion of (2.5) as

$$\frac{1-W_-}{1-\sqrt{\rho_1}W_-}(1-\rho_1-\rho_2)\rho_1^{m/2}\left[\frac{\sqrt{\rho_2}\rho_1^{1/4}W_-}{1-\sqrt{\rho_1}W_-} - m\sqrt{\frac{\rho_2}{\rho_1}}\rho_1^{1/4}\right] I, \tag{2.11}$$

where

$$I = \frac{1}{2\pi i}\int_\Gamma w^{-n-1}\sqrt{W_- - w}\,dw$$

and the contour Γ is sketched in Figure 2.2. Setting $w = W_-(1+\zeta/n)$ and parame-

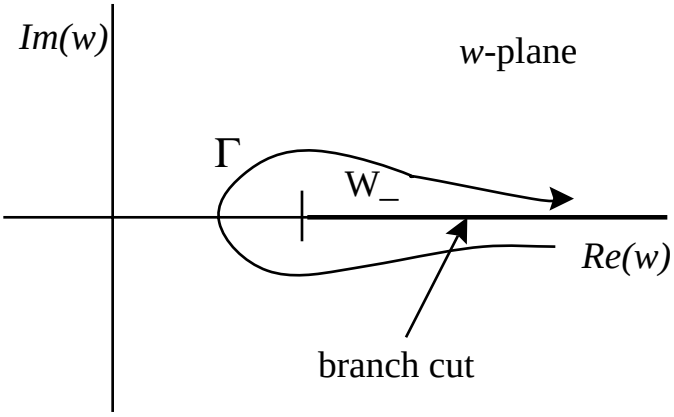


FIG. 2.2. *A sketch of the contour* Γ *in the complex* w*-plane.*

terizing Γ as a real integral, we are led to

$$I \sim -\frac{1}{\pi}\frac{\sqrt{W_-}}{n^{3/2}}W_-^{-n}\int_0^\infty e^{-\zeta}\sqrt{\zeta}d\zeta.$$

The last integral is equal to $\sqrt{\pi}/2$, and hence (2.11) gives the leading term for $n \to \infty$ with m fixed.

Now consider the limit $m \to \infty$ with n fixed. The major contribution comes from small values of w. We scale $w = u/m$ and use

$$[\tilde{z}(w)]^{-m} = [\tilde{z}(0) + \tilde{z}'(0)w + \cdots]^{-m} \sim [\tilde{z}(0)]^{-m}\exp\left[-\frac{\tilde{z}'(0)}{\tilde{z}(0)}mw\right].$$

The integral (2.5) thus becomes (apart from the factor $[\tilde{z}(0)]^{-m}$)

$$\frac{1-\rho_1-\rho_2}{2\pi i}\int_C m^n \exp\left[-\frac{\tilde{z}'(0)}{\tilde{z}(0)}u\right]u^{-n-1}du = \frac{1-\rho_1-\rho_2}{n!}m^n\left(\frac{-\tilde{z}'(0)}{\tilde{z}(0)}\right)^n,$$

where

$$\frac{-\tilde{z}'(0)}{\tilde{z}(0)} = \frac{\lambda_2}{\sqrt{(\mu+\lambda_1+\lambda_2)^2-4\lambda_1\mu}} > 0.$$

We summarize our main results below.

THEOREM 2.2. *For* $(\rho_1,\rho_2) \in R_A$*, the steady state probabilities in* (2.5) *have the asymptotic expansions*

(a) $m,\ n \to \infty$ *with* $0 < m/n < \infty$*:*

$$p(L+m,n) \sim \frac{1-\rho_1-\rho_2}{1-\rho_1 w_0 z_0}\left(\frac{1-w_0}{\sqrt{2\pi\mathcal{D}}}\right)w_0^{-n}z_0^{-m},$$

$$w_0 = w_0\left(\frac{m}{n}\right) = \frac{1}{\lambda_2}\frac{(\mu+\lambda_1+\lambda_2)^2-4\lambda_1\mu}{\sqrt{(\mu+\lambda_1+\lambda_2)^2\dfrac{m^2}{n^2}-4\lambda_1\mu\left(\dfrac{m^2}{n^2}-1\right)}+\mu+\lambda_1+\lambda_2},$$

$$z_0 = z_0\left(\frac{m}{n}\right) = \frac{1}{2\lambda_1}\left[\left(\frac{m}{n} - 1\right)\lambda_2 w_0 + \mu + \lambda_1 + \lambda_2\right],$$

$$\mathcal{D} = n\left[1 + \frac{n}{m} + \frac{2\mu}{\lambda_2}\frac{n^2}{m^2}\frac{1}{z_0 w_0}\right];$$

(b) $n \to \infty$ *with* $m = \mathcal{O}(1)$*:*

$$p(L+m,n) \sim \frac{1-\rho_1-\rho_2}{2\sqrt{\pi}}\left(\frac{W_- - 1}{\sqrt{\rho_1}W_- - 1}\right)\frac{\sqrt{\rho_2}}{\rho_1^{1/4}}\left[m + \frac{\sqrt{\rho_1}W_-}{\sqrt{\rho_1}W_- - 1}\right]\frac{\rho_1^{m/2}}{n^{3/2}}\sqrt{W_-}W_-^{-n},$$

$$W_- = \frac{1}{\lambda_2}\left[\mu + \lambda_1 + \lambda_2 - 2\sqrt{\lambda_1\mu}\right];$$

(c) $m \to \infty$ *with* $n = \mathcal{O}(1)$*:*

$$p(L+m,n) \sim (1-\rho_1-\rho_2)\frac{m^n}{n!}[\tilde{z}(0)]^{-m}\left(\frac{\lambda_2}{\sqrt{S}}\right)^n,$$

$$\tilde{z}(0) = \frac{\mu + \lambda_1 + \lambda_2 + \sqrt{S}}{2\lambda_1}, \qquad S = (\mu + \lambda_1 + \lambda_2)^2 - 4\lambda_1\mu.$$

We note that (a) gives the probability of having simultaneously large queue lengths for both customer classes. Part (b) gives the probability of having many type-2 (e.g., voice) customers, while part (c) gives the probability of having many type-1 (e.g., data) customers above the threshold L. We can easily show that $\sqrt{\rho_1}W_- > 1$ precisely for $(\rho_1, \rho_2) \in R_A$, so that the formula in (b) is positive for all $m \geq 0$, as it must be.

We next consider the region R_B. Now the integrand has a pole at w_* in addition to the saddle at w_0, whose location changes with m/n. As m/n increases from zero to infinity, the saddle moves from W_- to 0 and coalesces with the pole when

$$\frac{m}{n} = \beta \equiv \frac{(\lambda_1+\lambda_2)^2 - \mu\lambda_1}{\mu\lambda_2} = \frac{(\rho_1+\rho_2)^2 - \rho_1}{\rho_2}.$$

Note that $\beta > 0$ precisely when $(\rho_1, \rho_2) \in R_B$. For $m/n < \beta$, we have $w_* < w_0$, while if $m/n > \beta$, we have $w_* > w_0$. For $m \to \infty$ and $n = \mathcal{O}(1)$, we again obtain the result in part (c) of Theorem 2.2. For $m, n \to \infty$ with $m/n \in (\beta, \infty)$, we shift the contour C into the steepest descent contour and obtain the same result as in Theorem 2.2(a). If $m, n \to \infty$ and $m/n \in (0, \beta)$, we must take into account the contribution from the pole at w_* in deforming the contour. The residue at this pole is equal to

$$(1-\rho_1-\rho_2)\frac{(1-w_*)}{-\rho_1(w\tilde{z})'(w_*)}[\tilde{z}(w_*)]^{-m}(\rho_1+\rho_2)^{n+1}. \tag{2.12}$$

From the definition of $\tilde{z}$ we easily obtain

$$\tilde{z}(w_*) = 1 + \frac{\lambda_2}{\lambda_1}, \qquad \tilde{z}'(w_*) = \frac{\lambda_2(1+\lambda_2/\lambda_1)^2}{\mu - \lambda_1(1+\lambda_2/\lambda_1)^2}.$$

By comparing (2.12) to the saddle point contribution, we can show that the contribution from the pole is larger. Thus, for $m/n \in (0, \beta)$ and $m \to \infty$, $p(L+m, n)$ is asymptotically given by the negative of (2.12). The negative sign arises due to the

fact that the saddle point contour ($|w| = w_0$) has a radius larger than the original contour C.

Finally, we consider the limit $m, n \to \infty$ with $m/n \approx \beta$. Note that the saddle point approximation, as given by Theorem 2.2(a) for $m/n > \beta$, becomes singular as $m/n \downarrow \beta$, and $1 - \rho_1 w_0 z_0$ vanishes in this limit. The case of a saddle coalescing with a pole is a standard problem in the asymptotic evaluation of integrals (see [11]). The appropriate expansion for $m/n \approx \beta$ may be obtained by expanding the integrand near $w = w_*$. In this limit we have

$$G(w) \sim \frac{1-\rho_1-\rho_2}{2\pi i}\left(\frac{1-w_*}{-\rho_1(w\tilde{z})'(w_*)}\right)\frac{1}{w-w_*},$$

$$\begin{aligned} n\log w \,+\, m\log[\tilde{z}(w)] &= n\log w_* + m\log[\tilde{z}(w_*)] + \left(\frac{n}{w_*} + m\frac{\tilde{z}'(w_*)}{\tilde{z}(w_*)}\right)(w-w_*) \\ &+\, \frac{1}{2}\left[\frac{-n}{w_*^2} + m\frac{\tilde{z}''(w_*)}{\tilde{z}(w_*)} - m\left(\frac{\tilde{z}'(w_*)}{\tilde{z}(w_*)}\right)^2\right](w-w_*)^2 + \mathcal{O}((w-w_*)^3) \\ &= -n\log(\rho_1+\rho_2) + m\log\left(1+\frac{\rho_2}{\rho_1}\right) + (\rho_1+\rho_2)\left(n - \frac{m}{\beta}\right)(w-w_*) \\ &-\, \frac{1}{2}(\rho_1+\rho_2)^2\left[n + \frac{2m\rho_1}{\rho_2\beta^3} + \frac{m}{\beta^2}\right](w-w_*)^2 + \mathcal{O}((w-w_*)^3). \end{aligned}$$

We set $w - w_* = (\rho_1+\rho_2)^{-1}[n + m/\beta^2 + 2m\rho_1/(\rho_2\beta^3)]^{-1/2}\eta$ and obtain

$$\int_C G(w)e^{mF(w;\alpha)}dw \sim (1-\rho_1-\rho_2)\frac{-\mu\lambda_2\beta}{(\lambda_1+\lambda_2)^2}(\rho_1+\rho_2)^n\left(\frac{\rho_1}{\rho_1+\rho_2}\right)^m J,$$

where J is the integral

$$J = \frac{1}{2\pi i}\int_{C'} e^{\Delta_0\eta}e^{\eta^2/2}\frac{d\eta}{\eta}, \qquad \Delta_0 = \frac{m/\beta - n}{\sqrt{n + m/\beta^2 + 2m\rho_1/(\rho_2\beta^3)}},$$

and we assume that $m/\beta - n = \mathcal{O}(\sqrt{m})$ so that $\Delta_0 = \mathcal{O}(1)$. Here C' is a vertical contour in the η-plane such that $\Re(\eta) < 0$. The last integral is easily evaluated as $J = -(2\pi)^{-1/2}\int_{\Delta_0}^{\infty} e^{-u^2/2}du$, and we thus have the desired expression for $p(L+m,n)$ that is valid for $m, n \to \infty$ with $m/n \approx \beta$. Below we summarize the results.

THEOREM 2.3. *For $(\rho_1, \rho_2) \in R_B$, the steady state probabilities in* (2.5) *have the asymptotic expansions*

(a) *$n \to \infty$ with $0 \le m/n < \beta$, $\beta = [(\rho_1+\rho_2)^2 - \rho_1]/\rho_2$:*

$$p(L+m,n) \sim (1-\rho_1-\rho_2)\frac{(\lambda_1+\lambda_2)^2 - \mu\lambda_1}{(\lambda_1+\lambda_2)^2}\left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^m(\rho_1+\rho_2)^n;$$

(b) *$m, n \to \infty$ with $m/n = \beta + \mathcal{O}(m^{-1/2})$:*

$$\begin{aligned} &p(L+m,n) \\ &\sim (1-\rho_1-\rho_2)\frac{(\lambda_1+\lambda_2)^2 - \mu\lambda_1}{(\lambda_1+\lambda_2)^2}\left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^m(\rho_1+\rho_2)^n\frac{1}{\sqrt{2\pi}}\int_{\Delta}^{\infty} e^{-u^2/2}du, \end{aligned}$$

$$\Delta = \left(\frac{m-n\beta}{\sqrt{n}}\right)\frac{1}{\sqrt{\beta^2+\beta+2\rho_1/\rho_2}};$$

(c) $m, n \to \infty$ *with* $\beta < m/n < \infty$*: the result is the same as Theorem* 2.2(a)*; for* $m \to \infty$ *with* $n = \mathcal{O}(1)$*, Theorem* 2.2(c) *applies.*

The expression in (a) shows that the solution is asymptotically of "product form" for this range of m, n. We have thus identified two regions in the parameter (ρ_1, ρ_2) space where the structure of the joint distribution is different. Theorems 2.2 and 2.3 furthermore show that the expansion of $p(L + m, n)$ depends on the ratio m/n for both of the regions in Figure 2.1.

We next consider the asymptotic behavior of the marginal distributions $\overline{p}(m)$ and $p(n)$. The former has a simple geometric form, but the latter is complicated, and thus we evaluate it asymptotically for $n \to \infty$. We also note that for either region R_A or R_B, the saddle point approximation (i.e., the formula in Theorem 2.2(a)) can be simplified for $m/n \approx (1 - \rho_1)/\rho_2$. To do so, we first compute the point at which $p(L + m, n)$ is maximal as a function of n, for a fixed large m. We write

$$w_0^{-n} z_0^{-m} = \exp\left(-m \log[z_0(x)] - \frac{m}{x} \log[w_0(x)]\right), \qquad x = \frac{m}{n}.$$

At the maximum we must have

$$\frac{z_0'(x)}{z_0(x)} + \frac{1}{x}\frac{w_0'(x)}{w_0(x)} - \frac{1}{x^2} \log[w_0(x)] = 0,$$

and we can easily show that this is satisfied if

$$x = \beta_* \equiv \frac{1 - \rho_1}{\rho_2}.$$

We then have $w_0(\beta_*) = 1$ and $z_0(\beta_*) = 1/\rho_1$. After a lengthy calculation, we find that

$$\frac{d^2}{dx^2}\left[\log[z_0(x)] + \frac{1}{x}\log[w_0(x)]\right]\Big|_{x=\beta_*} = \frac{\rho_2^3}{1-\rho_1}\frac{1}{(1-\rho_1)^2 + \rho_2(1+\rho_1)}$$

and

$$\frac{1-\rho_1-\rho_2}{1-\rho_1 w_0(x) z_0(x)}\frac{1-w_0(x)}{\sqrt{2\pi\mathcal{D}}}\Big|_{x=\beta_*} = \frac{1-\rho_1}{\sqrt{2\pi m}}\frac{(1-\rho_1)^{3/2}}{\sqrt{\rho_2}\sqrt{(1-\rho_1)^2 + \rho_2(1+\rho_1)}}.$$

Using these expressions in Theorem 2.2(a), we obtain a simple form for $p(L+m, n)$ for $m, n \to \infty$ with $m/n \approx \beta_*$, which we summarize below. We also give the expansions for the marginal $p(n)$ as $n \to \infty$. These are easily obtained from (2.8) by identifying the singularity of the integrand closest to the origin, and this is $w = W_-$ for R_A and $w = w_*$ for R_B.

THEOREM 2.4. *Further asymptotic properties of the distribution are as follows:*

(a) $m, n \to \infty$ *with* $m/n = \beta_* + \mathcal{O}(m^{-1/2})$*,* $\beta_* = (1 - \rho_1)/\rho_2$*:*

$$p(L+m, n) \sim (1-\rho_1)\rho_1^m \frac{1}{\sqrt{2\pi m}} \frac{(1-\rho_1)^{3/2}}{\sqrt{\rho_2}\sqrt{(1-\rho_1)^2 + \rho_2(1+\rho_1)}} \times \exp\left[-\frac{1}{2m}\frac{(1-\rho_1)^3}{\rho_2[(1-\rho_1)^2 + \rho_2(1+\rho_1)]}\left(n - \frac{\rho_2 m}{1-\rho_1}\right)^2\right];$$

(b) $n \to \infty$ *with* $(\rho_1, \rho_2) \in R_A$*:*

$$p(n) \sim \frac{(1-W_-)^2}{(1-\sqrt{\rho_1}W_-)^2}\frac{(1-\rho_1-\rho_2)\sqrt{W_-}}{(1-\sqrt{\rho_1})^2}\frac{\sqrt{\rho_2}\rho_1^{1/4}}{2\sqrt{\pi}}\frac{W_-^{-n}}{n^{3/2}};$$

(c) $n \to \infty$ *with* $(\rho_1, \rho_2) \in R_B$*:*

$$p(n) \sim \frac{1-\rho_1-\rho_2}{\rho_2(\rho_1+\rho_2)}[(\rho_1+\rho_2)^2-\rho_1](\rho_1+\rho_2)^n.$$

We note that $\beta_* > \beta$ for all $\rho_1+\rho_2 < 1$. Thus when $m/n \approx \beta_*$, we are always in the case where the leading term in the expansion of $p(L+m,n)$ comes from the saddle point. Part (a) shows that if the number of type-1 customers above L is large and $= m$, then the number of type-2 customers is also likely to be large and close to $\rho_2 m/(1-\rho_1)$. Furthermore, there is a Gaussian spread about this mean value. From part (a) we also conclude that, for $m \to \infty$, $\sum_{n=0}^{\infty} p(L+m,n) \sim (1-\rho_1)\rho_1^m$. This result is of course not just asymptotic, but exact for all $m \geq 0$. This completes our analysis of the tail behavior of the joint distribution and its marginals.

2.3. Heavy traffic diffusion approximation. We study the system when it becomes close to unstable, i.e., as $\rho_1+\rho_2 \uparrow 1$. If $\rho_1+\rho_2 \uparrow 1$ for a fixed $\rho_2 \in (0,1)$, the probability mass becomes concentrated in the range where n is large but $m = \mathcal{O}(1)$. More precisely, we define $\varepsilon = 1-\rho_1-\rho_2 \to 0^+$ and obtain the limiting distribution from (2.5) as

$$p(L+m,n) \sim \varepsilon(1-\rho_1)\rho_1^m e^{-Y}, \qquad Y = \varepsilon n.$$

Thus, in this limit there tend to be only a few type-1 customers above the threshold L, while the number of type-2 customers tends to be large, of the order $\mathcal{O}(\varepsilon^{-1})$. Since the latter customers may represent voice messages, this situation is clearly not desirable. We next obtain another heavy traffic limit, whose behavior is much less trivial than that above.

We again define the small positive parameter ε by $1 = \rho_1+\rho_2+\varepsilon$, but we now assume that

$$\rho_2 = \varepsilon b = \mathcal{O}(\varepsilon), \qquad 1-\rho_1 = \varepsilon(b+1) = \mathcal{O}(\varepsilon),$$

where $b > 0$. This assumption means that the traffic intensity of type-2 customers is relatively small, while the type-1 customer queue is close to instability. Even though ρ_2 is small, with this scaling both of the queue lengths will tend to be large, so that it seems appropriate to still classify this limit as one of "heavy traffic." Even though the load of type-2 customers is small, large queue lengths tend to develop since the server is devoting a lot of time trying to service the large backlog of type-1 customers. We thus scale m and n to be large, with

$$m = \frac{X}{\varepsilon}, \qquad n = \frac{Y}{\varepsilon}.$$

In (2.5) we scale $w = 1-\varepsilon s$, and for $\varepsilon \to 0^+$ we have

$$\begin{aligned}
\tilde{z}(w) - 1 &\sim \frac{\varepsilon}{2}\left[b+1+\sqrt{(b+1)^2+4bs}\right], \\
w^{-n-1} &\sim e^{sY}, \\
[\tilde{z}(w)]^{-m} &\sim \exp\left\{-\frac{X}{2}\left[b+1+\sqrt{(b+1)^2+4bs}\right]\right\},
\end{aligned}$$

with $p(L,0) = \varepsilon$. We can also approximate the contour C by a vertical Bromwich contour Br in the complex s-plane. We thus obtain the limiting density in the form

$$p(L+m,n) = p\left(L+\frac{X}{\varepsilon}, \frac{Y}{\varepsilon}\right) \sim \varepsilon^2 P(X,Y), \tag{2.13}$$

where

$$P(X,Y)=\frac{1}{2\pi i}\int_{Br}\frac{2se^{sY}\exp\left\{-\frac{X}{2}\left[b+1+\sqrt{(b+1)^2+4bs}\right]\right\}}{b+1+2s-\sqrt{(b+1)^2+4bs}}ds \tag{2.14}$$

and, on Br, $\Re(s)>0$. We refer to P as the heavy traffic diffusion approximation.

An alternate approach to obtaining P is to introduce the heavy traffic scaling into the difference equations. From these we find that P in (2.13) satisfies the parabolic partial differential equation (PDE)

$$P_{XX}+(b+1)P_X-bP_Y=0,\qquad X,Y>0,$$

with boundary conditions (BC)

$$\begin{gathered}P_X(0,Y)+P_Y(0,Y)+(b+1)P(0,Y)=0,\qquad Y>0,\\ P(X,0)=0,\qquad X>0.\end{gathered}$$

Solving this problem using, e.g., Laplace transforms, we regain the expression in (2.14). The approximation (2.13) may be refined to the series $p(L+m,n)=\varepsilon^2[P(X,Y)+\varepsilon P^{(1)}(X,Y)+\varepsilon^2P^{(2)}(X,Y)+\cdots]$, but the calculation of the higher order terms $P^{(j)}$ is tedious. Below we summarize the final results for the diffusion model and its marginals and also give several alternate representations for P.

THEOREM 2.5. *In the heavy traffic limit where $\rho_1+\rho_2=1-\varepsilon$ and $\rho_2=\varepsilon b$, $b>0$, we have for $\varepsilon\to0^+$, $p(L+m,n)\sim\varepsilon^2P(X,Y)$, where*

$$\begin{aligned}P(X,Y)&=\frac{1}{2\pi i}\int_{Br}\frac{2se^{sY}\exp\left\{-\frac{X}{2}\left[b+1+\sqrt{(b+1)^2+4bs}\right]\right\}}{b+1+2s-\sqrt{(b+1)^2+4bs}}ds\\
&=(b-1)e^{-bX}e^{-Y}I\{b>1\}+e^{-(b+1)X/2}\frac{(b+1)^3}{4\pi}i\\
&\quad\times\int_{-\infty}^{\infty}\frac{\sinh\eta}{4b-(b+1)^2\cosh\eta}\left[1-\frac{b+1}{2b}\cosh\eta+i\sqrt{2}\sinh\left(\frac{\eta}{2}\right)\right]\\
&\quad\times\exp\left[-\frac{(b+1)^2}{4b}Y\cosh\eta-\frac{X}{\sqrt{2}}(b+1)i\sinh\left(\frac{\eta}{2}\right)\right]d\eta\\
&=(b-1)e^{-bX}e^{-Y}I\{b>1\}\\
&\quad+\frac{e^{-(b+1)X/2}e^{-Y}}{2\sqrt{\pi b}}\int_Y^{\infty}\frac{A(X,u)}{\sqrt{u}}\exp\left[-\frac{bX^2}{4u}-\frac{(b-1)^2}{4b}u\right]du,\end{aligned}$$

$$A(X,u)=\frac{b}{u}\left(\frac{3X}{2u}+1\right)-\left(\frac{b^2-1}{4}\frac{X}{u}+\frac{b^2X^2}{2u^2}+\frac{b^2X^3}{4u^3}\right).$$

If $b=1$, the above simplifies to

$$P(X,Y)=\frac{1}{\sqrt{\pi Y}}\left(1+\frac{X}{2Y}\right)e^{-X}e^{-Y}\exp\left(-\frac{X^2}{4Y}\right),\qquad b=1.$$

The marginals are given by

$$\overline{P}(X) = \int_0^\infty P(X,Y)dY = (b+1)e^{-(b+1)X},$$
$$P(Y) = \int_0^\infty P(X,Y)dX$$
$$= \left[\frac{b-1}{b}I\{b>1\} + \frac{1}{2\sqrt{\pi b}}\int_Y^\infty u^{-3/2}\exp\left[-\frac{(b-1)^2}{4b}u\right]du\right]e^{-Y},$$

where $I\{\cdot\}$ is the indicator function. When $b = 1$, the Y-marginal is $P(Y) = e^{-Y}/\sqrt{\pi Y}$.

From the result for P, we see that for any $X > 0$ the conditional density $P(Y|X) = P(X,Y)/\overline{P}(X)$ is continuous. However, when $X = 0$, the conditional density carries mass at the origin, and we have

$$P(Y|0) = \frac{1}{b+1}\delta(Y)$$
$$+ \frac{b}{b+1}e^{-Y}\left[\frac{b-1}{b}I\{b>1\} + \frac{1}{2\sqrt{\pi b}}\int_Y^\infty u^{-3/2}\exp\left(-\frac{(b-1)^2}{4b}u\right)du\right].$$

Thus the mass at $Y = 0$ is $1/(b+1)$, and the mass for $Y > 0$ is $b/(b+1)$. We can also study the tail behavior of the diffusion approximation as X and/or $Y \to \infty$. The qualitative structure of $P(X,Y)$ in this limit is similar to that in Theorems 2.2–2.4. Below we summarize the results, omitting the derivations.

THEOREM 2.6. *Asymptotic expansions of $P(X,Y)$ and $P(Y)$ in Theorem* 2.5 *are as follows:*

(a) $Y \to \infty$*:*

$$b>1: \quad P(Y) \sim \frac{b-1}{b}e^{-Y},$$
$$b=1: \quad P(Y) \sim \frac{e^{-Y}}{\sqrt{\pi Y}},$$
$$b>1: \quad P(Y) \sim \frac{2\sqrt{b}}{\sqrt{\pi}(b-1)^2}Y^{-3/2}\exp\left[-\frac{(b+1)^2}{4b}Y\right];$$

(b) $X, Y \to \infty$ *with* $Y - bX/(b+1) = \sqrt{X}U = \mathcal{O}(\sqrt{X})$*:*

$$P\left(\frac{bX}{b+1} + \sqrt{X}U|X\right) \sim \frac{(b+1)^{3/2}}{2\sqrt{\pi X}b}\exp\left[-\frac{(b+1)^3}{4b^2}U^2\right];$$

(c) $Y \to \infty$, $X = \mathcal{O}(1)$, $b < 1$*:*

$$P(X,Y) \sim \frac{\sqrt{b}}{2\sqrt{\pi}}\left(\frac{b+1}{1-b}\right)Y^{-3/2}\left[X + \frac{4b}{1-b^2}\right]\exp\left[-\frac{X}{2}(b+1) - \frac{Y}{4b}(b+1)^2\right];$$

(d) $X, Y \to \infty$ *with* $0 < Y/X < \infty$, $b < 1$*:*

$$P(X,Y) \sim K(X,Y)e^{-\phi(X,Y)},$$
$$\phi(X,Y) = \frac{X}{2}(b+1) + \frac{Y}{4b}(b+1)^2 + \frac{bX^2}{4Y},$$
$$K(X,Y) = \frac{\sqrt{b}X}{2\sqrt{\pi}Y^{3/2}}\left(\frac{bX+(b+1)Y}{bX-(b-1)Y}\right);$$

(e) $X, Y \to \infty$ *with* $0 < Y/X < \infty$, $b > 1$*:*

$$Y/X > \frac{b}{(b-1)}: \quad P(X,Y) \sim (b-1)e^{-bX}e^{-Y},$$

$$Y/X < \frac{b}{(b-1)}: \quad P(X,Y) \sim K(X,Y)e^{-\phi(X,Y)},$$

$$Y/X \approx \frac{b}{(b-1)}: \quad P(X,Y) \sim (b-1)e^{-bX}e^{-Y}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\Lambda} e^{-u^2/2}du,$$

$$\Lambda = \frac{(b-1)^{3/2}}{b\sqrt{2X}}\left(Y - \frac{bX}{b-1}\right) = \mathcal{O}(1).$$

Note that these results are closely analogous to the asymptotic results for the discrete probabilities $p(L+m, n)$. Theorems 2.5 and 2.6 show that the structure of the problem is nontrivial in the heavy traffic limit where $\rho_1 + \rho_2 \uparrow 1$ and $\rho_2 \to 0^+$. We also note that this limit corresponds to the vicinity of the "corner" point $(\rho_1, \rho_2) = (1, 0)$ in the parameter space in Figure 2.1. The curve that separates R_A from R_B passes through this point. The equation $(\rho_1+\rho_2)^2 = \rho_1$ is the same as $(1-\varepsilon)^2 = 1-\varepsilon(b+1)$, which as $\varepsilon \to 0^+$ becomes $b = 1 + \mathcal{O}(\varepsilon)$. This further explains the dichotomous behavior of P, according to whether $b > 1$ or $b < 1$.

3. $M_1, M_2/M/1$ priority queue with a dynamic scheduling policy. We now consider the full problem (1.1)–(1.9) in which the server serves type-1 customers when $N_2(t) = 0$. We are able to use the results in section 2 as part of our analysis here. For example, the solution in Theorem 3.1 uses the results in Theorem 2.5, and the asymptotic results in Theorems 3.2(a) and 3.3(a) rely on the asymptotic results in Theorem 2.5 in section 2.

3.1. Exact solution. In this subsection, we derive the joint queue length distribution. We use the same notation as in the previous section and introduce the following probability generating functions:

$$G(z,w) = \sum_{n=0}^{\infty}\sum_{m=L}^{\infty} p(m,n)z^{m-L}w^n,$$

$$H_j(w) = \sum_{n=0}^{\infty} p(j,n)w^n, \qquad 0 \le j \le L.$$

From (1.1), (1.2), (1.4), and (1.5), we obtain

$$\begin{aligned}&\left[\lambda_1 z + \lambda_2 w - (\mu + \lambda_1 + \lambda_2) + \frac{\mu}{z}\right]G(z,w)\\ &\qquad = -\lambda_1 H_{L-1}(w) + \mu\left(\frac{1}{z} - \frac{1}{w}\right)H_L(w) + \frac{\mu}{w}p(L,0).\end{aligned} \tag{3.1}$$

In order to determine $G(z,w)$ in (3.1), we must know $H_{L-1}(w), H_L(w)$, and $p(L,0)$. We first compute $H_{L-1}(w)$ and $p(L,0)$, and then derive $H_L(w)$ by using the analyticity of $G(z,w)$.

From (1.3) and (1.6), we obtain the following equation, which relates $H_{m-1}(w)$ and $H_m(w)$:

$$\begin{aligned}&\left[\lambda_2 w - (\mu + \lambda_1 + \lambda_2) + \frac{\mu}{w}\right]H_m(w) + \lambda_1 H_{m-1}(w)\\ &\qquad = -\mu p(m+1,0) + \frac{\mu}{w}p(m,0), \qquad 1 \le m \le L-1.\end{aligned} \tag{3.2}$$

From (1.7) and (1.8), we obtain an equation for $H_0(w)$,

$$\Big[\lambda_2 w - (\mu+\lambda_1+\lambda_2) + \frac{\mu}{w}\Big] H_0(w) = -\mu p(1,0) + \mu\left(\frac{1}{w}-1\right)p(0,0), \tag{3.3}$$

which can also be written as

$$\frac{\lambda_2}{w}(w-w_-)(w-w_+)H_0(w) = -\mu p(1,0) + \mu\left(\frac{1}{w}-1\right)p(0,0), \tag{3.4}$$

where w_- and w_+ are zeros of $L(w) \equiv w^2 - \frac{\mu+\lambda_1+\lambda_2}{\lambda_2}w + \frac{\mu}{\lambda_2} = 0$ and $|w_-| < 1$; i.e.,

$$w_- = \frac{(\mu+\lambda_1+\lambda_2) - \sqrt{(\mu+\lambda_1+\lambda_2)^2 - 4\mu\lambda_2}}{2\lambda_2},$$
$$w_+ = \frac{(\mu+\lambda_1+\lambda_2) + \sqrt{(\mu+\lambda_1+\lambda_2)^2 - 4\mu\lambda_2}}{2\lambda_2}.$$

Letting $w = w_-$ in (3.4), we obtain

$$p(1,0) = \left(\frac{1}{w_-}-1\right)p(0,0), \tag{3.5}$$

and then (3.4) yields

$$H_0(w) = \frac{1}{(1-w/w_+)}p(0,0). \tag{3.6}$$

Next we will calculate $H_j(w)$ for $1 \le j \le L-1$. We first take $L = \infty$ and show how to use this solution to solve the problem for finite L. If $L = \infty$, we introduce the generating functions

$$H(z,w) = \sum_{m=0}^{\infty} H_m(w) z^m,$$
$$F(z) = \sum_{m=0}^{\infty} p(m,0) z^m.$$

Then, using (3.2) and (3.3) with $L = \infty$, we obtain

$$\Big[\lambda_2 w - (\mu+\lambda_1+\lambda_2) + \lambda_1 z + \frac{\mu}{w}\Big] H(z,w)$$
$$= \mu\Big(\frac{1}{w}-\frac{1}{z}\Big)F(z) + \mu\Big(\frac{1}{z}-1\Big)p(0,0).$$

Upon dividing by w/μ, the above may be rewritten as

$$\rho_2(w-w_-(z))(w-w_+(z))H(z,w)$$
$$= \Big(1-\frac{w}{z}\Big)F(z) + w\Big(\frac{1}{z}-1\Big)p(0,0), \tag{3.7}$$

where $w_-(z)$ and $w_+(z)$ are zeros of the equation $w^2 - \frac{\mu+\lambda_1+\lambda_2-\lambda_1 z}{\lambda_2}w + \frac{\mu}{\lambda_2} = 0$ and $|w_-(z)| < 1$; hence

$$w_-(z) = \frac{(\mu+\lambda_1+\lambda_2-\lambda_1 z) - \sqrt{(\mu+\lambda_1+\lambda_2-\lambda_1 z)^2 - 4\mu\lambda_2}}{2\lambda_2},$$
$$w_+(z) = \frac{(\mu+\lambda_1+\lambda_2-\lambda_1 z) + \sqrt{(\mu+\lambda_1+\lambda_2-\lambda_1 z)^2 - 4\mu\lambda_2}}{2\lambda_2}.$$

Setting $w = w_-(z)$ in (3.7), we obtain $F(z)$ as

$$F(z) = \frac{w_-(z)(1-1/z)}{1-w_-(z)/z} p(0,0). \tag{3.8}$$

Substituting (3.8) into (3.7), we obtain

$$H(z,w) = p(0,0)\frac{1}{\rho_2}\left(\frac{z-1}{z-w_-(z)}\right)\frac{1}{w_+(z)}\left(\frac{1}{1-w/w_+(z)}\right). \tag{3.9}$$

Finally, by comparing the coefficients of (3.9) to those of the probability generating function, we obtain $p(m,n)$ as

$$p(m,n) = \frac{p(0,0)}{2\pi i}\int_C \left(\frac{1-z}{1-\rho_2 z w_+(z)}\right)\frac{1}{[w_+(z)]^n}\frac{1}{z^{m+1}}dz \qquad (L=\infty).$$

Note that the above integral is the same as (2.5) with m and n (and also λ_1 and λ_2) interchanged.

The above expression for $p(m,n)$ satisfies (1.1)–(1.8) for all m,n with $0 \le m \le L-1$. Hence for $L < \infty$ and $0 \le m \le L-1$ we obtain

$$p(m,n) = \frac{p(0,0)}{2\pi i}\int_C \left(\frac{1-z}{1-\rho_2 z w_+(z)}\right)\frac{1}{[w_+(z)]^n}\frac{1}{z^{m+1}}dz \tag{3.10}$$

and

$$H_m(w) = \frac{p(0,0)}{2\pi i}\int_C \frac{1-z}{1-\rho_2 z w_+(z)}\left(\frac{w_+(z)}{w_+(z)-w}\right)\frac{1}{z^{m+1}}dz. \tag{3.11}$$

Using (3.10) and (3.11), we can easily show that (3.2)–(3.6) are satisfied. In (3.2) and (3.3) we let $w=1$, and we sum from $m=0$ to $L-1$ to find that $p(L,0)$ is given by

$$\begin{aligned} p(L,0) &= \rho_1 H_{L-1}(1) \\ &= \rho_1 \frac{p(0,0)}{2\pi i}\int_C \frac{1-z}{1-\rho_2 z w_+(z)}\left(\frac{w_+(z)}{w_+(z)-1}\right)\frac{1}{z^L}dz. \end{aligned} \tag{3.12}$$

We next derive the probability generating function $G(z,w)$. Equation (3.1) can be rewritten as

$$\begin{aligned} &\rho_1(z-z_*(w))(z-\tilde{z}(w))G(z,w) \\ &\quad = -\rho_1 z H_{L-1}(w) + z\left(\frac{1}{z}-\frac{1}{w}\right)H_L(w) + \frac{z}{w}p(L,0), \end{aligned} \tag{3.13}$$

where $H_{L-1}(w)$ and $p(L,0)$ are now known in terms of $p(0,0)$, and $z_*(w)$ and $\tilde{z}(w)$ are the same as in section 2. In (3.13), we set $z = z_*(w)$ and thus obtain $H_L(w)$. Then, after some calculation, we obtain the joint probability generating function as

$$G(z,w) = \frac{p(L,0)-\rho_1 w H_{L-1}(w)}{\rho_1(w-z_*(w))}\left(\frac{1}{z-\tilde{z}(w)}\right), \tag{3.14}$$

where $p(L,0)$ and $H_{L-1}(w)$ are given by (3.12) and (3.11), respectively. Finally, we obtain the joint queue length probabilities $p(m,n)$ by inverting $G(z,w)$:

$$\begin{aligned} p(m,n) = \frac{p(0,0)}{(2\pi i)^2}\rho_1 \int_C\int_C &\left(\frac{1-z}{1-\rho_2 z w_+(z)}\right)\frac{[w_+(z)]^2}{(w_+(z)-1)}\frac{(1-w)}{(w_+(z)-w)}\frac{1}{z^L} \\ &\times \frac{1}{(1-\rho_1 w\tilde{z}(w))}\frac{1}{[\tilde{z}(w)]^{m-L}}\frac{1}{w^{n+1}}dzdw, \qquad m \ge L,\ n \ge 0. \end{aligned} \tag{3.15}$$

The unknown constant $p(0,0)$ can be obtained by the normalization condition

$$G(1,1) + \sum_{m=0}^{L-1} H_m(1) = 1.$$

A routine calculation yields

$$p(0,0) = 1 - \rho_1 - \rho_2. \tag{3.16}$$

We have thus obtained the following result.

THEOREM 3.1. *The joint probabilities $p(m,n)$ for the queue lengths are given by*
(a) *for $0 \le m \le L-1, n \ge 0$,*

$$p(m,n) = \frac{p(0,0)}{2\pi i} \int_C \left(\frac{1-z}{1-\rho_2 z w_+(z)} \right) \frac{1}{[w_+(z)]^n} \frac{1}{z^{m+1}} dz,$$

(b) *for $m \ge L, n \ge 0$,*

$$p(m,n) = \frac{p(0,0)}{(2\pi i)^2} \rho_1 \int_C \int_C \left(\frac{1-z}{1-\rho_2 z w_+(z)} \right) \frac{[w_+(z)]^2}{(w_+(z)-1)} \frac{(1-w)}{(w_+(z)-w)} \frac{1}{z^L} \times \frac{1}{(1-\rho_1 w \tilde{z}(w))} \frac{1}{[\tilde{z}(w)]^{m-L}} \frac{1}{w^{n+1}} dz dw,$$

where

$$p(0,0) = 1 - \rho_1 - \rho_2,$$

and $w_+(z)$ and $\tilde{z}(w)$ are given below (3.7) *and below* (2.2)*, respectively.*

To study the tail probabilities, it is reasonable to assume that $L \to \infty$ and then to scale m and n using L. The expansions for Theorem 3.1(a) as m and/or $n \to \infty$ follow immediately from the results of section 2.2, simply by interchanging $m \leftrightarrow n$ and $\lambda_1 \leftrightarrow \lambda_2$ (or $\rho_1 \leftrightarrow \rho_2$). The expansion of the double integral in part (b) is more complicated, and we have thus far not been able to enumerate all the different cases. However, we were able to obtain detailed results in the heavy traffic case; these we discuss next.

3.2. Heavy traffic behavior. We study $p(m,n)$ in the heavy traffic case, where $N_1(t)$ and $N_2(t)$ are likely to be large. There are now two distinct nontrivial heavy traffic limits. We call these **HTL 1** and **HTL 2**, and the precise scalings are

HTL 1. $\rho_1 + \rho_2 = 1 - \varepsilon$, $\rho_1 = \varepsilon a = \mathcal{O}(\varepsilon)$, $L = A/\varepsilon = \mathcal{O}(\varepsilon^{-1})$,

HTL 2. $\rho_1 + \rho_2 = 1 - \varepsilon$, $\rho_2 = \varepsilon b = \mathcal{O}(\varepsilon)$, $L = A/\varepsilon = \mathcal{O}(\varepsilon^{-1})$.

Note that **HTL 2** has the same scaling of ρ_1 and ρ_2 as we used for the simplified model in section 2. In both these limits we obtain a nontrivial, two-dimensional structure to $p(m,n)$. To achieve this it is also necessary to scale the threshold L to be large.

Below we give only the final results. The calculations are very similar to those presented in section 2.

THEOREM 3.2. *In the heavy traffic limit* **HTL 2**, *we have as $\varepsilon \to 0^+$*
(a) *$m - L = (X - A)/\varepsilon$, $n = Y/\varepsilon$:*

$$p(m,n) \sim \varepsilon^2 e^{-A} P(X - A, Y; b),$$

where P is the density in Theorem 2.5*;*

(b) $m \le L-1$, $n = \mathcal{O}(1)$, $m = X/\varepsilon$, $X > 0$:

$$p(m,n) \sim \varepsilon(\varepsilon b)^n e^{-X};$$

(c) $m, n = \mathcal{O}(1)$:

$$p(m,n) \sim \varepsilon^{n+1} b^n F(m,n), \qquad F(m,n) = \frac{1}{2\pi i}\int_C \frac{1}{(1-z)(2-z)^n z^{m+1}} dz;$$

(d) $m - L = \zeta/\sqrt{\varepsilon} = \mathcal{O}(\varepsilon^{-1/2})$, $m \ge L$, $n = \mathcal{O}(1)$:

$$p(m,n) \sim \varepsilon e^{-A} \frac{1}{2\pi i}\int_C u^{-n-1} e^{-\sqrt{b}\sqrt{1-u}\zeta} du.$$

Here the contours C are small loops about the origin.

We see from Theorem 3.2 that in the limit **HTL 2** the limiting distribution becomes a two-dimensional diffusion in the region $\{X > A,\ Y > 0\}$ coupled to a one-dimensional diffusion along $\{0 < X < A,\ n = 0\}$. The total masses in the respective regions are easily obtained from (a) and (b):

$$\sum_{n=0}^{\infty}\sum_{m=L}^{\infty} p(m,n) \sim e^{-A}\int_0^{\infty}\int_A^{\infty} P(X-A, Y; b)\, dX\, dY = e^{-A},$$

$$\sum_{n=0}^{\infty}\sum_{m=0}^{L-1} p(m,n) \sim \sum_{m=0}^{L-1} p(m,0) \sim \int_0^A e^{-X} dX = 1 - e^{-A}.$$

The masses in regions (c) and (d) are asymptotically $o(1)$ as $\varepsilon \to 0^+$. Theorem 3.2 is easily established from the exact representations in Theorem 3.1.

An alternate approach to the heavy traffic limit would be to analyze the difference equations directly, with the **HTL 2** scaling. Using this approach, it is easy to show that if $p(m,n) \sim \varepsilon^2 Q(X,Y)$, then Q satisfies the PDE and BC:

$$\begin{gathered} Q_{XX} + (b+1)Q_X - bQ_Y = 0, \qquad X > A,\ Y > 0, \\ Q_X(A,Y) + Q_Y(A,Y) + (b+1)Q(A,Y) = 0, \qquad Y > 0, \\ Q(X,0) = 0, \qquad X > A. \end{gathered}$$

Furthermore, by assuming a priori that $p(m,n) \sim \varepsilon^{n+1} R_n(X)$, we find that $R_0(X)$ satisfies

$$R_0''(X) + R_0'(X) = 0, \qquad 0 < X < A.$$

Then by considering separately the scale $m, n = \mathcal{O}(1)$, we derive the BC: $R_0'(0) + R_0(0) = 0$. It follows that $R_0(X) = Ke^{-X}$ and that $Q(X,Y) = K_* P(X-A, Y)$. However, we have the two constants K and K_*, and normalization of $p(m,n)$ will give only one relation between them. Hence a second relation must be obtained; i.e., the two-dimensional diffusion must somehow be coupled to the one-dimensional one. By considering the scale $m - L = \mathcal{O}(\varepsilon^{-1/2})$ and $n = \mathcal{O}(1)$, we obtain another expansion that asymptotically matches to $R_0(X)$ and to $Q(X,Y)$ in the appropriate limits. This yields the additional relation $K_* = R_0(A) = e^{-A}R_0(0) = Ke^{-A}$, and then normalization shows that $K = 1$. The merit of the direct approach sketched

here is that it should be possible to generalize to more complicated models, e.g., ones with general service time distributions.

Next we give the **HTL 1** results.

THEOREM 3.3. *In the heavy traffic limit* **HTL 1**, *we have as* $\varepsilon \to 0^+$

(a) $m = X/\varepsilon < L$, $n = Y/\varepsilon$:

$$p(m,n) \sim \varepsilon^2 P(Y,X;a) = \frac{\varepsilon^2}{2\pi i}\int_{Br} \frac{2\theta e^{\theta X}\exp[-\frac{Y}{2}(a+1+\sqrt{(a+1)^2+4a\theta})]}{a+1+2\theta-\sqrt{(a+1)^2+4a\theta}}d\theta;$$

(b) $m \geq L$, $m-L = \mathcal{O}(1)$, $n = Y/\varepsilon$:

$$p(m,n) \sim (\varepsilon a)^{m+1-L}\frac{1}{2\pi i}\int_{Br}\frac{8\theta e^{\theta A}}{a+1+2\theta-\sqrt{(a+1)^2+4a\theta}} \times \frac{e^{-Y}-e^{-Y(a+1+\sqrt{(a+1)^2+4a\theta})/2}}{[a+1+\sqrt{(a+1)^2+4a\theta}][a-1+\sqrt{(a+1)^2+4a\theta}]}d\theta;$$

(c) $m \geq L$, $m-L = \mathcal{O}(1)$, $n = \mathcal{O}(1)$:

$$p(m,n) \sim \varepsilon(\varepsilon a)^{m+1-L}\frac{1}{2\pi i}\int_C \frac{1}{(1-w)^2(2-w)^{m-L}w^{n+1}}dw \times \frac{1}{2\pi i}\int_{Br}\frac{4\theta e^{\theta A}}{[a+1+2\theta-\sqrt{(a+1)^2+4a\theta}][a+1+\sqrt{(a+1)^2+4a\theta}]}d\theta.$$

Thus, now the limiting distribution behaves as a two-dimensional diffusion in the range $\{0 < X < A, Y > 0\}$, coupled to a one-dimensional diffusion along $\{m = L, Y > 0\}$. The mass in the first region is

$$M_-(A) = \sum_{n=0}^{\infty}\sum_{m=0}^{L-1} p(m,n) \sim \frac{4}{2\pi i}\int_{Br}\frac{e^{\theta A}-1}{[a+1+2\theta-\sqrt{(a+1)^2+4a\theta}][a+1+\sqrt{(a+1)^2+4a\theta}]}d\theta,$$

where we have used part(a) of Theorem 3.3. The remaining mass is, using part (b) with $m = L$,

$$M_+(A) = \sum_{n=0}^{\infty}\sum_{m=L}^{\infty} p(m,n) \sim \sum_{n=0}^{\infty} p(L,n) \sim \frac{a}{2\pi i}\int_{Br}\frac{8\theta e^{\theta A}}{[a+1+2\theta-\sqrt{(a+1)^2+4a\theta}][a+1+\sqrt{(a+1)^2+4a\theta}]^2}d\theta.$$

Using contour integration, it is possible to show that these two expressions indeed sum to one.

Finally, we discuss the direct approach to the diffusion model. In the limit **HTL 1** we find that for $x < A$ ($m < L$) we have $p(m,n) \sim \varepsilon^2\bar{Q}(X,Y)$, where

$$\bar{Q}_{YY} + (a+1)\bar{Q}_Y - a\bar{Q}_X = 0, \qquad 0 < X < A,\ Y > 0,$$
$$\bar{Q}_X(X,0) + \bar{Q}_Y(X,0) + (a+1)\bar{Q}(X,0) = 0, \qquad 0 < X < A,$$
$$\bar{Q}(0,Y) = 0,$$

and thus $\bar{Q}(X,Y) = K_1 P(Y,X;a)$, where P is as in Theorem 2.5. For $Y = \varepsilon n > 0$ and $m = L, L+1, \ldots$ we use $p(m,n) \sim \varepsilon^{l+1} Q_l(Y)$, where $l = m - L$. Then we find that $Q_0(Y)$ satisfies the forced one-dimensional diffusion equation

$$Q_0''(Y) + Q_0'(Y) = -a\bar{Q}(A,Y), \qquad Y > 0.$$

From a careful consideration of the scale $m - L = \mathcal{O}(1)$, $m \geq L$, $n = \mathcal{O}(1)$, we obtain the boundary condition $Q_0(0) = 0$. Then we can easily solve for $Q_0(Y)$ in terms of the previously obtained $\bar{Q}(A,Y)$. Now both diffusions are known up to a common multiplicative constant, which follows from normalization.

4. Discussion and numerical results. We demonstrate the usefulness of both our exact and asymptotic results. The exact results for the two models are given in Theorems 2.1 and 3.1. In both cases, the exact formulas for the stationary probabilities are given in terms of complex contour integrals. These integrals can be evaluated by computing the residue at zero. However, for m and/or n large, the residue calculation is not feasible. The asymptotic results for the tail probabilities (cf. Theorems 2.2–2.4) and the heavy traffic diffusion results (cf. Theorems 2.5, 2.6, 3.2, 3.3) provide good approximations precisely when the computation of the exact solution is difficult.

We consider several examples of the simplified model in section 2. The marginal probability $p(n)$ for the number in the high priority queue is given by the exact formula (2.8), while its asymptotic approximation is given in Theorem 2.4 (b) and (c). For each of the tables, we evaluate the integral in (2.8) by computing the residue at $w = 0$ using the symbolic computation program *Maple*.

In Table 4.1, we consider a system with $\lambda_1 = 1/4$, $\lambda_2 = 1/2$, and $\mu = 1$. Thus, $\rho_1 = 1/4$ and $\rho_2 = 1/2$ so that $(\rho_1, \rho_2) \in R_B$ (cf. Figure 2.1). As we see from Table 4.1, the asymptotic expansion is quite accurate for $n > 5$, where the relative error is less than 5%. In Table 4.2, we consider the same queue as in Table 4.1 except that $\lambda_2 = 2/3$. Again the asymptotic results are extremely accurate.

In Table 4.3, we consider a system with $\lambda_1 = 3/4$, $\lambda_2 = 1/150$, and $\mu = 1$. Thus, $\rho_1 = 3/4$ and $\rho_2 = 1/150$ so that $(\rho_1, \rho_2) \in R_A$. For this case the results are not as accurate as for those in Tables 4.1 and 4.2. As n increases, the accuracy of the asymptotic expansion increases. For $n > 30$ the calculation of the exact formula is difficult, and so we no longer have a basis of comparison. If we choose (ρ_1, ρ_2) closer to the curve separating regions R_A and R_B in Figure 2.1, then the error is larger for each n. However, for sufficiently large n, we expect the asymptotic result to be very accurate. The difference in the size of the errors in these examples is due to the fact that in region R_A the error (in Theorem 2.4(b)) is $\mathcal{O}(1/n)$, while in region R_B the error is exponentially small. In fact, it can be shown that the error term for R_B is the same as the leading term for the region R_A. Neither expansion is valid at or near the transition curve that separates R_A and R_B. There a new expansion must be constructed. This new expansion involves hypergeometric functions, and we do not consider it here.

As a last example, we compare the heavy traffic diffusion approximation for the marginal probability $p(n)$, as given in Theorem 2.5, to the exact expression (2.8). In Figure 4.1, we present graphs of both (2.8) and the heavy traffic approximation in Theorem 2.5 when $\rho_1 = 4/5$ and $\rho_2 = 1/10$, so that $\varepsilon = 1/10$ and $b = 1$. For these parameters, the heavy traffic approximation is $\varepsilon P(Y) = \varepsilon e^{-Y}/\sqrt{\pi Y}$. As we see qualitatively, the heavy traffic result provides an accurate approximation if $n > 4$. In fact, the error is less than 3% for all $n > 4$.

TABLE 4.1

$\rho_1 = 1/4,\ \rho_2 = 1/2$—Region R_B			
n	Exact	Asympt.	Rel. err.
0	0.296535	0.208333	0.297441
1	0.195083	0.15625	0.199061
2	0.135762	0.117188	0.136819
3	0.097346	0.087891	0.097131
4	0.070956	0.065918	0.070998
5	0.052214	0.049438	0.053155
6	0.038647	0.037079	0.040569
7	0.028712	0.027809	0.031446
8	0.021385	0.020857	0.024685
9	0.015955	0.015643	0.019580
10	0.011919	0.011732	0.015669
11	0.008912	0.008799	0.012633
12	0.006668	0.006599	0.010251
13	0.004991	0.004949	0.008366
14	0.003738	0.003712	0.006861
15	0.0028	0.002784	0.005652
16	0.002098	0.002088	0.004674
17	0.001572	0.001566	0.003879
18	0.001178	0.001175	0.003229
19	0.000883	0.000881	0.002697
20	0.000662	0.000661	0.002258
21	0.000496	0.000496	0.001895
22	0.000372	0.000372	0.001594
23	0.000279	0.000279	0.001343
24	0.000209	0.000209	0.001135

TABLE 4.2

$\rho_1 = 1/4,\ \rho_2 = 2/3$—Region R_B			
n	Exact	Asympt.	Rel. err.
0	0.096987	0.080492	0.170066
1	0.08173	0.073785	0.097212
2	0.071742	0.067636	0.057233
3	0.064245	0.062	0.034951
4	0.058116	0.056833	0.022081
5	0.052856	0.052097	0.014355
6	0.048216	0.047756	0.009552
7	0.044061	0.043776	0.006477
8	0.040308	0.040128	0.004460
9	0.036899	0.036784	0.003110
10	0.033793	0.033719	0.002192
11	0.030957	0.030909	0.001559
12	0.028365	0.028333	0.001118
13	0.025993	0.025972	0.000807
14	0.023822	0.023808	0.000586
15	0.021833	0.021824	0.000427
16	0.020011	0.020005	0.000313
17	0.018342	0.018338	0.000230
18	0.016813	0.01681	0.000170
19	0.015411	0.015409	0.000126
20	0.014126	0.014125	0.000094
21	0.012949	0.012948	0.000070
22	0.011869	0.011869	0.000052
23	0.01088	0.01088	0.000039
24	0.009973	0.009973	0.000029

TABLE 4.3

$\rho_1 = 3/4,\ \rho_2 = 1/150$—Region R_A			
n	Exact	Asympt.	Rel. err.
1	0.078876	0.226964	1.877476
2	0.011338	0.021732	0.916778
3	0.001986	0.003204	0.612972
4	0.000385	0.000564	0.462403
5	7.96015e-05	0.000109	0.372002
6	1.71561e-05	2.25008e-05	0.311530
7	3.81328e-06	4.83583e-06	0.268156
8	8.67638e-07	1.07195e-06	0.235488
9	2.01078e-07	2.433e-07	0.209978
10	4.72974e-08	5.626e-08	0.189494
11	1.12622e-08	1.3207e-08	0.172677
12	2.70941e-09	3.13917e-09	0.158619
13	6.57536e-10	7.53991e-10	0.146690
14	1.60781e-10	1.82718e-10	0.136439
15	3.95733e-11	4.46202e-11	0.127533
16	9.79651e-12	1.09693e-11	0.119723
17	2.43757e-12	2.71258e-12	0.112818
18	6.09288e-13	6.74281e-13	0.106669
19	1.52919e-13	1.68388e-13	0.101158
20	3.85217e-14	4.22272e-14	0.096190
21	9.7365e-15	1.06292e-14	0.091689
22	2.46845e-15	2.68467e-15	0.087591
23	6.27566e-16	6.80184e-16	0.083845
24	1.59958e-16	1.7282e-16	0.080407
25	4.0868e-17	4.40246e-17	0.077240
26	1.04642e-17	1.12419e-17	0.074313
27	2.68481e-18	2.87705e-18	0.071601
28	6.90144e-19	7.37819e-19	0.069080
29	1.77717e-19	1.89576e-19	0.066731

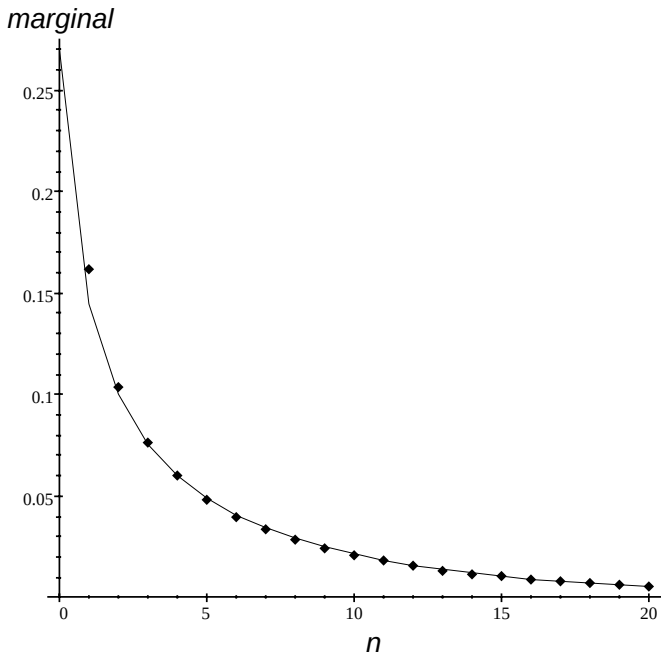


FIG. 4.1. *Graphs of the exact formula (−) for the marginal probability* $p(n)$ *given by* (2.8) *and the heavy traffic result* $(\cdots)$ *from Theorem* 2.5 *when* $\rho_1 = 4/5$ *and* $\rho_2 = 1/10$.

REFERENCES

[1] N. K. Jaiswal, *Priority Queues*, Academic Press, New York, 1968.

[2] H. Takagi, *Queueing Analysis (A Foundation of Performance Evaluation)*, Vol. 1, North–Holland, Dordrecht, The Netherlands, 1993.

[3] L. Takacs, *Priority queues*, Oper. Res., 12 (1964), pp. 63–74.

[4] A. Sugahara, T. Takine, Y. Takahashi, and T. Hasegawa, *Analysis of a nonpreemptive priority queue with SPP arrivals of high class*, Perform. Eval., 21 (1995), pp. 215–238.

[5] J. Y. Lee and Y. H. Kim, *Performance analysis of a hybrid priority control scheme for input and output queueing ATM switches*, in Proceedings of the IEEE INFOCOM'98, San Francisco, CA, 1998, pp. 1470–1477.

[6] S. S. Fratini, *Analysis of a dynamic priority queue*, Comm. Statist. Stochastic Models, 6 (1990), pp. 415–444.

[7] D. S. Lee and B. Sengupta, *Queueing analysis of a threshold based priority scheme for ATM networks*, IEEE/ACM Trans. Networking, 1 (1993), pp. 709–717.

[8] B. D. Choi and D. I. Choi, *An analysis of M,MMPP/G/*1 *finite queue with QLT scheduling policy and Bernoulli schedule*, IEICE Trans. Commun, E81-B (1998), pp. 13–22.

[9] R. Chipalkatti, J. F. Kurose, and D. Towsley, *Scheduling policies for real-time and nonreal-time traffic in a statistical multiplexer*, in Proceedings of the IEEE INFOCOM'89, Ottawa, ON, 1989, pp. 774–783.

[10] W. Fischer and K. Meier-Hellstern, *The Markov-modulated Poisson process (MMPP) cookbook*, Perform. Eval., 18 (1992), pp. 149–171.

[11] N. Bleistein and R. A. Handelsman, *Asymptotic Expansions of Integrals*, Holt, Rinehart and Winston, New York, 1975.

SIAM J. APPL. MATH.
Vol. 63, No. 2, pp. 423–458

MODELLING THE DYNAMICS OF TURBULENT FLOODS*

Z. MEI†, A. J. ROBERTS‡, AND ZHENQUAN LI‡

Abstract. Consider the dynamics of turbulent flow in rivers, estuaries, and floods. Based on the widely used k-ϵ model for turbulence, we use the techniques of center manifold theory to derive dynamical models for the evolution of the water depth and of vertically averaged flow velocity and turbulent parameters. This new model for the shallow water dynamics of turbulent flow resolves the vertical structure of the flow and the turbulence, includes interaction between turbulence and long waves, and gives a rational alternative to classical models for turbulent environmental flows.

Key words. k-ϵ turbulence, water waves, floods, center manifold

AMS subject classifications. 58F39, 76F20, 76B15

PII. S0036139999358866

1. Introduction.

Consider the dynamics of a turbulent flow over ground, as occurs in rivers, channels, or floods. In such flows it is the large-scale horizontal variations which are important; the vertical structure of velocity and turbulence may be expected to be determined by the local conditions of the horizontal flow. In this situation we may seek a model of the flow which involves only "coarse" depth-averaged quantities. Such models have been constructed before; for example, Fredsoe and Deigaard [15, pp. 37–39] depth-average the k-equation model of turbulent flow to model the dynamics of breakers on a beach, whereas depth-averaged k-ϵ equations have been used by Rastogi and Rodi [37] to model heat and mass transfer in open channels and by Keller and Rodi [24] to investigate flood plain flows. The need for such sophisticated models was also indicated by Peregrine [35, p. 97], commenting that an empirical friction law derived from channel flow underestimates the turbulence in breakers and surf, and Mei [29, p. 485], observing that eddy viscosities need to be different in and outside of the surf zone.

However, the recent development of center manifold theory and related techniques presages a much deeper understanding of the process of modelling nonlinear dynamics and foresees the systematic reduction of many nonlinear problems to an underlying low-dimensional system. For example, the process of depth-averaging has been shown to be deficient as a modelling paradigm [44]. In the context of turbulent flow, we show in section 3 how the mean motions may be determined by a few critical modes which have a nontrivial structure in the vertical; e.g., as a first approximation the horizontal velocity u and turbulent energy density k are taken to have a cube-root dependence. Moreover, the amplitude of these modes and their evolution may be expressed in terms of depth-averaged quantities. We derive the following coupled nonlinear set of equations to model the turbulent, large-scale flow of water over ground (see (28)):

$$\frac{\partial \eta}{\partial t} \sim -\frac{\partial(\eta \bar{u})}{\partial x}, \tag{1a}$$

*Received by the editors July 7, 1999; accepted for publication (in revised form) April 9, 2002; published electronically November 19, 2002.
http://www.siam.org/journals/siap/63-2/35886.html
†Generation5: Data Modeling and Statistical Analysis Inc., North York, ON M3B 2R7, Canada (zhen@generation5.net).
‡Department of Mathematics & Computing, University of Southern Queensland, Toowoomba, Queensland 4350, Australia (aroberts@usq.edu.au).

$$\frac{\partial \bar{u}}{\partial t} \sim -1.030\frac{\tilde{\nu}\bar{u}}{\eta^2} + (0.0504 - 0.243\,\tilde{\lambda})\frac{\tilde{\nu}\bar{u}^3}{\eta^2\bar{k}}$$

$$+ 0.961\,g\left(\theta - \frac{\partial \eta}{\partial x}\right) - 1.105,\bar{u}\frac{\partial \bar{u}}{\partial x} + 1.44\frac{\partial}{\partial x}\left(\tilde{\nu}\frac{\partial \bar{u}}{\partial x}\right), \tag{1b}$$

$$\frac{\partial \bar{k}}{\partial t} \sim -0.993\,\bar{\epsilon} - 0.0927\frac{\bar{k}^3}{\eta^2\bar{\epsilon}} + (0.589 + 0.516\,\tilde{\lambda})\frac{\tilde{\nu}\bar{u}^2}{\eta^2} - 1.106\,\bar{u}\frac{\partial \bar{k}}{\partial x}$$

$$+ 1.31\frac{\partial}{\partial x}\left(\tilde{\nu}\frac{\partial \bar{k}}{\partial x}\right), \tag{1c}$$

$$\frac{\partial \bar{\epsilon}}{\partial t} \sim -2.101\frac{\bar{\epsilon}^2}{\bar{k}} + (1.552 - 3.215\,\tilde{\lambda})\frac{\tilde{\nu}\bar{\epsilon}\bar{u}^2}{\bar{k}\eta^2}$$

$$- 0.173\,\tilde{\lambda}\bar{\epsilon}\frac{\partial \bar{u}}{\partial x} + 0.533\,\tilde{\lambda}\frac{\bar{\epsilon}\bar{u}}{\bar{k}}\frac{\partial \bar{k}}{\partial x} - (1 + 0.735\,\tilde{\lambda})\bar{u}\frac{\partial \bar{\epsilon}}{\partial x}$$

$$+ 0.81\frac{\partial}{\partial x}\left(\tilde{\nu}\frac{\partial \bar{\epsilon}}{\partial x}\right). \tag{1d}$$

Here η is the water depth and $\bar{u}$, $\bar{k}$, and $\bar{\epsilon}$ are depth-averaged flow velocity, turbulent energy, and turbulent dissipation, respectively. The other variables appearing are $\tilde{\nu} = C_\mu \bar{k}^2/\bar{\epsilon}$, measuring the local eddy diffusivity (see (24)), and $\tilde{\lambda} = \eta^2\bar{\epsilon}^2/\bar{k}^3$, being proportional to the ratio of the vertical mixing time to the turbulent eddy time (see (25)). For example, in section 5.3 this model is used to predict the flow after a dam breaks. See in Figure 6 the formation of a turbulent bore rushing downstream from the dam. The turbulence in the bore is generally highest near the front and decays away behind as seen in Figure 7. This gives one example of the variations in spatial structure of the turbulence that underlies shallow water flows.

Modelling turbulent flow is one of the major challenges in fluid dynamics. While large eddies, which can be as large as the flow domain, extract energy from the mean flow and feed it into turbulent motion, the eddies also act as a vortex stretching mechanism and transfer the energy to the smallest scales where viscous dissipation takes place. It is the scale at which the dissipation occurs that determines the rate of energy dissipation. However, the inflow of energy into turbulent motion is a characteristic of only the large-scale motion. In other words, the turbulent but small-scale motion is often dominated and determined by the large-scale motion and can be treated as a perturbation of the mean flow. The coupling of energy transportation and energy dissipation with the mean flow is adequately described by widely used second-order closure models. In particular, the most popular choice is the two equation k-ϵ model; see, for example, Launder, Reece, and Rodi [26], Hanjalić and Launder [19], Rodi [50], and Speziale [52] for reviews.

We base our analysis upon the k-ϵ model (section 2) for the turbulence underlying the free surface of a fluid in a channel or river or over a flood plain. The resultant model (1) is basically a model for the evolution of vertically averaged quantities; the model resolves large-scale, compared to the depth, dynamical structures in the horizontal. It is important to distinguish between models obtained by depth-averaged equations (which are known to be incorrect [44] for other similar long-wave dynamics) and our model, which is, for convenience, written in terms of depth-averaged quantities.

In section 4 we derive the "coarse" low-dimensional model (1) from the "fine" k-ϵ equations. Despite the well-recognized limitations of the k-ϵ equations as a model of turbulence, we anticipate that the information retained in our coarse model is

reasonably insensitive to deficiencies in the k-ϵ dynamics. Further modelling may be done via more sophisticated Reynolds stress models for channel flow, such as that described by Gibson and Rodi [17]. One aspect to note is that the model we derive has *no* adjustable parameters—all constants are determined from values established for the k-ϵ turbulence model and its boundary conditions. Thus the model predictions are definitive. We describe some example solutions in section 5 to illustrate the dynamical predictions of the model.

Penultimately, in Appendix A we comment on the status of the theory of center manifolds in this development of a low-dimensional model of turbulent flow using center manifold techniques. Finally, in Appendix B we list the computer algebra program used to perform the intricate algebra in constructing the model.

2. The k-ϵ model of turbulent flow. Consider the two-dimensional inviscid k-ϵ model of turbulent flow over rough ground. Distance parallel to the ground's slope is measured by x, while we measure distance normal to the slope by y. Molecular dissipation is neglected because we anticipate little direct effect for it in flood waves of a depth $\mathcal{O}$(metre) over ground with roughness which may be many times the length-scale of viscous dissipation. Turbulent eddies are proposed to be the dominant mechanism for dispersion and dissipation. We denote the ensemble mean velocity components and pressure by u, v, and p, respectively; that is, for simplicity, we omit any distinguishing overbars (instead overbars will later be used to denote depth-averaged quantities). Then the incompressible k-ϵ model (with ensemble means) is

$$\text{(2a)} \qquad \begin{bmatrix} 0 \\ \frac{\partial \boldsymbol{u}}{\partial t} \end{bmatrix} = \begin{bmatrix} \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} \\ \boldsymbol{F}(p, \boldsymbol{u}) \end{bmatrix},$$

where the vector $\boldsymbol{u} = (u, v, \eta, k, \epsilon)$[1] is formed from the velocities u and v in the lateral and normal directions, respectively, the height of the free surface $y = \eta(x, t)$, the turbulent energy density k, and its dissipation rate ϵ. The nonlinear model governing the evolution of the unknowns $\boldsymbol{u}$ and p is

$$\text{(2b)} \quad \boldsymbol{F}(p, \boldsymbol{u}) = \begin{bmatrix} -u\frac{\partial u}{\partial x} - v\frac{\partial u}{\partial y} - \frac{\partial p}{\partial x} + g\sin\theta - \frac{2}{3}\frac{\partial k}{\partial x} + 2\frac{\partial}{\partial x}\left(\nu\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left\{\nu\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right)\right\} \\ -u\frac{\partial v}{\partial x} - v\frac{\partial v}{\partial y} - \frac{\partial p}{\partial y} - g\cos\theta - \frac{2}{3}\frac{\partial k}{\partial y} + 2\frac{\partial}{\partial y}\left(\nu\frac{\partial v}{\partial y}\right) + \frac{\partial}{\partial x}\left\{\nu\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right)\right\} \\ -u(x, \eta, t)\frac{\partial \eta}{\partial x} + v(x, \eta, t) \\ -u\frac{\partial k}{\partial x} - v\frac{\partial k}{\partial y} + \left\{\frac{\partial}{\partial x}\left(\frac{\nu}{\sigma_k}\frac{\partial k}{\partial x}\right) + \frac{\partial}{\partial y}\left(\frac{\nu}{\sigma_k}\frac{\partial k}{\partial y}\right)\right\} + P_h - \epsilon \\ -u\frac{\partial \epsilon}{\partial x} - v\frac{\partial \epsilon}{\partial y} + \left\{\frac{\partial}{\partial x}\left(\frac{\nu}{\sigma_\epsilon}\frac{\partial \epsilon}{\partial x}\right) + \frac{\partial}{\partial y}\left(\frac{\nu}{\sigma_\epsilon}\frac{\partial \epsilon}{\partial y}\right)\right\} + \frac{C_{\epsilon 1}}{T}P_h - C_{\epsilon 2}\frac{\epsilon^2}{k} \end{bmatrix}.$$

Here the eddy viscosity

$$\text{(3)} \qquad \nu = C_\mu \frac{k^2}{\epsilon}$$

is a result of the turbulent mixing and varies in space and time. Later we use the approximate values of the constants

$$\text{(4)} \qquad C_\mu = 0.09\,, \quad \sigma_k = 1\,, \quad \sigma_\epsilon = 1.3\,, \quad C_{\epsilon 1} = 1.44\,, \quad C_{\epsilon 2} = 1.92\,,$$

[1] We adopt the notation that a vector in parentheses, such as $(u, v, \eta, k, \epsilon)$, is a shorthand for the corresponding column vector.

in order to form definite models. Also

$$P_h = \nu \left[2 \left(\frac{\partial u}{\partial x} \right)^2 + 2 \left(\frac{\partial v}{\partial y} \right)^2 + \left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} \right)^2 \right]$$

describes the generation of turbulence through instabilities associated with mean-velocity gradients. T is the time-scale of the turbulent eddies and is frequently defined as

$$T = \max \left\{ k/\epsilon, C_T \sqrt{\nu_m/\epsilon} \right\};$$

the cutoff at viscous time-scales $\sqrt{\nu_m/\epsilon}$ is to avoid a singularity in turbulent production at a wall; see [14, p. 470], for example. However, here we eschew the incorporation of direct viscous effects and so avoid this singularity by using $T = \bar{k}/\bar{\epsilon}$ as the typical turbulent time-scale, where the overbars denote depth averages. The downward slope of the bed, θ, is assumed to be small and to have negligible variation.

The boundary conditions on the bottom and the free surface are important in the details of the construction of the low-dimensional model. The following arguments lead to the given boundary conditions:

- The standard condition is that, in view of the extremely low density of air, the pressure of the air on the fluid surface is effectively constant, which we take to be zero without loss of generality. Thus the normal stress of the fluid across the free surface should vanish:

$$p + \frac{2}{3}k - \frac{2\nu}{1+\eta_x^2} \left[\frac{\partial v}{\partial y} + \frac{\partial u}{\partial x} - \eta_x \left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y} \right) \right] = 0 \quad \text{on } y = \eta\,. \tag{5}$$

  This is only approximately true—corrections should exist of the order of $\overline{p'\eta'}$ in terms of fluctuations about the ensemble means and similarly for other equations involving the free surface. However, the time-scale of gravity waves, $\sqrt{\ell/(2\pi g)}$, associated with the turbulent length-scale, $\ell \propto k^{3/2}/\epsilon$, should be typically much shorter than the turbulent eddy turn-over time, $\ell/\sqrt{k}$ (true for the scaling introduced in section 3.2), and, as in [22, section 2.3], we expect there to be little interaction between the turbulent fluctuations and the free surface dynamics.
- In this work we assume that the horizontal extent of the flood waves is small enough so that wind stress is negligible. This is in contrast to large-scale geophysical simulations, such as that by Arnold and Noye [2], where the wind stress is very important. Thus the fluid surface is free of tangential stress:

$$\left(1 - \eta_x^2\right) \left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} \right) + 2\eta_x \left(\frac{\partial v}{\partial y} - \frac{\partial u}{\partial x} \right) = 0 \quad \text{on } y = \eta\,. \tag{6}$$

  A wind stress could be incorporated into the model by appropriately replacing the zero right-hand side.
- Symmetry conditions for turbulent variables k and ϵ on the free surface (see Arnold and Noye [3] or Fredsoe and Deigaard [15, p. 117] for examples) lead to

$$\frac{\partial k}{\partial n} = 0 \quad \text{on } y = \eta\,, \tag{7}$$

$$\text{and} \quad \frac{\partial \epsilon}{\partial n} = 0 \quad \text{on } y = \eta\,. \tag{8}$$

These assert that the energy in turbulent eddies cannot be lost or gained by transport through the free surface, and similarly for the turbulent dissipation. More sophisticated models of the free surface effect upon turbulence by Gibson and Rodi [17, p. 238] use these zero net flux boundary conditions. Spilling breakers on the water surface could perhaps be modelled by a turbulent production term on the right-hand side of these boundary conditions.

- At the ground, $y = 0$, there must be no flow across the flat bottom:

$$v(x, 0, t) = 0\,. \tag{9}$$

- Other boundary conditions on the ground are more arguable (compare our treatment with that of Arnold and Noye [3, 4]). We are interested only in the flow outside of any molecular boundary layer that may exist on the stream bed—we imagine that the structure of any ordered viscous layer will be broken up by the roughness. This is supported by recent experiments by Krogstad and Antonia [25] who show that roughness of a wall, even on a scale 1/100th the thickness of a turbulent boundary layer, tends to *reduce* the overall anisotropy of the turbulence. A major limitation of the k-ϵ model is the high level of anisotropy near a wall, so such a reduction in anisotropy due to roughness will favor the k-ϵ model. Note that we treat the ground as $y = 0$ in the mathematical model even though we imagine it to be rough. In effect, ensemble means are also done over all realizations of a "rough" bed with mean slope θ and hence mean position $y = 0$.
  We suppose that the bottom inhibits the turbulence in its immediate neighborhood so that the turbulent energy falls to zero:

$$k = 0 \quad \text{on } y = 0\,. \tag{10}$$

In using the k-ϵ model for near-wall turbulence, Durbin [12, 13] asserts that $\partial k/\partial y = 0$ on the wall as well. However, Figure 1 by Durbin [12] shows this latter condition is significant only for the viscous boundary layer—a layer we ignore due to the roughness of the ground. Instead of this requirement, we place the fairly weak constraint on the turbulent dissipation:

$$\nu \frac{\partial \epsilon}{\partial y} \to 0 \quad \text{as } y \to 0\,. \tag{11}$$

This is weak because $\nu \propto k^2 \to 0$ as $y \to 0$. In essence, this condition asserts that the bed does not directly act as a source or sink of turbulent dissipation.

- Although $\nu \propto k^2 \to 0$ as $y \to 0$, we suppose that ν approaches zero slowly enough so that turbulence is still an effective mixing mechanism near the bed. Thus, the ensemble mean horizontal velocity should also vanish on the bed (as also used by Lin and Falconer [27, p. 740]):

$$u = 0 \quad \text{on } y = 0\,. \tag{12}$$

These three boundary conditions on the bed are the same as those used by "low Reynolds" k-ϵ turbulent models [33, p. 64]. The difference here is that we do not include the near-wall dependence upon local Reynolds numbers $R_t = k^2/\epsilon\nu_m$ and $R_y = \sqrt{k}y/\nu_m$ because these involve molecular viscosity ν_m.

The boundary conditions are different to those we used in an earlier treatment of this problem [30]. The difference occurs because there we assumed that the stress $\partial u/\partial y$ is small near the bottom and is more appropriate to weakly turbulent flows. Here we seek the dynamics of flows with a strong level of turbulence leading to the boundary condition (12).

3. Basis of the low-dimensional model. In this paper we consider flows that vary "slowly" in the x and t directions. In this context, the meaning of "slow" is that the dynamics are slower relative to the vertical mixing time induced by the turbulence. In particular, the derivatives of the flow variables with respect to x and t are small quantities that can be collected with the "nonlinear" part of the equations and treated as perturbations. Hence we rewrite the equations as

$$\begin{bmatrix} 0 \\ \dot{u} \\ \dot{v} \\ \dot{\eta} \\ \dot{k} \\ \dot{\epsilon} \end{bmatrix} = \begin{bmatrix} 0 & 0 & \frac{\partial}{\partial y} & 0 & 0 & 0 \\ 0 & \frac{\partial}{\partial y}\left(\nu\frac{\partial}{\partial y}\right) & 0 & 0 & 0 & 0 \\ -\frac{\partial}{\partial y} & 0 & 2\frac{\partial}{\partial y}\left(\nu\frac{\partial}{\partial y}\right) & 0 & -\frac{2}{3}\frac{\partial}{\partial y} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{\sigma_k}\frac{\partial}{\partial y}\left(\nu\frac{\partial}{\partial y}\right) & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{\sigma_\epsilon}\frac{\partial}{\partial y}\left(\nu\frac{\partial}{\partial y}\right) \end{bmatrix} \begin{bmatrix} p \\ u \\ v \\ \eta \\ k \\ \epsilon \end{bmatrix}$$

$$+ \begin{bmatrix} \frac{\partial u}{\partial x} \\ -u\frac{\partial u}{\partial x} - v\frac{\partial u}{\partial y} - \frac{\partial p}{\partial x} + g\sin\theta - \frac{2}{3}\frac{\partial k}{\partial x} + 2\frac{\partial}{\partial x}\left(\nu\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(\nu\frac{\partial v}{\partial x}\right) \\ -u\frac{\partial v}{\partial x} - v\frac{\partial v}{\partial y} - g\cos\theta + \frac{\partial}{\partial x}\left[\nu\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right)\right] \\ -u(x,\eta,t)\frac{\partial \eta}{\partial x} + v(x,\eta,t) \\ -u\frac{\partial k}{\partial x} - v\frac{\partial k}{\partial y} + \frac{\partial}{\partial x}\left(\frac{\nu}{\sigma_k}\frac{\partial k}{\partial x}\right) + P_h - \lambda\epsilon \\ -u\frac{\partial \epsilon}{\partial x} - v\frac{\partial \epsilon}{\partial y} + \frac{\partial}{\partial x}\left(\frac{\nu}{\sigma_\epsilon}\frac{\partial \epsilon}{\partial x}\right) + \frac{C_{\epsilon 1}}{T}P_h - C_{\epsilon 2}\lambda\frac{\epsilon^2}{k} \end{bmatrix}$$

$$= \mathcal{L}(p, \boldsymbol{u}) + \mathcal{F}(p, \boldsymbol{u}, \lambda)\,. \tag{13}$$

Treating the time and lateral variations and the "nonlinear" terms as small, we see that $\mathcal{L}(p, \boldsymbol{u})$ comprises the leading term in the equation.

With a little adaptation, the operator $\mathcal{L}$ has a critical space of equilibrium points parametrized by the water depth η and the mean fields $\bar{u}$, $\bar{k}$, and $\bar{\epsilon}$. To ensure that the turbulent energy and turbulent dissipation are critical modes of $\mathcal{L}$ and thus retained in our model of turbulent floods, we need k and ϵ to be conserved to leading order. The parameter $\lambda \in [0, 1]$ is an artificial parameter which we use to adjust the rate of decay of turbulent energy and its dissipation; by making λ small we initially neglect the natural turbulent decay, but when $\lambda = 1$ we recover the standard k-ϵ model. This is reasonable because the combined effect of $\dot{k} = -\epsilon$ and $\dot{\epsilon} = -C_{\epsilon 2}\epsilon^2/k$ is a *slow* algebraic decay of turbulence [28, p. 277] that is appropriate for the long-term center manifold dynamics.

3.1. Vertical mixing. The operator $\mathcal{L}$ is considered to be the dominant feature of the k-ϵ model (2). It is primarily composed of the differential operator

$$\frac{\partial}{\partial y}\left(\nu\frac{\partial}{\partial y}\right),$$

which represents vertical mixing by turbulent eddies. By identifying this as the dominant term in the equations we are physically supposing that turbulent mixing in the

vertical is stronger than the other processes that redistribute momentum and turbulent energy. For example, Lin and Falconer [27, p. 738] comment that "the shallow regions of tidal embayments are usually well mixed."

The boundary conditions on the bottom, (10) and (11), in conjunction with the k and ϵ components of $\mathcal{L}$ in (13) admit homogeneous solutions $k \propto y^{1/3}$ and ϵ constant. Such a cube-root profile in the vertical fits with arguments that the turbulence should be weaker near the bottom due to its constraining effects. Given such a profile, the turbulent diffusivity $\nu \propto y^{2/3}$ and so the horizontal velocity component of $\mathcal{L}$ also admits homogeneous solutions $u \propto y^{1/3}$. Although our long-wave model will be expressed in terms of depth-averaged quantities, we base the vertical structure that they measure on these cube-root profiles.

Traditionally, many theoretical approaches have assumed constant or near constant vertical profiles (see [15, section 4.3.1] or [36, p. 670] for examples), as indeed we also have in an earlier treatment of this problem [30]. However, as seen in experiments the horizontal velocity profile is typically curved (see [51, Fig. 12], [7], or [54, Fig. 2]) as is the turbulent energy (see [15, Fig. 4.25]). A logarithmic profile is a well-established approximation; here we work with the cube-root profile as it is compatible with the k-ϵ equations, is analytically tractable, and is a rough initial approximation to the logarithmic profile.

These cube-root profiles result in downwards turbulent transport of momentum and turbulent energy with constant flux and eventual removal from the fluid at the bed. In order to maintain flow at leading order (only) we modify the conservative free surface boundary conditions (6)–(7) to supply the requisite flux at leading order and to remove the supply at higher order. We replace (6)–(7) with the boundary conditions that on $y = \eta$

$$(1 - a\gamma)\left[\left(1 - \eta_x^2\right)\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right) + 2\eta_x\left(\frac{\partial v}{\partial y} - \frac{\partial u}{\partial x}\right)\right] = \frac{1-\gamma}{3\eta}u, \tag{14}$$

$$(1 - a\gamma)\left[\frac{\partial k}{\partial y} - \frac{\partial \eta}{\partial x}\frac{\partial k}{\partial x}\right] = \frac{1-\gamma}{3\eta}k, \tag{15}$$

where the artificial parameter γ, as for λ introduced earlier, is small in the asymptotic scheme but eventually will be set to 1 to recover the desired boundary conditions (6)–(7). Such manipulation of the governing equations has worked well in developing analogous models of the laminar viscous flow of a thin fluid layer [44, 47]; the resulting model was found to be independent of the two different manipulations. The particular choice (14)–(15) here enables us to develop the necessary asymptotic expansions purely in polynomials of $y^{1/3}$.

The center manifold analysis forms a power series in γ which needs to be summed for $\gamma = 1$. The parameter $a \neq 1$ is free for us to choose. Initially we omitted a, equivalent to choosing $a = 0$, but after two years of exploration we determined that an Euler transformation of the series in γ greatly improves the convergence; e.g., see Hinch [21, Chap. 8]. Introducing $a \neq 0$ is equivalent to such an Euler transformation. Another view of $a \neq 0$ is that of an over-relaxation parameter in an iterative scheme. In this problem it appears that $a \approx 1/2$ is a good compromise between conflicting influences and is used henceforth.

From the special structure of the vertical mixing operator $\mathcal{L}$ with the modified boundary conditions, we deduce that there are four critical modes of interest in the long-term dynamics. These modes correspond to the leading-order conservation, and hence long-life, of fluid mass, momentum, turbulent energy, and turbulent dissipation.

These modes span the space $\mathcal{M}_0$:

$$(p, \boldsymbol{u}) = \left(g(\eta - y),\, U\tfrac{4}{3}\left(\tfrac{y}{\eta}\right)^{1/3},\, 0,\, H,\, K\tfrac{4}{3}\left(\tfrac{y}{\eta}\right)^{1/3},\, E \right), \tag{16}$$

where U, H, K, and E are arbitrary functions of x and t. Note that the turbulent diffusivity, ν, then also varies slowly in x and t; to leading order it is

$$\nu_0 = C_\mu \frac{16K^2}{9E} \left(\frac{y}{\eta}\right)^{2/3}.$$

As proven in Appendix A, the dominant operator $\mathcal{L}$, linearized about the space of equilibrium points, has eigenvalues which are all negative (due to the decaying dynamics of turbulent dispersion), except for the four zero eigenvalues corresponding to the four conserved modes. By continuity in the "nonlinear" perturbation $\mathcal{F}$, the center manifold is also exponentially attractive, at least for small enough nonlinearity. Since all other modes decay exponentially quickly, the long time behavior of the flow is determined by the functions $U(x,t)$, $H(x,t)$, $K(x,t)$, and $E(x,t)$. Respectively, these represent the vertically averaged horizontal velocity, the surface elevation, the vertically averaged turbulent energy, and the vertically averaged turbulent dissipation. In essence, we construct a "vertically averaged" model, but in u and k there is structure in the vertical, roughly proportional to $y^{1/3}$, whose amplitude we measure by the vertical average.

3.2. Approximating the center manifold. Center manifold techniques systematically develop such a model in the vertically averaged quantities. Based on the relatively low-dimensional space of exponentially attractive equilibria $\mathcal{M}_0$, center manifold theory [9] suggests that the nonlinear terms $\mathcal{F}$ "bend" $\mathcal{M}_0$ to a nearby manifold $\mathcal{M}$ of slow evolution. Further, $\mathcal{M}$ will similarly attract exponentially quickly all solutions in its vicinity; in standard formulations $\mathcal{M}$ is called the center manifold. Once on $\mathcal{M}$, solutions evolve slowly according to a low-dimensional system of evolution equations—these evolution equations form the simplified model of the original dynamics. This general approach to forming low-dimensional models of dynamical systems is reviewed by Roberts [45].

We have never found it profitable to decompose dynamics into that on the null space $\mathcal{M}_0$ and its complement. Many do this in order to write the center manifold as a graph. We view the complement space as an artifice of the linearization, whereas we are interested in a physically meaningful parametrization of the center manifold $\mathcal{M}$. Thus the appropriate geometric picture is simply the curving center manifold $\mathcal{M}$ embedded in the original physical state space. Hence we construct the model in terms of easily understandable physical quantities, namely the vertical averages. Nonetheless, the physically relevant complement space at any point on $\mathcal{M}$ is the local direction of projection of initial conditions [41, 11, 48]. Determining this projection is difficult for nonlinear long-wave approximations and is beyond the scope of this paper.

In this problem, the center manifold $\mathcal{M}$ is parametrized by the four "amplitudes," U, H, K, and E, which are functions of x and evolve in time. Due to the difficult nature of the nonlinear terms in the k-ϵ model (2) we have to be very careful about these amplitudes and their derivatives. We introduce two independent small parameters: δ as an amplitude scale and ϑ to scale spatial derivatives. Then we treat the flow fields and derivatives as

$$u, k = \mathcal{O}(\delta^2), \quad \epsilon = \mathcal{O}(\delta^3), \quad \eta = h + \mathcal{O}(\delta^2), \quad \text{and} \quad \frac{\partial}{\partial x} = \mathcal{O}(\delta\vartheta). \tag{17}$$

Note that in this scaling, the turbulent length-scale $\ell \propto k^{3/2}/\epsilon = \mathcal{O}(1)$. This ensures that the turbulent eddies modelled by k and ϵ are not asymptotically larger than the water depth; large-scale horizontal eddies are resolved by variations in the amplitudes of the model. Also from these scalings, the turbulent diffusivity $\nu = \mathcal{O}(k^2/\epsilon) = \mathcal{O}(\delta)$ and thus the time-scale of vertical mixing is $\mathcal{O}(1/\delta)$. Consequently, we must consider horizontal scales larger than $\mathcal{O}(1/\delta)$, which accounts for requiring the product $\delta\vartheta$ in the scaling of horizontal derivatives. In standard applications of center manifold theory, we are free to scale the amplitudes in any reasonable fashion or, indeed, to treat the amplitudes as independent; as discussed in [40] a change in the scaling just reorders the appearance of the same set of terms in the model. In this application of center manifold techniques, physical considerations and the nonstandard nonlinearities place the constraints on the scaling that derivatives $\frac{\partial}{\partial x} = \mathcal{O}(\delta\vartheta)$ and that the slope $\theta = \mathcal{O}(\delta^3\vartheta)$. Nonetheless, we exploit usefully some of the capability of center manifold techniques to treat amplitudes as independent variables by the above introduction of two independent small parameters, δ and ϑ.

Two other independent small parameters in this problem are γ, the artificial forcing at the free surface, and λ, an artificial adjustment of turbulent interaction. We treat all four of these parameters as independently small.

In terms of the field variables, we define the four amplitudes U, H, K, and E in terms of physical quantities such that

$$\bar{u} = \delta^2 U\,, \tag{18a}$$

$$\eta = h + \delta^2 H\,, \tag{18b}$$

$$\bar{k} = \delta^2 K\,, \tag{18c}$$

$$\bar{\epsilon} = \delta^3 E\,, \tag{18d}$$

where the overbar denotes a vertical average over the whole fluid depth at any x and t; for example,

$$\bar{u} = \frac{1}{\eta}\int_0^\eta u\,dy. \tag{19}$$

Denoting the collective amplitudes by $\boldsymbol{s}(x,t) = (U, H, K, E)$, we pose the low-dimensional assumption that the evolution of the physical variables may be expressed in terms of the evolution of the four amplitudes (effectively equivalent to the "slaving" principle of synergetics [18]):

$$(p, \boldsymbol{u}) = \mathcal{V}(y, \boldsymbol{s}) \quad \text{such that} \quad \frac{\partial \boldsymbol{s}}{\partial t} = \mathcal{G}(\boldsymbol{s})\,. \tag{20}$$

In general, we cannot find these functions $\mathcal{V}$ and $\mathcal{G}$ exactly, as this would be tantamount to exactly solving the original equations. Instead we determine asymptotic approximations in the four small parameters.

It would be decidedly awkward to explicitly write out an asymptotic expansion in the four asymptotic parameters. But it is also inappropriate to link their relative magnitudes into one parameter as we need to find relatively high-order in γ but not in the others. Thus we apply an iterative algorithm in computer algebra to find the center manifold and the evolution thereon, which is based directly upon the approximation theorem 3 in [9, 46] and its variants, as explained in detail by Roberts [46]. An outline of the procedure follows.

The aim is to find the functional $\mathcal{V}$ and evolution $\mathcal{G}$ such that the pressure, velocity, and turbulence fields described by (20) form actual solutions of the scaled turbulent equations—this ensures fidelity between our model and the fluid dynamics of the k-ϵ equations. Suppose that at some stage in an iterative scheme we have some approximation, $\tilde{\mathcal{V}}$ and $\tilde{\mathcal{G}}$. We then seek a correction, $\mathcal{V}'$ and $\mathcal{G}'$, to obtain a more accurate solution to the turbulence equations. Substituting

$$(p, \boldsymbol{u}) = \tilde{\mathcal{V}} + \mathcal{V}' \quad \text{such that} \quad \frac{\partial \boldsymbol{s}}{\partial t} = \tilde{\mathcal{G}} + \mathcal{G}'$$

into the scaled turbulence equations, then rearranging, dropping products of corrections, and using a leading-order approximation wherever factors multiply corrections (see [46] for details), we obtain a system of equations for the corrections which is of the form

$$\mathcal{L}\mathcal{V}' = \tilde{\boldsymbol{R}} + \mathcal{E}\mathcal{G}' \,, \tag{21}$$

where $\mathcal{E} = \partial\mathcal{V}/\partial\boldsymbol{s}|_{\boldsymbol{s}=\boldsymbol{0}}$ is the basis for the linear subspace $\mathcal{M}_0$ in (16), and, most importantly, $\tilde{\boldsymbol{R}}$ is the residual of the scaled k-ϵ equations using the current approximation $\tilde{\mathcal{V}}$ and $\tilde{\mathcal{G}}$. This homological equation is solved by choosing corrections $\mathcal{G}'$ to the evolution so that the right-hand side is in the range of $\mathcal{L}$; then the correction to the fields $\mathcal{V}'$ is determined. The solution is made unique by requiring that the amplitudes $\boldsymbol{s}$ have some specific physical meaning, here the vertical averages of the fields as in (18). Then the current approximation $\tilde{\mathcal{V}}$ and $\tilde{\mathcal{G}}$ is updated. The iteration is repeated until the residual of the governing equations, $\tilde{\boldsymbol{R}}$, becomes zero to some order of error, whence the center manifold model will be accurate to the same order of error (by the approximation theorem 3 of Carr [9]).

A computer algebra program was written to perform all the necessary detailed algebra for this physical problem. A listing is given in Appendix B. The very important feature of this iteration scheme is that it is performed until the residuals of the actual governing equations are zero to some order of error. Thus the correctness of the results that we present here is based only upon the correct evaluation of the residuals and upon sufficient iterations to drive these to zero. The key to the correctness of the results produced by the computer program is the proper coding of the k-ϵ turbulence equations. These can be seen in the computed residuals within the iterative loop of the program.

4. Constructing the low-dimensional model. As a first step in constructing a dynamical model we discard any variation in x and any influence of slope θ. Thus we first examine the dynamics of a uniform layer of turbulent fluid, k and ϵ nonzero, slowly decelerating, $u \to 0$, due to turbulent drag on the bed.

4.1. The physical fields to low-order. As discussed in section 3.2 on the vertical mixing operator, the leading-order approximation to the shape of the center manifold is just the solutions to $\mathcal{L}\mathcal{V} = 0$. We deduce that

$$u \approx \delta^2 U(x,t)\tfrac{4}{3}\left(\tfrac{y}{\eta}\right)^{1/3}, \quad v \approx 0, \quad k \approx \delta^2 K(x,t)\tfrac{4}{3}\left(\tfrac{y}{\eta}\right)^{1/3},$$
$$\epsilon \approx \delta^3 E(x,t), \quad \text{and} \quad \nu \approx \delta C_\mu \tfrac{16K^2}{9E}\left(\tfrac{y}{\eta}\right)^{2/3}.$$

At higher orders in the small parameters δ, ϑ, λ, and γ, we construct more refined descriptions of the fluid flow and its dynamics through the evolution of the amplitudes.

However, we leave the influence of spatial variations through nonzero ϑ until the next section.

By iterations of the scheme outlined in the previous section we obtain a basic description of the turbulence production and decay. The nonlinear processes and boundary condition corrections modify the cube-root profile and simultaneously determine the slow evolution of the amplitudes.

It is useful to record the asymptotic expansions directly in terms of physical quantities η, $\bar{u}$, $\bar{k}$, and $\bar{\epsilon}$ rather than the corresponding artificially scaled quantities H, U, K, and E. We find the following expressions for the first significant modifications to the fields within the fluid, written in terms of a scaled vertical coordinate $\zeta = y/\eta$ which ranges from $\zeta = 0$ at the bed to $\zeta = 1$ at the fluid surface:

$$v \approx 0\,, \tag{22a}$$

$$u \approx \upsilon_0(\zeta)\bar{u} + \left[C_{\epsilon 1}\sigma_\epsilon \upsilon_1(\zeta) - \sigma_k \upsilon_2(\zeta)\right] \frac{\bar{u}^3}{\bar{k}} + \left(C_{\epsilon 2}\sigma_\epsilon - \sigma_k\right) \upsilon_3(\zeta) \frac{\lambda\eta^2 \bar{u}\bar{\epsilon}}{\tilde{\nu}\bar{k}}\,, \tag{22b}$$

$$k \approx \upsilon_0(\zeta)\bar{k} + \left[C_{\epsilon 1}\sigma_\epsilon \upsilon_1(\zeta) + \sigma_k \upsilon_2(\zeta)\right] \bar{u}^2 + \left(C_{\epsilon 2}\sigma_\epsilon + \sigma_k\right) \upsilon_3(\zeta) \frac{\lambda\eta^2 \bar{\epsilon}}{\tilde{\nu}}\,, \tag{22c}$$

$$\epsilon \approx \bar{\epsilon} + C_{\epsilon 1}\sigma_\epsilon \epsilon_p(\zeta) \frac{\bar{u}^2\bar{\epsilon}}{\bar{k}} + C_{\epsilon 2}\sigma_\epsilon \epsilon_d(\zeta) \frac{\lambda\eta^2\bar{\epsilon}^2}{\tilde{\nu}\bar{k}}\,, \tag{22d}$$

$$\nu \approx \nu_0(\zeta)\frac{\bar{k}^2}{\bar{\epsilon}} + \left[-C_{\epsilon 1}\sigma_\epsilon \nu_1(\zeta) + \sigma_k \nu_2(\zeta)\right] \frac{\bar{\epsilon}\eta^2}{\bar{k}} + \left[C_{\epsilon 2}\sigma_\epsilon \nu_3(\zeta) - \sigma_k \nu_4(\zeta)\right] \frac{\bar{\epsilon}^3\eta^4}{\bar{k}^4}\,. \tag{22e}$$

These expressions are correct to errors $\mathcal{O}\big(\delta^6 + \lambda^3 + \gamma^3, \vartheta\big)$, where, for example, a multinomial term

$$\delta^a \lambda^b \gamma^c \vartheta^d = \mathcal{O}\big(\delta^A + \lambda^B + \gamma^C, \vartheta^D\big) \quad \text{if} \quad \frac{a}{A} + \frac{b}{B} + \frac{c}{C} \geq 1 \text{ and } d \geq D\,.$$

The vertical structure functions occurring in the expressions on the right-hand side of (22) are as follows:

- For the turbulent dissipation,

$$\epsilon_p(\zeta) = \frac{4}{9}\zeta^{4/3} - \frac{8}{9}\zeta^{2/3} + \frac{12}{35}\,,$$
$$\epsilon_d(\zeta) = -\frac{243}{512}\zeta^{4/3} + \frac{81}{128}\zeta - \frac{405}{3584}\,,$$

  as shown in Figure 1. See the effect of turbulent dissipation production, at a rate proportional to $\bar{u}^2$, through the velocity shear. Since velocity shear is largest near the bed, as seen in the shape of $\epsilon_p(\zeta)$, this enhances turbulent dissipation ϵ near the bed.

  However, the natural turbulent dissipation within the fluid causes a greater decay of turbulent dissipation near the bed, due to the smaller turbulent energy there, and so counters this enhancement. Being proportional to $1/\tilde{\nu}$, this effect on the ϵ-profile is greatest in weakly turbulent flows.

- For the turbulent energy density,

$$\upsilon_0(\zeta) = \frac{4}{3}\zeta^{1/3} + \gamma\left(\frac{1}{12}\zeta^{1/3} - \frac{1}{6}\zeta^{5/3}\right),$$

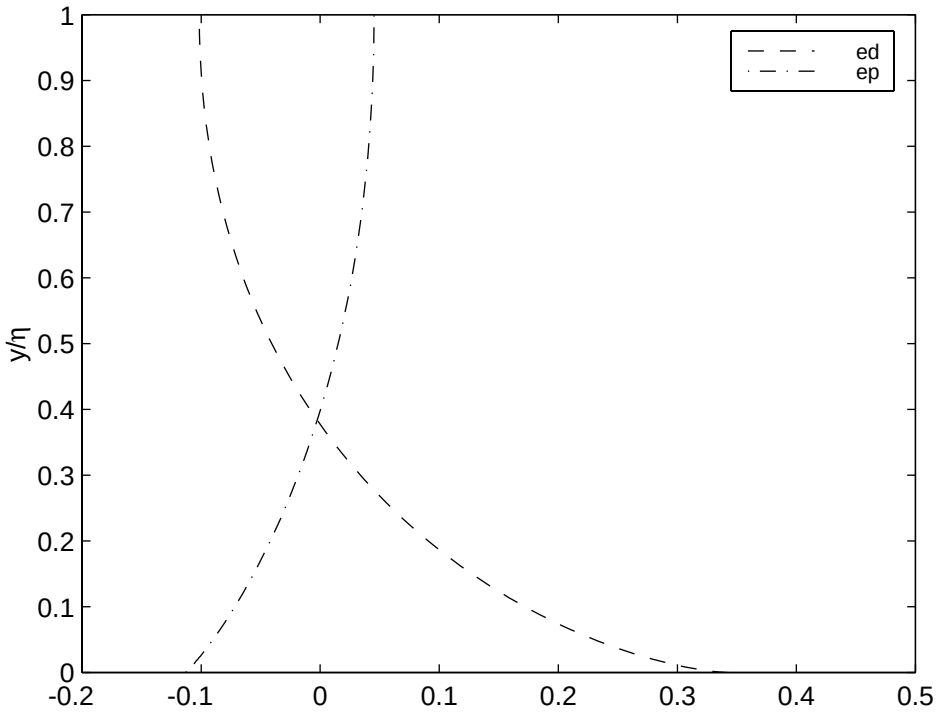


FIG. 1. *The vertical structure of the turbulent dissipation field within the fluid as a function of the scaled vertical coordinate* $\zeta = y/\eta$: $-\cdot-\cdot-\cdot-$, $\epsilon_p(\zeta)$; $- - - -$, $\epsilon_d(\zeta)$.

$$\upsilon_1(\zeta) = \frac{16}{135}\zeta^{5/3} - \frac{32}{135}\zeta + \frac{8}{81}\zeta^{1/3},$$
$$\upsilon_2(\zeta) = \frac{1}{9}\zeta^{5/3} - \frac{4}{9}\zeta^{2/3} + \frac{3}{10}\zeta^{1/3},$$
$$\upsilon_3(\zeta) = -\frac{27}{256}\zeta^{5/3} + \frac{9}{64}\zeta^{4/3} - \frac{99}{3584}\zeta^{1/3},$$

as shown in Figure 2. Observe the cube-root structure in the vertical is modified to $\upsilon_0(\zeta)$. As shown in Figure 2, when γ is set to 1 to recover the original boundary conditions from (15), the cube-root dependence is maintained near the bed but is effectively flattened near the fluid surface to closely approximate the absence of turbulent energy flux through the free surface. This correction is simultaneously determined with a corresponding decay term in the evolution equations, as seen below, due to the removal of the sustaining flux. We look even closer at the effects of modifying the free surface boundary conditions in section 4.2.

Also the effect of turbulent production, proportional to $\bar{u}^2$, through the velocity shear is largest near the bed; as seen in the shape of $\upsilon_1(\zeta)$ and $\upsilon_2(\zeta)$, this enhances turbulent energy k near the bed.

However, the natural turbulent dissipation within the fluid causes a relatively greater decay of turbulent energy near the bed as compared with the body of the fluid, as seen in $\upsilon_3(\zeta)$, and so counters this enhancement. Being proportional to $1/\tilde{\nu}$, this effect on the k-profile is greatest in weakly turbulent

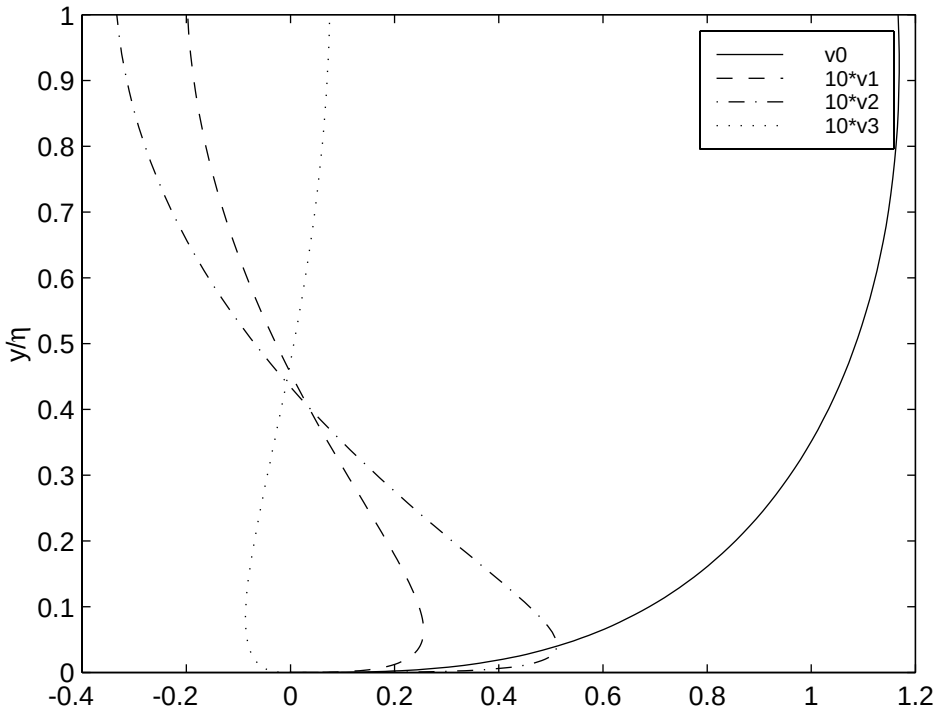


FIG. 2. *The vertical structure functions for the turbulent energy density and horizontal velocity fields within the fluid as a function of the scaled vertical coordinate* $\zeta = y/\eta$: ———, $\upsilon_0(\zeta)$ *(with* $\gamma = 1$*);* – – –, $10 \times \upsilon_1(\zeta)$; –·–·–·–, $10 \times \upsilon_2(\zeta)$; ·····, $10 \times \upsilon_3(\zeta)$.

flows.

- The basic cube-root structure of the horizontal velocity is modified in exactly the same way and for the same reasons as for the basic turbulent energy k-profile.
  Modifications of the velocity profile due to the turbulent production and dissipation occur, but they occur primarily through the indirect effects of modifications to the turbulent diffusivity profile $\nu(\zeta)$. These are weak due to the subtractions in (22b).
- The corresponding vertical structure of the turbulent mixing coefficient ν is shown in Figure 3, where the five components are

$$\begin{aligned}
\nu_0(\zeta) &= \frac{20}{9}\zeta^{2/3} - \frac{8}{9}\zeta^2, \\
\nu_1(\zeta) &= \frac{64}{135}\zeta^2 - \frac{128}{135}\zeta^{4/3} + \frac{2944}{8505}\zeta^{2/3}, \\
\nu_2(\zeta) &= -\frac{32}{27}\zeta + \frac{4}{5}\zeta^{2/3} + \frac{8}{27}\zeta^2, \\
\nu_3(\zeta) &= \frac{57}{448}\zeta^{2/3} - \frac{3}{4}\zeta^{5/3} + \frac{9}{16}\zeta^2, \\
\nu_4(\zeta) &= \frac{33}{224}\zeta^{2/3} + \frac{9}{16}\zeta^2 - \frac{3}{4}\zeta^{5/3}.
\end{aligned}$$

Simultaneously with the determination of the above fields, the solvability con-

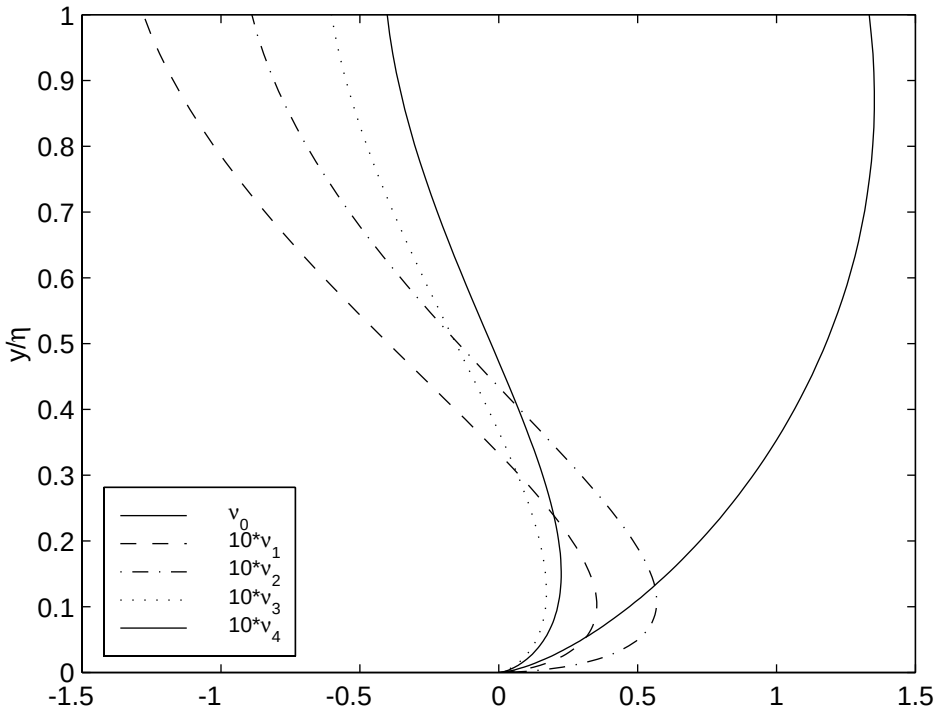


FIG. 3. *The vertical structure functions for the turbulent mixing coefficient within the fluid as a function of the scaled vertical coordinate* $\zeta = y/\eta$: ——— *(right),* $\nu_0(\zeta)$ *(with* $\gamma = 1$*);* – – –, $10 \times \nu_1(\zeta)$; –·–·–, $10 \times \nu_2(\zeta)$; ·····, $10 \times \nu_3(\zeta)$; ——— *(left),* $\nu_4(\zeta)$.

dition for the linear equations of the form (21) supplies terms in the asymptotic, low-dimensional evolution equations for the amplitudes of the four critical modes. Writing these in terms of physical variables we find, with errors $\mathcal{O}\big(\delta^6 + \lambda^3 + \gamma^3, \vartheta\big)$, that

$$\frac{\partial \eta}{\partial t} \sim 0, \tag{23a}$$

$$\frac{\partial \bar{u}}{\partial t} \sim -\frac{56\gamma\tilde{\nu}}{81\eta^2}\bar{u} - \frac{\lambda}{16}\left[C_{\epsilon 2}\sigma_\epsilon - \sigma_k\right]\frac{\bar{\epsilon}\bar{u}}{\bar{k}} + \frac{16}{81}\left[\frac{32}{45}C_{\epsilon 1}\sigma_\epsilon - \sigma_k\right]\frac{\tilde{\nu}\bar{u}^3}{\eta^2\bar{k}}, \tag{23b}$$

$$\frac{\partial \bar{k}}{\partial t} \sim -\lambda\left[\frac{7}{8} + \frac{C_{\epsilon 2}\sigma_\epsilon}{16\sigma_k}\right]\bar{\epsilon} - \frac{56\gamma\tilde{\nu}}{81\sigma_k\eta^2}\bar{k} + \frac{16}{243}\left[7 + \frac{32C_{\epsilon 1}\sigma_\epsilon}{15\sigma_k}\right]\frac{\tilde{\nu}}{\eta^2}\bar{u}^2, \tag{23c}$$

$$\frac{\partial \bar{\epsilon}}{\partial t} \sim -\lambda\frac{9}{8}C_{\epsilon 2}\frac{\bar{\epsilon}^2}{\bar{k}} + \frac{256}{243}C_{\epsilon 1}\frac{\tilde{\nu}\bar{\epsilon}}{\eta^2\bar{k}}\bar{u}^2, \tag{23d}$$

where

$$\tilde{\nu}(x,t) = C_\mu\frac{\bar{k}^2}{\bar{\epsilon}} \tag{24}$$

is a measure of the local turbulent diffusivity. These form a crude approximation to the evolution equations for the four amplitudes of the model when there are no horizontal variations.

TABLE 1
Terms in the series expansions in γ of selected coefficients in the model for homogeneous turbulent decay. The last row is the sum of known terms at $\gamma = 1$.

	In $\partial\bar{u}/\partial t$			In $\partial\bar{k}/\partial t$	
	$-\tilde{\nu}\bar{u}/\eta^2$	$-\tilde{\nu}\bar{u}\bar{\lambda}/\eta^2$	$\tilde{\nu}\bar{u}^3/\eta^2\bar{k}$	$\tilde{\nu}\bar{u}^2/\eta^2$	$-\bar{\epsilon}$
1	0	+1.03889	+0.06542	+0.72385	+1.03100
γ	+0.69136	−0.86096	+0.01388	−0.13682	−0.05665
γ^2	+0.22783	+0.14923	−0.01564	−0.02193	+0.01602
γ^3	+0.07479	+0.02700	−0.00921	+0.00569	+0.00250
γ^4	+0.02448	+0.00462	−0.00333	+0.00814	+0.00027
γ^5	+0.00801	+0.00069	−0.00080	+0.00549	−0.00002
γ^6	+0.00262	+0.00006	−0.00003	+0.00300	−0.00003
γ^7	−0.00086	−0.00002	+0.00017	+0.00147	−0.00001
$\sum$	1.02995	0.35950	0.05040	0.58892	0.99307

The two fifth-order evolution equations (23c)–(23d) summarize the turbulent production and decay processes. Setting the artificial parameters $\lambda = \gamma = 1$ to approximate the dynamics of the original problem, observe, first, the natural decay of turbulent energy and dissipation in the body of the fluid, second, decay of turbulent energy with coefficient proportional to $\tilde{\nu}/\eta^2$ via turbulent mixing transporting energy to the bed, and third, the generation of turbulence energy and its dissipation through the shear in the vertical, proportional to $\tilde{\nu}\bar{u}^2/\eta^2$.

The horizontal velocity evolution (23b) similarly includes terms $\tilde{\nu}\bar{u}/\eta^2$, which represents the effective drag of the bottom via turbulent mixing to the bed, and weak cubic, proportional to $\tilde{\nu}\bar{u}^3/(\eta^2\bar{k})$, and linear, proportional to $\bar{\epsilon}\bar{u}/\bar{k}$, modification of this drag through changes in the stress tensor in the momentum equations. (Note that the coefficients are the difference of two terms, and that with the usual values for the parameters (4) there is significant cancellation.)

The free surface stays horizontal, (23a), because there are no horizontal gradients until we look at order ϑ effects in a later section.

4.2. Convergence in the artificial parameters. One limitation on the accuracy of the above model is that even within the k-ϵ model of turbulence the coefficients are only approximate. This is due to both the modification of the free surface boundary conditions on u and k to (14)–(15) and the introduction of λ in (13). Although setting $\gamma = 1$ recovers the original boundary conditions (6)–(7) and $\lambda = 1$ recovers the original k-ϵ model, there is no certainty that this will give a model which is a good approximation to the "true" system. In essence the coefficients in the model are multivariable Taylor series in γ and in λ. In this subsection we present evidence that these series converge for $\gamma = \lambda = 1$, and so we can form a reasonable model.

Arbitrarily high-order terms in the center manifold expansion may be computed in principle. Our computer algebra program currently is limited by memory and time constraints to about 8th order in γ and lower orders in other parameters.[2] This is only attainable by the simplification of setting the k-ϵ parameters to the conventional numerical values in (4). By executing the REDUCE program and discarding terms $\mathcal{O}\big(\delta^6, \gamma^8, \lambda^2, \vartheta\big)$, we discover more terms in the series in γ. Listed in Table 1 are the expansions of some of the coefficients appearing later in the models.

Look down the columns in the table and see that the coefficients in each series

[2] In some applications [31, 43, 55] such routine computations can be performed to 30th order and are convincingly used to show the convergence or otherwise of the series expansions.

generally decrease by at least a factor of two. This suggests that the radii of convergence of the series in γ are roughly two or more. Thus simply evaluating the series at $\gamma = 1$ is reasonably good—some are shown in the bottom line of the table.[3]

The convergence in the parameter λ is problematical because it seems to appear always in the combination $\lambda\eta^2\bar{\epsilon}/(\tilde{\nu}\bar{k})$, that is $\lambda\eta^2\bar{\epsilon}^2/\bar{k}^3$. Thus convergence depends upon the properties of the solution which are generally unknown beforehand. We suggest that truncating to linear terms in λ forms an adequate approximation. It seems at least self-consistent to do this as later homogeneous solutions, namely (30)–(33), show a balance for the relatively small value $\eta^2\bar{\epsilon}^2/\bar{k}^3 \approx 0.2$. Thus the nonlinear terms in λ are generally expected to have a negligible influence in most flows of interest.[4]

First, the series are summed for $\gamma = 1$ as discussed in the previous subsection. Then introducing

$$\tilde{\lambda} = \frac{\lambda\eta^2\bar{\epsilon}^2}{\bar{k}^3} = C_\mu\lambda\frac{\eta^2/\tilde{\nu}}{\bar{k}/\bar{\epsilon}} \propto \lambda\frac{\text{vertical mixing time}}{\text{turbulent eddy time}}\,, \tag{25}$$

for brevity we write the model of decay of homogeneous turbulent flow as follows, with errors $\mathcal{O}\big(\delta^6, \lambda^2, \vartheta\big)$,

$$\frac{\partial\eta}{\partial t} = 0\,, \tag{26a}$$

$$\frac{\partial\bar{u}}{\partial t} = -(1.030 + 0.359\,\tilde{\lambda})\frac{\tilde{\nu}\bar{u}}{\eta^2} + (0.0504 - 0.243\,\tilde{\lambda})\frac{\tilde{\nu}\bar{u}^3}{\eta^2\bar{k}}\,, \tag{26b}$$

$$\frac{\partial\bar{k}}{\partial t} = -(0.0927 + 0.993\,\tilde{\lambda})\frac{\bar{k}^3}{\eta^2\bar{\epsilon}} + (0.589 + 0.516\,\tilde{\lambda})\frac{\tilde{\nu}\bar{u}^2}{\eta^2}\,, \tag{26c}$$

$$\frac{\partial\bar{\epsilon}}{\partial t} = -2.101\,\tilde{\lambda}\frac{\bar{k}^2}{\eta^2} + (1.552 - 3.215\,\tilde{\lambda})\frac{\tilde{\nu}\bar{\epsilon}\bar{u}^2}{\bar{k}\eta^2}\,. \tag{26d}$$

Note that setting $\lambda = 1$ to recover the original problem is just equivalent to using $\tilde{\lambda} = \eta^2\bar{\epsilon}^2/\bar{k}^3$. Observe that the first terms on the right-hand sides of the above equations represent decay terms through, for example, in the $\bar{u}$ and $\bar{k}$ equations, turbulent transport to the stream bed. The last terms in the $\bar{k}$ and $\bar{\epsilon}$ equations represent the production of turbulence through the velocity shear.

One feature of the model derived here is that it has no adjustable coefficients. All constants are derived from well-known physical parameters and accepted constants of the k-ϵ equations. Despite its relative complexity, the model has been systematically derived and the constants which appear are well defined. However, there are adjustable parameters, namely, the order of truncation of the series expansions. The model (26), for example, contains just the low-order terms in expansions in δ and λ.

[3]Actually, the introduction of the parameter a, and the selection of $a = 1/2$, was motivated by our original series in γ exhibiting singularities for $\gamma \approx -1$, as indicated by Domb–Sykes plots [21]. These singularities ruined the convergence at $\gamma = 1$. However, an Euler transform of the series to accelerate convergence is precisely equivalent to choosing nonzero a, and a few numerical experiments lead to $a \approx 1/2$ causing good convergence.

[4]We noticed in simulations that if ever $\eta^2\bar{\epsilon}^2/\bar{k}^3$ happened to become as large as approximately 3 at any point in x and t, then the dynamics became rapidly unstable. It may be that higher-order terms in the parameter λ, perhaps formed into a Padé approximant, could stabilize such a local instability. However, this aspect has not been explored as it is likely to involve infeasible amounts of algebraic computation.

4.3. Dynamics of spatial structure. The leading-order effect of horizontal gradients, such as that due to a sloping free surface, is found by computing terms of order ϑ in the asymptotic expansions. We describe these in this subsection.

Dominantly, horizontal gradients affect the velocity and pressure fields. By computing terms to order $\delta^3\vartheta$ we find that the velocity fields (22a)–(22b) are modified to

$$v = -\zeta^{4/3}\eta\frac{\partial\bar{u}}{\partial x} + \mathcal{O}\big(\delta^6 + \lambda^3 + \gamma^3 + \vartheta^3\big), \tag{27a}$$

$$u = \cdots + 3v_3(\zeta)\frac{g\eta^2}{\tilde{\nu}}\frac{\partial\eta}{\partial x} + \mathcal{O}\big(\delta^6 + \lambda^3 + \gamma^3 + \vartheta^3\big), \tag{27b}$$

where the $\cdots$ indicate the terms on the right-hand side of (22b) and where $v_3(\zeta)$ is drawn in Figure 2. The shape of v is required by the continuity equation. The modification to u asserts reasonably that at low levels of turbulence, large $1/\tilde{\nu}$, horizontal accelerations through decreasing depth, $\eta_x < 0$, cause the fluid to respond with a flatter profile through a subtraction of $v_3(\zeta)$ from $v_0(\zeta)$, as seen in Figure 2.

The structure of the fields within the fluid rapidly become more complicated at higher order. We do not detail the fields any more.

By executing the computer algebra program and discarding generated terms $\mathcal{O}\big(\delta^6, \gamma^6, \lambda^2, \vartheta^2\big)$, we discover first-order effects of horizontal variations with sufficient terms in the series in γ to sum them reliably for $\gamma = 1$. We find the same production and decay terms identified in (26) and, in addition, extra terms in the horizontal gradients. Using the accepted values (4) for the constants of the k-ϵ equations, we obtain the following model with our best estimates of its coefficients:

$$\frac{\partial\eta}{\partial t} \sim -\frac{\partial(\eta\bar{u})}{\partial x}, \tag{28a}$$

$$\begin{aligned}\frac{\partial\bar{u}}{\partial t} \sim{}& -(1.030 + 0.359\,\tilde{\lambda})\frac{\tilde{\nu}\bar{u}}{\eta^2} + (0.0504 - 0.243\,\tilde{\lambda})\frac{\tilde{\nu}\bar{u}^3}{\eta^2\bar{k}} \\ &+ \left[0.961 - 0.019\,\tilde{\lambda} - (0.019 - 0.087\,\tilde{\lambda})\frac{\bar{u}^2}{\bar{k}}\right] g\left(\theta - \frac{\partial\eta}{\partial x}\right) \\ &- (1.105 + 0.104\,\tilde{\lambda})\bar{u}\frac{\partial\bar{u}}{\partial x} - (0.032 - 0.056\,\tilde{\lambda})\frac{\bar{u}^2}{\bar{k}}\frac{\partial\bar{k}}{\partial x} \\ &+ (0.025 - 0.041\,\tilde{\lambda})\frac{\bar{u}^2}{\bar{\epsilon}}\frac{\partial\bar{\epsilon}}{\partial x},\end{aligned} \tag{28b}$$

$$\begin{aligned}\frac{\partial\bar{k}}{\partial t} \sim{}& -0.0927\frac{\bar{k}^3}{\eta^2\bar{\epsilon}} - 0.993\,\bar{\epsilon} + (0.589 + 0.516\,\tilde{\lambda})\frac{\tilde{\nu}\bar{u}^2}{\eta^2} \\ &- (0.025 + 0.011\,\tilde{\lambda})g\bar{u}\left(\theta - \frac{\partial\eta}{\partial x}\right) \\ &- (1.106 - 0.065\,\tilde{\lambda})\bar{u}\frac{\partial\bar{k}}{\partial x} - (0.030 + 0.056\,\tilde{\lambda})\bar{k}\frac{\partial\bar{u}}{\partial x} \\ &+ (0.025 - 0.060\,\tilde{\lambda})\frac{\bar{u}\bar{k}}{\bar{\epsilon}}\frac{\partial\bar{\epsilon}}{\partial x},\end{aligned} \tag{28c}$$

$$\begin{aligned}\frac{\partial\bar{\epsilon}}{\partial t} \sim{}& -2.101\frac{\bar{\epsilon}^2}{\bar{k}} + (1.552 - 3.215\,\tilde{\lambda})\frac{\tilde{\nu}\bar{\epsilon}\bar{u}^2}{\bar{k}\eta^2} \\ &+ \left(-0.006 + 0.562\,\tilde{\lambda}\right) g\frac{\bar{u}\bar{\epsilon}}{\bar{k}}\left(\theta - \frac{\partial\eta}{\partial x}\right)\end{aligned}$$

$$- 0.173\,\tilde{\lambda}\bar{\epsilon}\frac{\partial \bar{u}}{\partial x} + 0.533\,\tilde{\lambda}\frac{\bar{\epsilon}\bar{u}}{\bar{k}}\frac{\partial \bar{k}}{\partial x} - (1 + 0.735\,\tilde{\lambda})\bar{u}\frac{\partial \bar{\epsilon}}{\partial x}. \tag{28d}$$

We expect that the coefficients in the above equations, when considered as a model of the k-ϵ equations given in section 2, are accurate as shown. Except for the surface equation (28a), the first line in each equation is the same as in the horizontally homogeneous model (23); subsequent lines detail the additional terms needed to begin modelling long waves. A simpler version of the above model, obtained by omitting terms with small coefficients, is recorded in the introduction as the model (1).

Equation (28a) is an exact statement of the conservation of water and is not modified by any higher-order effects. To order $\delta^3\vartheta$, it may be written as

$$\frac{\partial \eta}{\partial t} + h\frac{\partial \bar{u}}{\partial x} = 0, \tag{29a}$$

which is a linear description of the conservation of water. Similarly, with $\theta = 0$ and in very low levels of turbulence ($\tilde{\nu} \approx 0$), the horizontal momentum equation (28b) may be written to order $\delta^3\vartheta$ as

$$\frac{\partial \bar{u}}{\partial t} = -0.961\,g\frac{\partial \eta}{\partial x}\,. \tag{29b}$$

This describes the horizontal acceleration due to slope of the fluid surface. These last two coupled equations form a standard description of linear wave dynamics except for one remarkable feature: the effect of gravity is reduced by the factor 0.961. For example, this would predict that even low levels of turbulence reduce the phase speed of waves by about two percent. As in thin films of viscous fluid [44], the phenomenon is due to the response of the fluid, approximately $v_0(\zeta)$ shown in Figure 2, being at an angle to the forcing 1 (either due to gravity or horizontal pressure gradients) when considered in the space of functions on $[0, \eta]$. Consequently, the forcing is less effective. Such a depression in phase speed may be observable in the propagation of long waves on turbulent flow.[5]

Returning to the order δ^5 momentum equation (28b), we note several interesting effects:

- The first line contains the turbulent drag terms identified in the previous subsection.
- The second lines describe the effects of surface and bed slope. Within the square brackets
  - the first term gives the depression of wave speed discussed above;
  - the second term very weakly enhances the phase speed correction in turbulent flow;
  - whereas it is difficult to ascribe one definite cause to the last term, coefficient modifications of the form $\bar{u}^2/\bar{k}$ are common in this model and reflect the relative importance of the turbulence on the mean flow.
- The third and fourth lines are dominated by the nonlinear advection term $\bar{u}\bar{u}_x$, with coefficient approximately 1.1. This coefficient is larger than 1 because of the shear: the maximum $u(y) > \bar{u}$ advects itself faster than $\bar{u}$. This third line also shows small "cross-talk" effects in the advection through the $\bar{u}^2\left(\log\bar{k}\right)_x$ and $\bar{u}^2\left(\log\bar{\epsilon}\right)_x$ terms.

[5]This modelling approach shows that where there is vertical or cross-sectional structure, depth-averaging or cross-sectional integration is generally unsound as a modelling tool. The reason is that it is the size and structure of the dynamical modes which determine the evolution (here approximately cube-root) and not the particular amplitudes used to measure the motion (here depth averages).

The dynamics of k and ϵ averages are given by equations (28c)–(28d).

- The first lines of each equation are the same turbulent production and decay terms identified in the previous subsection.
- The next line in each equation may arise from the modification of the turbulent production through the change of the velocity profile, seen in (27b), due to horizontal acceleration.
- The remaining terms simply represent horizontal advection by the fluid velocity. Note that different properties are advected at different effective speeds[6] as indicated by the different coefficients of the $\bar{u}\partial/\partial x$ terms—1.1 for $\bar{u}$ and $\bar{k}$, and 1 for $\bar{\epsilon}$.

The model, (1), reported in the introduction is a simplified version of (28). The solutions described in the next section show that the terms neglected from (28) in writing (1) are relatively small, contributing at most a few percent in the numerical balance of the terms, and so may be neglected at least for initial exploration.

The model (28) is purely hyperbolic, so a spatial diffusion is incorporated into (1) to help stabilize simulations. It was physically appealing to incorporate the turbulent diffusion $(\tilde{\nu}\bar{u}_x)_x$ into the $\bar{u}$ equation, and similarly for $\bar{k}$ and $\bar{\epsilon}$. The coefficients of these terms were determined by executing the computer algebra program to higher order in spatial derivatives but lower order in the artificial parameter λ. That the coefficients are larger than that of turbulent diffusion in the k-ϵ-model is due to the same process as that giving the enhanced Trouton vicosity in laminar flows, e.g., [38]. Such diffusion made hardly any difference to the simulations as a whole yet usefully avoided the generation of unphysical and ruinous spikes in the numerical solution. Higher-order terms in lateral derivatives may be able to refine the long-wave expansion employed here and controlled by the parameter ϑ. Although the long-wave approximation has been shown to converge in some simple dynamical circumstances [31, 43, 55], such higher-order derivative terms may easily destabilize the model (1), would involve enormous algebra to compute, and probably add little to the structurally stable model (1). An important parameter is the rate of attraction to the center manifold of the model, here locally proportional to $\tilde{\nu}/\eta^2$: when this is large the long-wave approximation is expected to be very good. Fortunately, in this application relatively rapid lateral variations are closely associated with the generation of turbulence; hence the local eddy diffusivity $\tilde{\nu}$ is typically large exactly where needed to resolve the lateral structure. Thus in this work we truncate the model (1) to the lateral effective diffusion terms.

5. Predictions of the new model. In this section we investigate some of the predictions the newly derived model (1) might make. We look at decaying turbulence, uniform flow on slopes, approximation to the St Venant equations, and a dam break simulation.

5.1. Decaying turbulence. Homogeneous turbulence decays algebraically. If there is no slope ($\theta = 0$), no variations in x, no mean flow ($\bar{u} = 0$), and no surface waves ($\eta = \text{const}$), then it is consistent to seek solutions of the model (1) in the form $\bar{k} \propto t^{-2}$ and $\bar{\epsilon} \propto t^{-3}$. Substituting and solving for the constants of proportionality, the model 1, or (28), predicts the turbulence decays according to

$$\bar{k} \sim 8.97\,\eta^2 t^{-2}\,, \quad \bar{\epsilon} \sim 12.8\,\eta^2 t^{-3}\,, \quad \tilde{\nu} \sim 0.565\,\eta^2 t^{-1}\,, \quad \tilde{\lambda} \sim 0.227 \tag{30}$$

[6] Though due to the nonlinear interaction terms we should really report on the speeds associated with the characteristics of the equations.

for large time t. The turbulence ultimately decays with the balance $\bar{\epsilon} \sim 0.48\,\bar{k}^{3/2}/\eta$.

However, the transients before this large time behavior may be long. There are two regimes of interest characterized by large and small $\tilde{\lambda}$ compared to 0.227. (Recall from (25) that $\tilde{\lambda}$ is the ratio of the vertical mixing time to the turbulent eddy mixing time.)

- For small $\tilde{\lambda}$, high turbulence $\bar{k}$, and low dissipation $\bar{\epsilon}$, the dissipation is roughly constant, actually
$$\bar{\epsilon} \approx \frac{1}{\sqrt{\bar{\epsilon}_\infty^{-2} + 15.1/k^3}}\,,$$
as the turbulence decays to (30) on a time-scale of approximately $2(\eta^2/\bar{\epsilon})^{1/3}$.
- For large $\tilde{\lambda}$, the vertical mixing time is relatively rapid and the turbulence decays with a different power law for some time. We find that $\bar{\epsilon} \propto \bar{k}^{2.11}$, which is only a little different from (30). The rate of decay towards (30) is relatively slow,
$$(\bar{k}, \bar{\epsilon}) \approx A\left(t^{-0.90}, 0.90\,t^{-1.90}\right),$$
and forms a long lasting transient.

The above results are for a stationary fluid. Instead, if the fluid is moving with uniform velocity on a horizontal bed, then the characteristics of the decaying bulk motion and turbulence are different in detail. We seek solutions of the model (1) in the form $\bar{k} \propto t^{-2}$ and $\bar{\epsilon} \propto t^{-3}$, as before, but now with $\bar{u} \propto t^{-1}$. Substituting and solving for the constants of proportionality, the model (1) may be rewritten as a generalized eigenvalue problem for $\tilde{\lambda}$. It is then straightforward to determine that the only positive solution is

$$\begin{aligned} &\bar{u} \sim 6.24\,\eta t^{-1}, \quad \bar{k} \sim 23.2\,\eta^2 t^{-2}, \quad \bar{\epsilon} \sim 49.6\,\eta^2 t^{-3}, \\ &\tilde{\nu} \sim 0.98\,\eta^2 t^{-1}, \quad \tilde{\lambda} \sim 0.198. \end{aligned} \tag{31}$$

Numerical solutions show that there are long lasting transients of a similar nature to those mentioned above for a stationary fluid. We do not elaborate further as this class of solutions is less likely to be of interest in applications.

5.2. Roll waves on turbulence flow down a slope. Water flowing down a slope generates turbulence that provides the drag to balance the gravitational forcing. But if we suppose only the turbulent parameters are in quasi equilibrium, but not the speed, then a St Venant approximation is obtained to the flow dynamics. When the flow is fast enough and the turbulence weak enough, we see the spontaneous development of turbulent roll waves. For flow down a slope the turbulent roll waves appear with a finite wavelength (see Figure 4), whereas the St Venant approximation predicts an infinite wavelength [34].

Let the downward bed slope be $\theta \neq 0$, but as in the previous subsection assume there are no variations in x, that is, just a mean flow ($\bar{u} \neq 0$) with no surface waves ($\eta = \text{const}$). Then it is consistent to seek solutions of the model (1) in the form $\bar{u} \propto \eta^{1/2}\theta^{3/2}$, $\bar{k} \propto \eta\theta$, and $\bar{\epsilon} \propto \eta^{1/2}\theta^{3/2}$. Substituting and solving for the constants of proportionality leads to a nonlinearly perturbed eigenvalue problem for $\tilde{\lambda}$ which is solved iteratively to give

$$\begin{aligned} &\bar{u} \approx 3.11\,\eta^{1/2}(g\theta)^{1/2}, \quad \bar{k} \approx 2.16\,\eta g\theta\,, \quad \bar{\epsilon} \approx 1.36\,\eta^{1/2}(g\theta)^{3/2}, \\ &\tilde{\nu} \approx 0.308\,\eta^{3/2}(g\theta)^{1/2}, \quad \tilde{\lambda} \approx 0.184. \end{aligned} \tag{32}$$

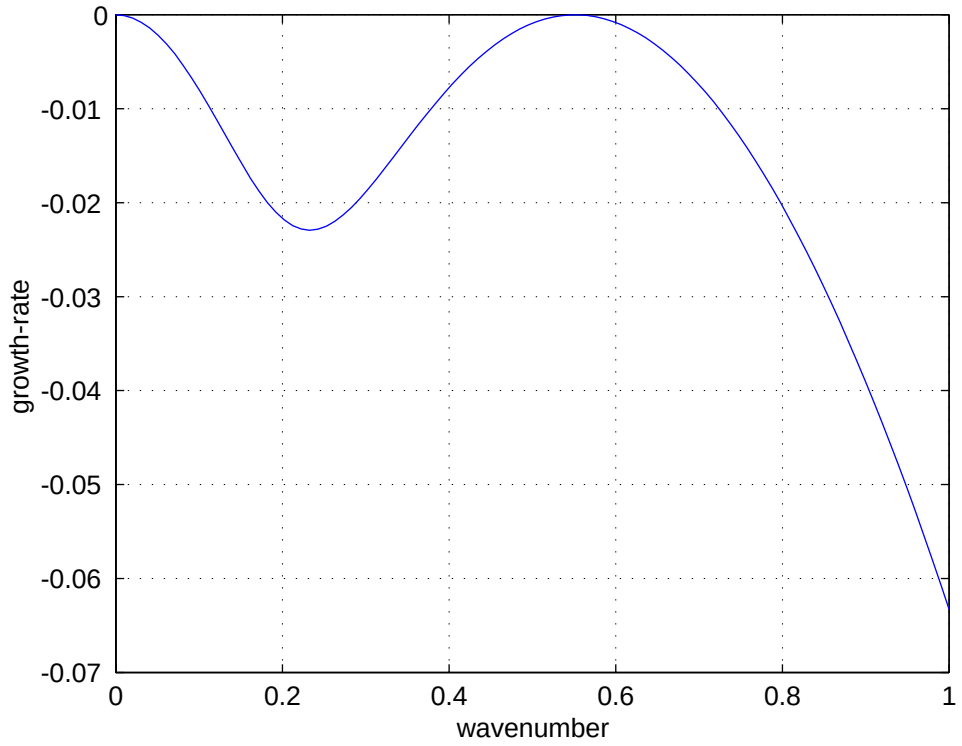


FIG. 4. *The leading growth-rate, the real part of the eigenvalue α, for turbulent roll waves down the critical slope $\theta = 0.602$ showing that an instability will first arise at finite wavenumber l.*

We expect this flow to be established on a time-scale of the vertical mixing, which is

$$T_{\rm mix} = \frac{\eta^2}{\tilde{\nu}} \approx 4\sqrt{\frac{\eta}{g\theta}}.$$

The flow (32) appears to have a high level of turbulence consistent with flow along a very rough channel.

Another interesting balance occurs when we assume that the production of turbulence parameters, $\bar{k}$ and $\bar{\epsilon}$, equals their dissipation through natural dissipation and bed drag. This leads to a reduced model in the form of St Venant's equation used in open channel flow, e.g., [34]. Assume that at all times the production and dissipation of $\bar{k}$ and $\bar{\epsilon}$ are given in the first lines on the right-hand sides of (1c) and (1d). That is, assume that the bed slope is small enough and that the flow is evolving slowly enough that spatial and temporal gradients are negligible. Then seek a balance with $\bar{k} \propto \bar{u}^2$ and $\bar{\epsilon} \propto \bar{u}^3$ to find

$$\bar{k} \approx 0.224\,\bar{u}^2, \quad \bar{\epsilon} \approx 0.0453\,\bar{u}^3/\eta, \quad \tilde{\nu} \approx 0.0992\,\eta\bar{u}, \quad \tilde{\lambda} \approx 0.184. \tag{33}$$

With this balance the momentum equation (1b) becomes

$$\frac{\partial \bar{u}}{\partial t} = -0.100\frac{\bar{u}^2}{\eta} + 0.96\,g\left(\theta - \frac{\partial \eta}{\partial x}\right) - 1.11\,\bar{u}\frac{\partial \bar{u}}{\partial x} + \frac{\partial}{\partial x}\left(0.143\,\bar{u}\eta\frac{\partial \bar{u}}{\partial x}\right), \tag{34}$$

which has exactly the same form as St Venant's equation for open channel flow except for the longitudinal diffusion term $\partial_x(\nu_0\bar{u}_x)$ for coefficient $\nu_0 = 0.143\,\bar{u}\eta$. Such a diffusion term, with previously unknown but assumed constant diffusivity, was included in St Venant's model by Needham and Merkin [34]—that $\nu_0 > \tilde{\nu}$ is due to the same process as that giving the Trouton vicosity in laminar flows, e.g., [38]. The three other coefficients are worthy of comment: the self-advection coefficient of (1.11) accounts for the vertical nonuniformity of the velocity profile with mean $\bar{u}$; the influence of gravity is reduced to $0.96\,g$ because, as explained earlier for (29b), the response of the fluid flow is not constant in the vertical so some of gravitational forcing is not used; and lastly the bed drag $\bar{u}^2/\eta$ has coefficient 0.100 which is larger than typical values. However, note that such drag coefficients have to vary depending upon the roughness of the channel bottom as often expressed by different roughness coefficients in Manning's law, e.g., [6, p. 246] or [23, p. 137]. We surmise that the flows we describe with model (1) have strong mixing in the vertical due to strong turbulence generated by a rough channel bed or other extremely turbulent flows such as breaking waves or dam spillways.

Needham and Merkin [34] analyzed a model close to (34) and deduced that the balanced turbulent flow (33) became unstable if the Froude number $F > 4$. The instability developed into roll waves propagating along the flow; long waves were most unstable. Here the Froude number $F \approx 0.96\,\theta/0.1$ and so we might expect roll waves to develop on flow down very steep slopes, $\theta > 0.42$, because of the effectiveness of turbulent damping on lesser slopes. Analysis of the full model (1), rather than the St Venant approximation (34), shows a slightly different picture. Seeking solutions to the model (1) linearized about the equilibrium flow (32) for nondimensional depth $\eta = 1$ and proportional to $\exp(ilx + \alpha t)$ leads to an eigenproblem for the complex eigenvalue α as a function of wavenumber l and the slope θ. Plotted in Figure 4 is the growth-rate, the real part of the eigenvalue α, as a function of wavenumber l for the critical slope $\theta_c = 0.602$: observe the zero growth-rate at wavenumber $l = 0.551$; for larger slopes θ, $\Re\{\alpha\}$ becomes positive here. In contrast to the St Venant model, the unstable roll waves are here predicted to have a finite wavelength $2\pi/l = 11.4$. The roll waves have similar shapes to that found by Needham and Merkin [34] and for many other roll waves on a fluid: see in Figure 5 the steepening at the front of the roll wave and the relatively longer tail. Here see we predict the turbulent intensity $\bar{k}$ is maximum a little behind the peak of the wave. In contrast to Needham and Merkin's model, we found no evidence for subcritical roll waves in our simulations. One point of interest arose in the eigenvalue computations: over a wide range of slopes and wavenumbers (< 1) we observed that the leading three eigenvalues were generally well separated from the fourth. For example, for slope $\theta = 0.1$, three eigenvalues had $\Re\{\alpha\} \approx -0.2$ whereas the other had $\Re\{\alpha\} \approx -0.9$. This suggests that for many purposes a three mode model of turbulent flow may be sufficient rather than the four mode model (1) derived here. We leave this for further research.

5.3. Simulate a dam breaking. One of the canonical flows of shallow water occurs after a dam breaks. Here we simulate such a flow and resolve the water slumping downstream and becoming extremely turbulent as it does so. For simplicity we use the model (1) reported in the introduction.

Imagine a dam at $x = 0$ initially holding back water of nondimensional depth $\eta = 1$. At time $t = 0$ the dam breaks and releases the water to rush downstream. To avoid overly poor conditioning in the numerics we let the water in front of the dam be of depth $\eta = 0.1$ (all quantities will be nondimensional so that in effect $g = 1$).

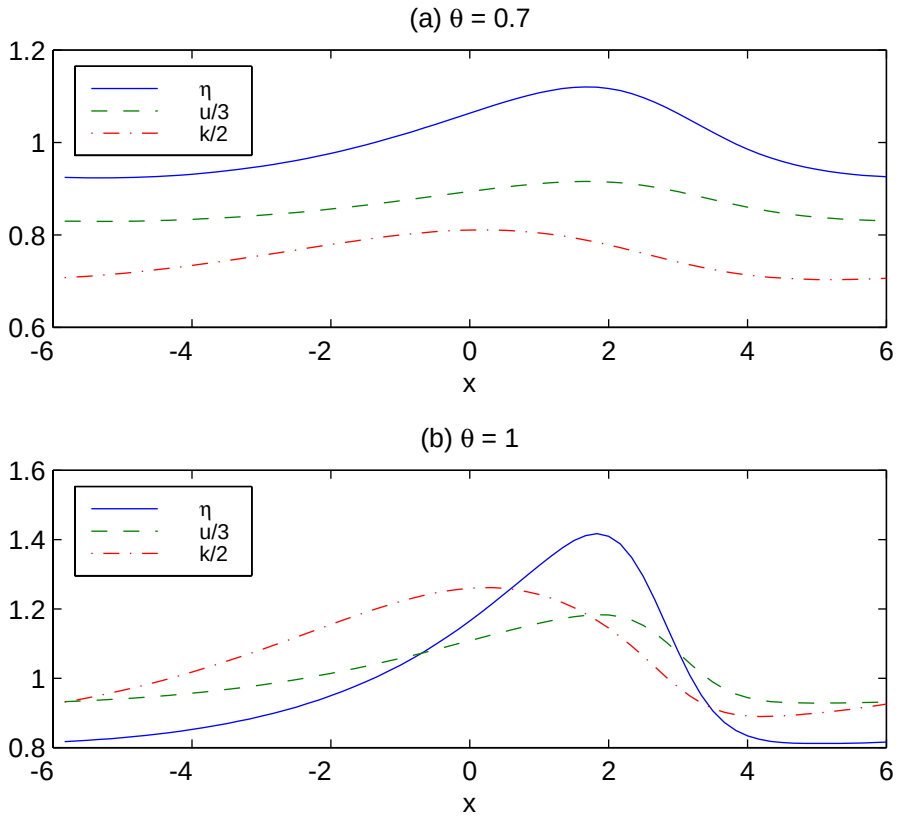


FIG. 5. *Profile of two roll waves on different slopes obtained by numerical simulation of model* (1) *until the solutions settled on a steadily propagating wave: solid, fluid depth* η*; dashed, mean velocity* $\bar{u}/3$*; and dot-dash, mean turbulent intensity* $\bar{k}/2$.

Also, to smooth the initial few time steps, we actually set the initial depth η to a tanh profile that smoothly varies between these extremes such that the water slope was a maximum of 2 (rather large under the slowly varying assumption) at the dam. The water is assumed initially quiescent, $\bar{u} = 0$ throughout, and has a low level of turbulence, somewhat arbitrarily chosen to be $\bar{k} = 0.0001$. Turbulent dissipation is initially set such that $\tilde{\lambda} = 0.227$ so that the balance of decaying turbulence (30) holds throughout.

The model (1) is simply discretized on a regular grid in space-time with a time step of $\Delta t = 1/10$ and space step of $\Delta x = 1/10$. The equations are discretized using a stencil 3 points wide in space using second-order accurate centered differencing in space. The resulting ordinary differential equations together with appropriate boundary conditions are integrated forward in time (with constant time step) as a set of differential algebraic equations using the robust, second-order, backward difference solver of [49]. The domain of simulation extended from six dam heights to the left behind the dam to six dam heights downstream. We integrated over a time $t = 6$, which is long enough for the disturbance to nearly reach the ends of the computational domain (linear waves on the dammed water having nondimensional speed 1).

The results of this simulation are shown in Figures 6 and 7. Observe that when

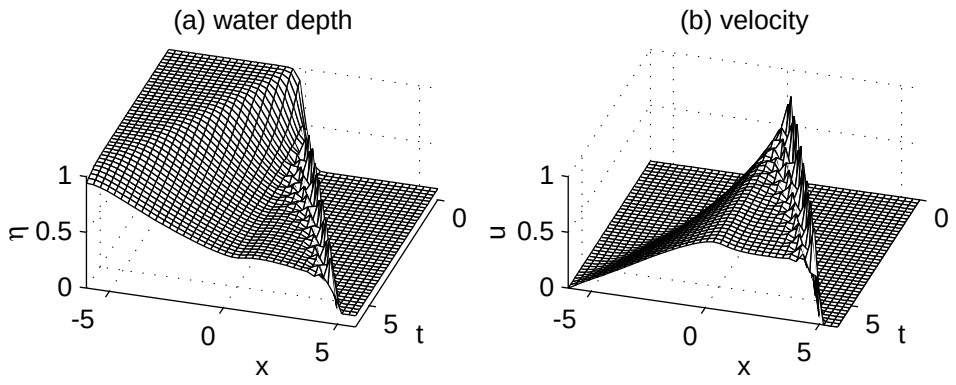


FIG. 6. *Simulation of dam breaking showing* (a) *the water depth* η *and* (b) *the mean downstream velocity* $\bar{u}$. *Observe the formation of a bore with superposed waves.*

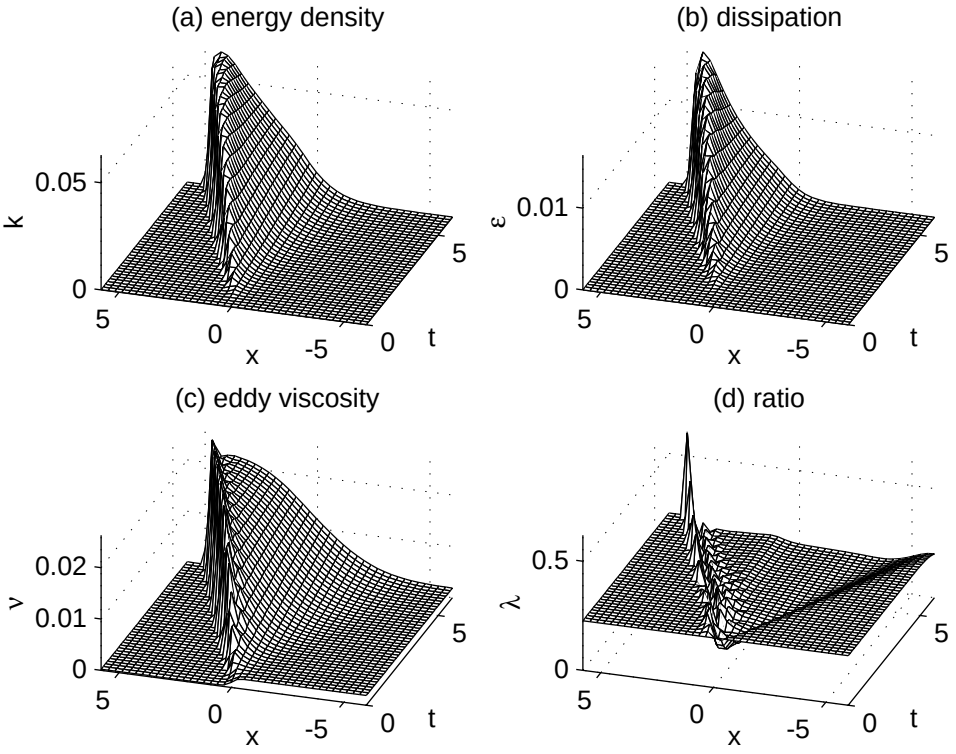


FIG. 7. *Simulation of dam breaking showing the time evolution of the turbulence parameters (note the view point is rotated from Figure* 6*):* (a) *the turbulent energy density* $\bar{k}$ *is highest just a little behind the bore and then tails away;* (b) *the turbulent dissipation* $\bar{\epsilon}$ *behaves similarly;* (c) *the turbulent eddy viscosity* $\tilde{\nu}$ *is greatest at the front of the bore; and* (d) *the parameter* $\tilde{\lambda}$, *apart from a peak at the front of the bore, is generally depressed from the decaying balance value of* 0.227 *in* (30).

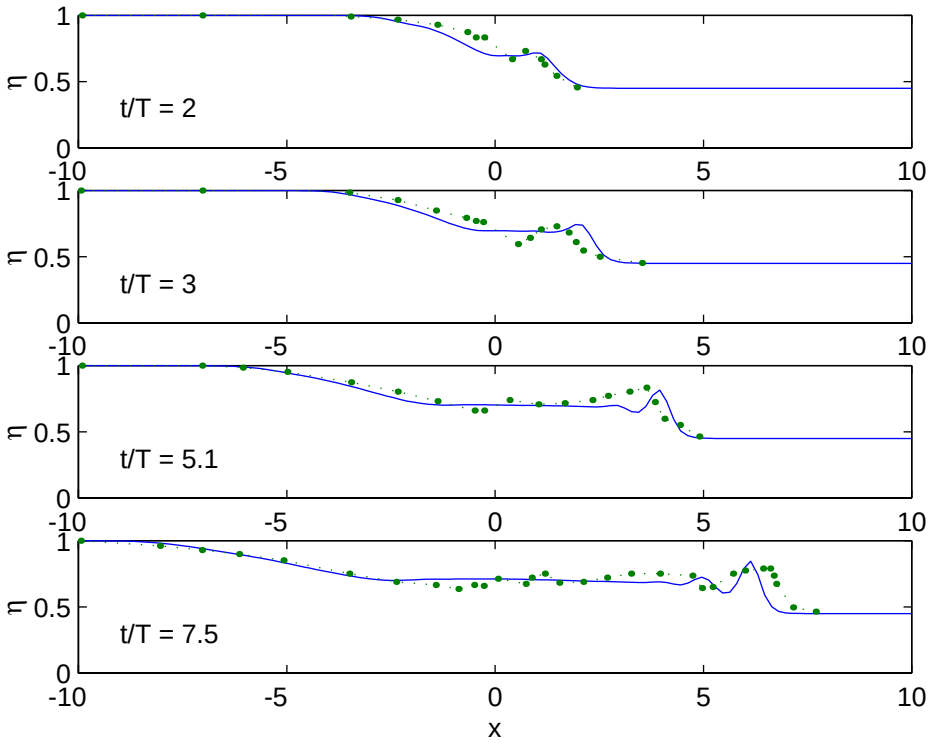


FIG. 8. *Compare model* (1) *(solid lines) with the experiments of Stansby, Chegini, and Barnes* [53, Fig. 8(c), p. 423] *(dots), where water of depth* 10 *cm* ($\eta = 1$) *behind a dam rushes forward over water of depth* 4.5 *cm* ($\eta = 0.45$) *after the dam (at* $x = 0$*) breaks. The time-scale* $T = \sqrt{H/g} =$ 0.1 *secs.*

the dam breaks, the water slumps down and rushes downstream in a turbulent bore. The bore appears undular but may be evolving towards a series of solitary waves—they cannot be differentiated on this time-scale. The turbulent structure shown in Figure 7 has interesting features. Apparently the energy density peaks a little behind the bore, then decays approximately linearly with distance. It appears that up to time $t = 7$ the generation of turbulence is still significantly greater than its decay as the peak is still growing. The turbulent dissipation and eddy viscosity behave similarly, though the eddy viscosity appears to peak much closer to the front of the turbulent bore. The parameter $\tilde{\lambda}$ plotted in Figure 7(d) reaffirms that relatively small values appear relevant to such flows.[7] The peak of $\tilde{\lambda}$ at the front of the bore predicts that there is a lot of vertical mixing at the front, but less so behind the bore where $\tilde{\lambda}$ is smaller. All of the above seem physically reasonable.

A further similar simulation matches the experiments of Stansby, Chegini, and Barnes [53]. In Figure 8 we plot a comparison of the water depth between the model (1) and the experiment reported in Figure 8(c) of Stansby, Chegini, and Barnes.[8] The model (1) was solved with $\Delta x = 1/6$, $\Delta t = 1/10$ and initially η a

[7] We find that our model (1) becomes quickly unstable if ever the parameter $\tilde{\lambda} > 3$ approximately.

[8] Unfortunately it does not seem reasonable to compare with the other two experiments of Stansby, Chegini, and Barnes [53] as in both it appears from our digitization that water is not conserved by

tanh profile, $\bar{u} = 0$, $\tilde{\lambda} = 0.227$, and $\bar{k} = 10^{-4}$, except for a bump in $\bar{k}$ to 0.018 to represent the initial plunging jet seen in the experiments. The location of the turbulent bore is reasonably well predicted. It appears that the height of the undulations in the model's bore is approximately that of the turbulent fluctuations seen in the water level.

This model that we have derived and solved in some cases of interest explicitly accounts for the spatio-temporal variations of the intensity and broad nature of the turbulence underlying the flow of shallow water.

Appendix A. Comments on theory in this application. This appendix addresses the connection to the rigorous theory of center manifolds in this application. We emphasize that throughout this paper we describe the application of center manifold techniques and *not* the application of the rigorous center manifold theory. There are two main reasons for this which we elaborate on below.

First, here we construct an infinite dimensional center manifold. At each point in x there are four degrees of freedom, parametrized by η, $\bar{u}$, $\bar{k}$, and $\bar{\epsilon}$; but there are an infinitude of x positions and so there is an infinite number of degrees of freedom in the mathematical model. However, there is currently very little theory on infinite dimensional center manifolds appearing via slowly varying approximations, e.g., [16] and what there is does not rigorously apply here, nor does it apply directly to many other physically interesting models such as dispersion in pipes [32], laminar long-wave, thin-film flows [44], and the dynamics in flow reactors [5]; the principal reason is that $\partial/\partial x$ is an unbounded operator, because of its small scale characteristics, whereas the slowly varying approximation treats $\partial/\partial x$ as small. Thus we use the formal technique of constructing complete low-dimensional models [10, 39, 40, 41, 42], techniques suggested and developed by standard applications of the theory. We expect that eventually theory will be developed which supports the application of center manifold concepts to slowly varying approximations.

But there is a second obstacle to supporting this model with theory. In standard applications of center manifold theory the nonlinear terms in the original problem are required to be smooth in the neighborhood of the equilibrium under consideration (here the origin, a state of no flow, and no turbulence). However, here many nonlinear terms are definitely not smooth; for example, turbulent dispersion terms such as $\frac{\partial}{\partial x}\left(C_\mu \frac{k^2}{\epsilon}\frac{\partial u}{\partial x}\right)$ and interaction terms such as $C_{\epsilon 2}\frac{\epsilon^2}{k}$ are unbounded as $(u, k, \epsilon) \to \mathbf{0}$. In rigorous applications of center manifold theory, one may choose the various critical modes and parameters to have any set of relative orders of magnitude. The resulting asymptotic expressions are the same [39], it is only the sequence in which the terms appear that changes with a change in relative orders of magnitude. Indeed, this reflects a very desirable property of a modelling procedure, namely, that the results are essentially independent of arbitrary human-made assumptions (such as order of magnitude) in the analysis. However, in this application the highly nonsmooth nature of the original equations means that in order to apply the center manifold techniques we need to choose carefully the various orders of magnitudes of the variables and parameters, via (17). The aim, as in all asymptotic analyses, is to obtain a tractable and physically relevant leading order problem. The techniques of center manifold

up to 6–7%, compared to better than 1% for their Figure 8(c). This is a significant discrepancy which, if associated with the identification of the bore, would correspond to an error in its location of up to $\Delta x \approx 1.5$ (15 cm). The discrepancy is likely to be due to entrained air [53, p. 422] which we have not attempted to model, and which is likely to be insignificant in a real dam break [53, p. 407].

theorems are then applied, it is just that the current center manifold theorems cannot give *rigorous* justification.

Notwithstanding these theoretical limitations we consider that the systematic techniques are applicable because of the attractivity of the manifold of equilibria, $\mathcal{M}_0$, of the linear operator $\mathcal{L}$. In the spirit of center manifold theory, we claim that the "small nonlinear" terms on the right-hand side of (13) just perturb the shape of $\mathcal{M}_0$ to a nearby manifold $\mathcal{M}$ and perturb the evolution thereon. Thus our last task, and the one fulfilled in this appendix, is to prove the exponential decay to $\mathcal{M}_0$ at a nonzero level of turbulence. The "small nonlinear" terms will affect the rate of decay to $\mathcal{M}$, perhaps slowing the attraction in some regimes, but by continuity, since $\mathcal{M}_0$ is exponentially attractive, $\mathcal{M}$ will be attractive for small enough nonlinearity.

Linearizing the k-ϵ equations (13) about $\mathcal{M}_0$,

$$(p_0, \boldsymbol{u}_0) = \left(g(h-y), \tfrac{4}{3}\left(\tfrac{y}{h}\right)^{1/3} U(x,t), 0, H(x,t), \tfrac{4}{3}\left(\tfrac{y}{h}\right)^{1/3} K(x,t), E(x,t) \right), \tag{35}$$

we obtain $(0, \partial \boldsymbol{u}/\partial t) = \mathcal{L}(p, \boldsymbol{u})$, where the linear operator $\mathcal{L}$ is

$$\begin{bmatrix}
0 & 0 & \frac{\partial \cdot}{\partial y} & 0 & 0 & 0 \\
0 & \frac{\partial}{\partial y}\left(\nu_0 \frac{\partial \cdot}{\partial y}\right) & 0 & 0 & C_\mu \frac{\partial}{\partial y}\left(\frac{\partial u_0}{\partial y}\frac{2k_0}{\epsilon_0}\cdot\right) & -C_\mu \frac{\partial}{\partial y}\left(\frac{\partial u_0}{\partial y}\frac{k_0^2}{\epsilon_0^2}\cdot\right) \\
-\frac{\partial \cdot}{\partial y} & 0 & 2\frac{\partial}{\partial y}\left(\nu_0 \frac{\partial \cdot}{\partial y}\right) & 0 & -\frac{2}{3}\frac{\partial \cdot}{\partial y} & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \frac{C_\mu}{\sigma_k}\frac{\partial}{\partial y}\left(\frac{k_0^2}{\epsilon_0}\frac{\partial \cdot}{\partial y} + \frac{2k_0}{\epsilon_0}\frac{\partial k_0}{\partial y}\cdot\right) & -\frac{C_\mu}{\sigma_k}\frac{\partial}{\partial y}\left(\frac{\partial k_0}{\partial y}\frac{k_0^2}{\epsilon_0^2}\cdot\right) \\
0 & 0 & 0 & 0 & 0 & \frac{C_\mu}{\sigma_\epsilon}\frac{\partial}{\partial y}\left(\frac{k_0^2}{\epsilon_0}\frac{\partial \cdot}{\partial y}\right)
\end{bmatrix},$$

subject to boundary conditions

$$p + \frac{2}{3}k = 0 \quad \text{on } y = h\,, \tag{36}$$

$$u = 0 \quad \text{on } y = 0\,, \tag{37}$$

$$\frac{\partial u}{\partial y} - \frac{u}{3h} = 0 \quad \text{on } y = h\,, \tag{38}$$

$$v = 0 \quad \text{on } y = 0\,, \tag{39}$$

$$k = 0 \quad \text{on } y = 0\,, \tag{40}$$

$$\frac{\partial k}{\partial y} - \frac{k}{3h} = 0 \quad \text{on } y = h\,, \tag{41}$$

$$y^{2/3}\frac{\partial \epsilon}{\partial y} \to 0 \quad \text{as } y \to 0\,, \tag{42}$$

$$\frac{\partial \epsilon}{\partial y} = 0 \quad \text{on } y = h\,. \tag{43}$$

We seek solutions proportional to $\exp(\lambda t)$. The first thing to note is that we address a generalized eigenproblem

$$\mathcal{L}\begin{bmatrix} p_0 \\ \boldsymbol{u} \end{bmatrix} = \begin{bmatrix} 0 \\ \lambda \boldsymbol{u} \end{bmatrix}$$

as the first row of $\mathcal{L}$ comes from the continuity equation.

Thus the first row, with the boundary condition (39), gives $v = 0$ for any eigenvalue λ. Furthermore, the third row, from the vertical momentum equation with the

pressure boundary condition (36), then gives that $p = -\frac{2}{3}k$ for any λ. These considerations give no constraint on the eigenvalue λ. Ignoring the p and v components of the eigenproblem then leads to a standard eigenproblem—one where the operator is in block, upper-triangular form. Thus consider the components in turn, starting from the last, and we show that all eigenvalues must be nonpositive and thus $\mathcal{M}_0$ is attractive.

- For any eigenfunction $\boldsymbol{u}$, if turbulent dissipation $\epsilon \neq 0$, then

$$\frac{h^{4/3}}{T\sigma_\epsilon}\frac{\partial}{\partial y}\left(y^{2/3}\frac{\partial \epsilon}{\partial y}\right) = \lambda\epsilon\,, \tag{44}$$

where

$$T = \frac{9h^2}{16\tilde{\nu}}\,,$$

is the time-scale of cross-depth turbulent diffusion. Multiplying (44) by ϵ, $\int_0^h \cdots dy$, and integrating by parts we deduce

$$\lambda = -\frac{h^{4/3}}{T\sigma_\epsilon}\frac{\int_0^h y^{2/3}\left(\frac{\partial \epsilon}{\partial y}\right)^2 dy}{\int_0^h \epsilon^2 dy} \leq 0\,,$$

provided that the boundary conditions (43) (assuming $\epsilon \not\to \infty$ for $y \to h$) and

$$\epsilon\frac{\partial \epsilon}{\partial y} = o\big(y^{-2/3}\big) \quad \text{as } y \to 0 \tag{45}$$

are satisfied. The equality $\lambda = 0$ holds if and only if $\partial\epsilon/\partial y \equiv 0$, which leads to

$$\epsilon = E(x,t)\,,$$

a function independent of y. Since (44) indicates nontrivial solutions near $y = 0$ are of the form $\epsilon \sim A + By^{1/3}$, then the condition (45) is effectively equivalent to (42).

Further, apply Sturm–Liouville theory to (44) under boundary conditions (43) and (42). Changing the vertical variable from $y = hz^3$ to z, the eigenvalue problem becomes

$$\frac{\partial^2 \epsilon}{\partial z^2} = 9T\sigma_\epsilon\lambda z^2\epsilon\,,$$

(a form of Bessel's equation [1, (9.1.51)]) with the following normal separate boundary conditions:

$$\frac{\partial \epsilon}{\partial z} = 0 \quad \text{on } y = 0 \text{ and } y = h.$$

Applying standard Sturm–Liouville theory, see, for example, Birkhoff and Rota [8, pp. 296] or Hartman [20, pp. 337 and the following ones], we see that the eigenvalues are discrete and must tend to infinity:

$$0 = \lambda_1 > \lambda_2 > \cdots \to -\infty.$$

Thus, linearly, solutions in the neighborhood of the manifold $\mathcal{M}_0$ are attracted exponentially quickly to it (at a rate at least as fast as $\exp(\lambda_2 t)$).
Sturm–Liouville theory may be also applied directly for the u and k components to show that any eigenvalues associated primarily with them are discrete. We do not record the details in the following.

- Similarly, for any eigenfunction $\boldsymbol{u}$, if turbulent dissipation $\epsilon = 0$ but the turbulent energy $k \neq 0$, then

$$\frac{h^{4/3}}{T\sigma_k}\frac{\partial}{\partial y}\left(y^{2/3}\frac{\partial k}{\partial y} + \frac{2k}{3y^{1/3}}\right) = \lambda k. \tag{46}$$

Multiplying this by $y^{2/3}$ we rewrite it as

$$\frac{h^{4/3}}{T\sigma_k} y^{-1/3}\frac{\partial}{\partial y}\left[y^2\frac{\partial}{\partial y}\left(y^{-1/3}k\right)\right] = \lambda y^{2/3}k.$$

Multiplying by k, $\int_0^h \cdots dy$ and integrating by parts we deduce that

$$\lambda = -\frac{h^{4/3}}{T\sigma_k}\frac{\int_0^h y^2\left[\frac{\partial}{\partial y}\left(y^{-1/3}k\right)\right]^2 dy}{\int_0^h y^{2/3}k^2 dy} \leq 0,$$

provided (41) and

$$k = o\left(y^{1/12}\right) \quad \text{as } y \to 0 \tag{47}$$

are satisfied. The equality $\lambda = 0$ holds here if and only if $\frac{\partial}{\partial y}\left(y^{-1/3}k\right) \equiv 0$, which together with boundary condition (40) implies that

$$k = K(x,t)\tfrac{4}{3}\left(\tfrac{y}{h}\right)^{1/3}$$

with a function $K(x,t)$ independent of y. Since the indicial equation of (46) indicates that nontrivial solutions near $y = 0$ are of the form $k \sim Ay^{-2/3} + By^{1/3}$, then the boundary condition (47) is equivalent to (40).

- The only possible eigenvalue associated with nonzero η is 0.
- Last, for any eigenfunction $\boldsymbol{u}$, if $\epsilon = k = \eta = 0$ but the horizontal velocity $u \neq 0$, then

$$\frac{h^{4/3}}{T}\frac{\partial}{\partial y}\left(y^{2/3}\frac{\partial u}{\partial y}\right) = \lambda u, \tag{48}$$

and we rewrite this as

$$\frac{h^{4/3}}{T} y^{-1/3}\frac{\partial}{\partial y}\left[y^{4/3}\frac{\partial}{\partial y}\left(y^{-1/3}u\right)\right] = \lambda u.$$

Multiplying by u, $\int_0^h \cdots dy$ and integrating by parts we deduce that

$$\lambda = -\frac{h^{4/3}}{T}\frac{\int_0^h y^{4/3}\left[\frac{\partial}{\partial y}\left(y^{-1/3}u\right)\right]^2 dy}{\int_0^h u^2 dy} \leq 0,$$

provided (38) and

$$u = o\left(y^{1/6}\right) \quad \text{as } y \to 0 \tag{49}$$

are satisfied. Similarly to the ϵ and k equations, the equality $\lambda = 0$ holds here if and only if $\frac{\partial}{\partial y}\left(y^{-1/3}u\right) \equiv 0$. This and the boundary condition (40) yields a unique solution

$$u = U(x,t)\tfrac{4}{3}\left(\tfrac{y}{h}\right)^{1/3}$$

with $U(x,t)$ independent of y. Since the indicial equation of (48) indicates nontrivial solutions near $y = 0$ are of the form $u \sim A + By^{1/3}$, then the boundary condition (49) is effectively equivalent to (37).

We have proven that *if the boundary conditions* (36)–(43) *are satisfied, then, except for the four-fold eigenvalue zero whose eigenfunctions span* $\mathcal{M}_0$*, the eigenvalues of* $\mathcal{L}$ *are negative and bounded away from* 0. Thus we expect the manifold (35) is locally attractive. Further, the time-scale of this attraction is the cross-depth turbulent diffusion time-scale T.

Appendix B. Computer algebra constructs the model.

Here we list the REDUCE[9] computer algebra program used to derive the long-wave models of turbulent flow.

The algorithm is the iterative algorithm described in [46, 48] adapted to this difficult asymptotic problem. The program refines the description of the center manifold and the evolution thereon until the residual of the governing differential equations are driven to zero, to some asymptotic error. The key to the correctness of the results is then in the correct coding of the residuals—see inside the iterative loop.

Note that because the thickness of the film is continuously varying in space and time, and because of the cube-root structure in the vertical, it is convenient to work with equations in terms of a scaled vertical coordinate $z = \sqrt[3]{y/\eta}$ so that the free surface of the film is always $z = 1$. However, the turbulence equations are not explicitly rewritten in this new coordinate because the computer handles all the necessary details of the transformation.

```
COMMENT Constructs a model of turbulent 2D flow of shallow water flow
on a flat slope based on the k-epsilon turbulence dynamical equations.
Calculates the centre manifold & reduced dynamic system on it for the
k-\epsilon model with the following boundary conditions for u, v, p, k
and \epsilon: u=v=k=deps/dy=0 at y=0, dk/dn=deps/dn=0 on y=eta. Fiddle
the free-surface BC to linearly force u & k at the surface. This gives
roughly cube-root profile in the vertical structure of u & k, whereas
eps is roughly constant.  Write results in terms of z=(y/eta)^(1/3).
Here scale derivatives as ddx*del for better control.

Created 11/11/94, last modified 8/6/99;

% improve output appearance
on div; off allfac; on list; on revpri;
factor es,ks,eta,us,del,ddx,df,g;

% maximum order of calculation in del: linear=3; non-trivial=5
o:=5;
```

[9] At the time of writing, information about REDUCE was available from Anthony C. Hearn, RAND, Santa Monica, CA 90407-2138, USA (reduce@rand.org).

```
% truncate in parameters: d/dx; BC kludge; k-eps fudge
let { ddx^2=>0, gamm^4=>0, lamb^2=>0};
% let ddx=>0; gamm:=del^2*gam; lamb:=lam*del^2; % for initial results
cgam:=(1-gamm*1/2); % equivalent to an Euler transform of gam series
theta:=del^3*ddx*thet;

% turbulence constants---remove to get general formulae
C_m:=9/100; C_e2:=192/100; C_e1:=144/100; s_k:=1; s_e:=13/10;

% FOR ALL q SUCH THAT  q>o LET del^q=0$
procedure ignore_order_gt(o); begin
IF o=3 THEN LET del^4=0;
IF o=4 THEN LET del^5=0;
IF o=5 THEN LET del^6=0;
IF o=6 THEN LET del^7=0;
IF o=7 THEN LET del^8=0;
IF o=8 THEN LET del^9=0;
IF o=9 THEN LET del^10=0;
IF o=10 THEN LET del^11=0;
IF o>=11 THEN LET del^12=0;
end;

% amplitudes and their dependences  ( eta=h+hs )
depend us,x,t;
depend ks,x,t;
depend es,x,t;
depend hs,x,t;
let {df(us,t) => gu,
     df(hs,t) => gh,
     df(ks,t) => gk,
     df(es,t) => ge
     };

% since z=(y/eta)^1/3 we need the following for d/dx, d/dt & d/dy
etax:=del^2*del*ddx*df(hs,x);
procedure dfdx(a);
begin scalar aa,bb;
      aa:=a*del^3*ddx; bb:=a*del*ddx;
      return df(bb,x) +( df(aa,eta) -z/3/eta*df(aa,z) )*df(hs,x);
end;
procedure dfdt(a);
begin scalar aa,ugh;
      aa:=a*del^2*del*ddx; ugh:=if gh=0 then 0 else gh/(del*ddx);
      return df(a,t) +( df(aa,eta) -z/3/eta*df(aa,z) )*ugh ;
end;
depend z,y;
let df(z,y) => 1/(3*eta*z^2);
fs:={z=1}$

% procedures to solve for cross-stream structures
operator iav;  linear iav;
operator ise;  linear ise;
operator isk;  linear isk;
operator isu;  linear isu;
operator isp;  linear isp;
operator isv;  linear isv;
let {iav(z~^p,z) => 1/(p+1)
    ,ise(z~^p,z) => ( z^(p+2) -3/(p+5) )/(p+1)/(p+2)
    ,isk(z~^p,z) => ( z^(p+2) -z*4/(p+5) )/(p+1)/(p+4)
    ,isu(z~^p,z) => ( z^(p+2) -z*4/(p+5) )/(p+1)/(p+2)
    ,isp(z~^p,z) => ( z^(p+1) -1 )/(p+1)
    ,isv(z~^p,z) => ( z^(p+1) )/(p+1)
```

```
 81     ,iav(z,z) => 1/2
 82     ,ise(z,z) => ( z^3 -1/2 )/6
 83     ,isk(z,z) => ( z^3 -z*2/3 )/10
 84     ,isu(z,z) => ( z^3 -z*2/3 )/6
 85     ,isp(z,z) => ( z^2 -1 )/2
 86     ,isv(z,z) => ( z^2 )/2
 87     ,iav(1,z) => 1
 88     ,ise(1,z) => ( z^2 -3/5 )/2
 89     ,isk(1,z) => ( z^2 -z*4/5 )/4
 90     ,isu(1,z) => ( z^2 -z*4/5 )/2
 91     ,isp(1,z) => ( z -1 )
 92     ,isv(1,z) => ( z )
 93     };
 94 procedure mean(a);   3*iav(a*z^2,z);
 95 procedure solv_e(a); 9/16*9*eta^2*s_e/anu*ise(z^2*a,z);
 96 procedure solv_k(a); 9/16*9*eta^2*s_k/anu*isk(z^2*a,z);
 97 procedure solv_u(a); 9/16*9*eta^2/anu*isu(z^2*a,z);
 98 procedure solv_p(a); 3*eta*isp(z^2*a,z);
 99 procedure solv_v(a); 3*eta*isv(z^2*a,z);
100
101 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
102 % initial approximation for the iteration
103
104 cbrt:=(4/3)*z$
105 vu:=del^2*cbrt*us;
106 vv:=0;
107 vp:=g*eta*(1-z^3) +del^2*8/9*ks*(-z);
108 vk:=del^2*cbrt*ks;
109 ve:=del^3*es;
110 vnu:=c_m*vk^2/ve;;
111 vmu:=c_e2*ve^2/vk;
112 vph:=0;
113
114 gu:=0;
115 gh:=0;
116 gk:=0;
117 ge:=0;
118
119 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
120 % iterate
121 anu:=c_m*ks^2/es$ % scaled typical diffusion
122 sinth:=theta-theta^3/6+theta^5/120-theta^7/5040$
123 costh:=1-theta^2/2+theta^4/24-theta^6/720$
124 repeat begin
125 write "
126  NEXT ITERATION
127 ----------------";
128
129 begin scalar Eqc;
130 ignore_order_gt(o);
131 % continuity equation for v
132 Eqc:=dfdx(vu)+df(vv,y);
133 ok:=if Eqc=0 then 1 else 0;
134 vv:=vv-solv_v(Eqc);
135 end;
136
137 % kinematic equation for eta
138 gh:=SUB(fs, vv-vu*etax )/del^2;
139
140 % mu equation for mu of sufficient order
141 begin scalar Eqmu;
142 ignore_order_gt(o+3);
```

```
Eqmu:=(vmu*vk-C_e2*ve^2)$
ok:=if ok and(Eqmu=0) then 1 else 0;
vmu:=vmu-Eqmu/(del^2*cbrt*ks);
end;

% ph production equation of order m-1
begin scalar Eqph;
ignore_order_gt(o-1);
Eqph:=(vph-(df(vu,y)+dfdx(vv))^2-2*df(vv,y)^2-2*dfdx(vu)^2)$
ok:=if ok and(Eqph=0) then 1 else 0;
vph:=vph-Eqph;
end;

% epsilon equation of order m+1 & BC of order m
begin scalar Eqeps,BCe0,BCeh,gep;
ignore_order_gt(o+1);
Eqeps:= -dfdt(ve) -lamb*vmu +C_e1*(del*es/ks)*vnu*vph
       +1/s_e*df(vnu*df(ve,y),y) -vu*dfdx(ve)-vv*df(ve,y)
       +1/s_e*dfdx(vnu*dfdx(ve))$
BCe0:=del*sub(z=0,z^2*df(ve,y))$
BCeh:=del*sub(fs, df(ve,y)-etax*dfdx(ve) )$
ok:=if ok and(Eqeps=0)and(BCe0=0)and(BCeh=0) then 1 else 0;
gep:=+mean(Eqeps)-160/117*anu/eta*BCeh;
ve:=ve-solv_e( Eqeps-gep )/del
      +27*eta*es/(32*ks^2*C_m*del)*BCe0*((1-z)^2-1/10);
ge:=ge+gep/del^3;
end;

% nu equation of order m+1
begin scalar Eqnu;
ignore_order_gt(o+1);
Eqnu:=vnu*ve-C_m*vk^2$
vnu:=vnu-Eqnu/(del^3*es);
end;

% k equation & (BC of order m-1)
begin scalar Eqk,BCk0,BCkh,gkp;
ignore_order_gt(o);
Eqk:= -dfdt(vk) -lamb*ve +vnu*vph +1/s_k*df(vnu*df(vk,y),y)
       -vu*dfdx(vk)-vv*df(vk,y) +1/s_k*dfdx(vnu*dfdx(vk))$
BCk0:=del*sub(z=0, vk )$
BCkh:=del*sub(fs, (df(vk,y)-etax*dfdx(vk))*cgam -(1-gamm)*vk/3/eta )$
ok:=if ok and(Eqk=0)and(BCk0=0)and(BCkh=0) then 1 else 0;
gkp:=7/4*mean(z^3*Eqk) -28/9*anu/s_k/eta*BCkh;
vk:=vk+solv_k( -Eqk+cbrt*gkp )/del;
gk:=gk+gkp/del^2;
end;
% patch up nu again because of significant changes
% nu equation of order m+1
begin scalar Eqnu;
ignore_order_gt(o+1);
Eqnu:=vnu*ve-C_m*vk^2$
ok:=if ok and(Eqnu=0) then 1 else 0;
vnu:=vnu-Eqnu/(del^3*es);
end;

% v equation of order m-1 for p
begin scalar Eqv,BCph;
ignore_order_gt(o);
Eqv:=del*( -dfdt(vv) -df(vp,y) -g*costh -2/3*df(vk,y)
           -vu*dfdx(vv)-vv*df(vv,y) +dfdx( vnu*(df(vu,y)+dfdx(vv)) )
           +2*df(vnu*df(vv,y),y) )$
```

```
BCph:=del*sub(fs, (vp+2/3*vk)*(1+etax^2)
             -2*vnu*(df(vv,y)+dfdx(vu)-etax*(dfdx(vv)+df(vu,y)))
             )$
ok:=if ok and(Eqv=0)and(BCph=0) then 1 else 0;
vp:=vp+(solv_p(Eqv) -BCph)/del;
end;

% u equation for u  (& BC of order m-1)
begin scalar Equ,BCu0,BCuh,gup;
Equ:= -dfdt(vu) +g*sinth -dfdx(vp) -2/3*dfdx(vk)
     -vu*dfdx(vu)-vv*df(vu,y) +2*dfdx(vnu*dfdx(vu))
     +df(vnu*(df(vu,y)+dfdx(vv)),y)$
BCu0:=del*sub(z=0, vu )$
BCuh:=del*sub(fs, -(1-gamm)*vu/3/eta
              +cgam*((df(vu,y)+dfdx(vv))*(1-etax^2)
              +2*etax*(df(vv,y)-dfdx(vu))) )$
ok:=if ok and(Equ=0)and(BCu0=0)and(BCuh=0) then 1 else 0;
gup:=5/4*mean(z*Equ) -20/9*anu/eta*BCuh;
vu:=vu+solv_u( -Equ+gup*cbrt )/del;
gu:=gu+gup/del^2;
end;

showtime;
end until ok;

end;
```

Acknowledgments. This research was supported by a grant from the University of Southern Queensland, by the Australian Research Council, and by the Volkswagen Foundation, Germany, I/69449.

REFERENCES

[1] M. Abramowitz and I. A. Stegun, eds., *Handbook of Mathematical Functions*, Dover, New York, 1965.

[2] R. Arnold and J. Noye, *Numerical modelling of long waves*, in Numerical Solutions of Partial Differential Equations, J. Noye, ed., North-Holland, Amsterdam, 1982, pp. 437–453.

[3] R. Arnold and J. Noye, *On the performance of turbulent energy closure schemes of wind driven flows in shallow seas*, in Computational Techniques and Applications: CTAC-83, J. Noye, ed., North-Holland, Amsterdam, 1984, pp. 425–437.

[4] R. Arnold and J. Noye, *Open boundary conditions for a tidal and storm surge model of Bass strait*, in Computational Techniques and Applications: CTAC-85, J. Noye, ed., North-Holland, Amsterdam, 1986, pp. 503–518.

[5] V. Balakotaiah and H. C. Chang, *Dispersion of chemical solutes in chromatographs and reactors*, Philos. Trans. Roy. Soc. London Ser. A, 351 (1995), pp. 39–75.

[6] P. B. Bedient and W. C. Huber, *Hydrology and Floodplain Analysis*, Addison-Wesley, Reading, MA, 1988.

[7] J. R. Bertschy, R. W. Chin, and F. H. Abernathy, *High-strain-rate free-surface boundary-layer flows*, J. Fluid Mech., 126 (1983), pp. 443–461.

[8] G. Birkhoff and G.-C. Rota, *Ordinary Differential Equations*, Xerox College Publishing, Lexington, MA, 1969.

[9] J. Carr, *Applications of Centre Manifold Theory*, Appl. Math. Sci. 35, Springer-Verlag, New York, 1981.

[10] P. H. Coullet and E. A. Spiegel, *Amplitude equations for systems with competing instabilities*, SIAM J. Appl. Math., 43 (1983), pp. 776–821.

[11] S. M. Cox and A. J. Roberts, *Initial conditions for models of dynamical systems*, Phys. D, 85 (1995), pp. 126–141.

[12] P. A. Durbin, *Near-wall turbulence closure modeling without "damping functions,"* Theoret. Comput. Fluid Dynamics, 3 (1991), pp. 1–13.

[13] P. A. Durbin, *Application of a near-wall turbulence model to boundary layers and heat transfer*, Int. J. Heat and Fluid Flow, 14 (1993), pp. 316–323.

[14] P. A. Durbin, *A Reynolds stress model for near-wall turbulence*, J. Fluid Mech., 249 (1993), pp. 465–498.
[15] J. Fredsoe and R. Deigaard, *Mechanics of Coastal Sediment Transport*, Adv. Series Ocean Eng. 3, World Scientific, Singapore, 1992.
[16] Th. Gallay, *A center-stable manifold theorem for differential equations in Banach spaces*, Comm. Math. Phys, 152 (1993), pp. 249–268.
[17] M. M. Gibson and W. Rodi, *Simulation of free surface effects on turbulence with a Reynolds stress model*, J. Hydraulic Res., 27 (1989), pp. 233–244.
[18] H. Haken, *Synergetics, An Introduction*, Springer-Verlag, Berlin, 1983.
[19] K. Hanjalić and B. E. Launder, *A Reynolds stress model of turbulence and its applications to thin shear flows*, J. Fluid Mech., 52 (1972), pp. 609–638.
[20] P. Hartman, *Ordinary Differential Equations*, 2nd ed., Birkhäuser, Boston, 1982.
[21] E. J. Hinch, *Perturbation Methods*, Cambridge Texts Appl. Math., Cambridge University Press, Cambridge, UK, 1991.
[22] B. R. Hodges and R. L. Street, *On simulation of turbulent nonlinear free-surface flows*, J. Comput. Physics, 151 (1999), pp. 425–457.
[23] M. Kay, *Practical Hydraulics*, E. and F. N. Spon, London, 1998.
[24] R. J. Keller and W. Rodi, *Prediction of flow characteristics in main channel/flood plain flows*, J. Hydraulic Res., 26 (1988), pp. 425–442.
[25] P.-A. Krogstad and R. A. Antonia, *Structure of turbulent boundary layers on smooth and rough walls*, J. Fluid Mech., 277 (1994), pp. 1–21.
[26] B. E. Launder, G. J. Reece, and W. Rodi, *Progress in the development of a Reynolds-stress turbulence closure*, J. Fluid Mech., 68 (1975), pp. 537–566.
[27] B. Lin and R. A. Falconer, *Three-dimensional layer-integrated modelling of estuarine flows with flooding and drying*, Estuarine, Coastal and Shelf Sci., 44 (1997), pp. 737–751.
[28] J. Mathieu and J. Scott, *An Introduction to Turbulent Flow*, Cambridge University Press, Cambridge, UK, 2000.
[29] C. C. Mei, *The Applied Dynamics of Ocean Surface Waves*, Adv. Series Ocean Eng. 1, World Scientific, Singapore, 1989.
[30] Z. Mei and A. J. Roberts, *Equations for turbulent flood waves*, in Structure and Dynamics of Nonlinear Waves in Fluids, A. Mielke and K. Kirchgässner, eds., World Scientific, Singapore, 1995, pp. 342–352.
[31] G. N. Mercer and A. J. Roberts, *A centre manifold description of contaminant dispersion in channels with varying flow properties*, SIAM J. Appl. Math., 50 (1990), pp. 1547–1565.
[32] G. N. Mercer and A. J. Roberts, *A complete model of shear dispersion in pipes*, Jap. J. Indust. Appl. Math., 11 (1994), pp. 499–521.
[33] B. Mohammadi and O. Pironneau, *Analysis of the k-epsilon Turbulence Model*, Masson, Paris, John Wiley and Sons, Chichester, UK, 1994.
[34] D. J. Needham and J. H. Merkin, *On roll waves down an open inclined channel*, Proc. Roy. Soc. London Ser. A, 394 (1984), pp. 259–278.
[35] D. H. Peregrine, *Equations for water waves and the approximations behind them*, in Waves on Beaches and Resulting Sediment, R. E. Meyer, ed., Academic Press, 1972, pp. 95–121.
[36] Th. Prokopiou, M. Cheng, and H. C. Chang, *Long waves on inclined films at high Reynolds number*, J. Fluid Mech., 222 (1991), pp. 665–691.
[37] A. K. Rastogi and W. Rodi, *Prediction of heat and mass transfer in open channels*, J. Hydraulics Div., 104 (1978), pp. 397–419.
[38] N. M. Ribe, *Bending and stretching of thin viscous sheets*, J. Fluid Mech., 433 (2001), pp. 135–160.
[39] A. J. Roberts, *Simple examples of the derivation of amplitude equations for systems of equations possessing bifurcations*, J. Austral. Math. Soc. B, 27 (1985), pp. 48–65.
[40] A. J. Roberts, *The application of centre manifold theory to the evolution of systems which vary slowly in space*, J. Austral. Math. Soc. Ser. B, 29 (1988), pp. 480–500.
[41] A. J. Roberts, *Appropriate initial conditions for asymptotic descriptions of the long term evolution of dynamical systems*, J. Austral. Math. Soc. Ser. B, 31 (1989), pp. 48–75.
[42] A. J. Roberts, *Boundary conditions for approximate differential equations*, J. Austral. Math. Soc. Ser. B, 34 (1992), pp. 54–80.
[43] A. J. Roberts, *The invariant manifold of beam deformations. Part* 1*: The simple circular rod*, J. Elasticity, 30 (1993), pp. 1–54.
[44] A. J. Roberts, *Low-dimensional models of thin film fluid dynamics*, Phys. Lett. A, 212 (1996), pp. 63–72.
[45] A. J. Roberts, *Low-Dimensional Modelling of Dynamical Systems*, Technical report, University of Southern Queensland, Toowoomba, Australia, 1997; also available online from

http://arXiv.org/abs/chao-dyn/?9705010.

[46] A. J. Roberts, *Low-dimensional modelling of dynamics via computer algebra*, Comput. Phys. Comm., 100 (1997), pp. 215–230.

[47] A. J. Roberts, *An accurate model of thin* $2d$ *fluid flows with inertia on curved surfaces*, in Free-Surface Flows with Viscosity, P. A. Tyvand, ed., Adv. Fluid Mech. 16, pp. 69–88.

[48] A. J. Roberts, *Computer algebra derives correct initial conditions for low-dimensional dynamical models*, Comput. Phys. Comm., 126 (2000), pp. 187–206.

[49] A. J. Roberts, *Solve Differential-Algebraic Equations in Matlab*, Technical report, University of Queensland, Toowoomba, Australia, 2000; also available online from http://www.sci.usq.edu.au/staff/robertsa/dae.dtx.

[50] W. Rodi, *Turbulence Models and Their Applications in Hydraulics*, International Association of Hydraulic Research, Delft, The Netherlands, 1980.

[51] K. Shiono and D. W. Knight, *Turbulent open-channel flows with variable depth across the channel*, J. Fluid Mech., 222 (1991), pp. 617–646.

[52] C. G. Speziale, *Analytical methods for the development of Reynolds-stress closures in turbulence*, Annu. Rev. Fluid Mech., 23 (1991), pp. 107–157.

[53] P. K. Stansby, A. Chegini, and T. C. D. Barnes, *The initial stages of dam-break flow*, J. Fluid Mech., 374 (1998), pp. 407–424.

[54] IB A. Svendsen, J. Veeramony, J. Bakunin, and J. T. Kirby, *The flow in weak hydraulic jumps*, J. Fluid Mech., 418 (2000), pp. 25–57.

[55] S. D. Watt and A. J. Roberts, *The accurate dynamic modelling of contaminant dispersion in channels*, SIAM J. Appl Math, 55 (1995), pp. 1016–1038.

SIAM J. APPL. MATH.
Vol. 63, No. 2, pp. 459–484

THE MCKEAN'S CARICATURE OF THE FITZHUGH–NAGUMO MODEL I. THE SPACE-CLAMPED SYSTEM*

ARNAUD TONNELIER†

Abstract. Within the context of Liénard equations, we present the FitzHugh–Nagumo model with an idealized nonlinearity. We give an analytical expression (i) for the transient regime corresponding to the emission of a finite number of action potentials (or spikes), and (ii) for the asymptotic regime corresponding to the existence of a limit cycle. We carry out a global analysis to study periodic solutions, the existence of which is linked to the solutions of a system of transcendental equations. The periodic solutions are obtained with the help of the harmonic balance method or as limit behavior of the transient regime. We show how the appearance of periodic solutions corresponds either to a fold limit cycle bifurcation or to a Hopf bifurcation at infinity. The results obtained are in agreement with local analysis methods, i.e., the Melnikov method and the averaging method. The generalization of the model leads us to formulate two conjectures concerning the number of limit cycles for the piecewise linear Liénard equations.

Key words. excitability, oscillations, limit cycle, piecewise linear model, bifurcation

AMS subject classifications. 34A05, 37G15, 34C05, 92C20

PII. S0036139901393500

1. Introduction.

We consider the autonomous system

$$\begin{aligned} \frac{dv}{dt} &= p(v) - w, \\ \frac{dw}{dt} &= bv, \end{aligned} \tag{1.1}$$

where $t \in \mathbb{R}$, $b > 0$, $v(t) \in \mathbb{R}$ represents the system status variable at time t, $w(t) \in \mathbb{R}$ represents an additional variable, and $p : \mathbb{R} \to \mathbb{R}$ is a given function. These equations are known as the Liénard system [23], [22]. Special cases of (1.1) provide mathematical models for many applications in science and engineering. We mention here biology [31], [17], electronics (e.g., the van der Pol model [38]), chemistry [19], and mechanics (for instance, damped mass spring systems).

In this paper we consider the case of a cubic-like function for p. System (1.1) then describes the behavior of an isolated excitable cell where v is the membrane potential and w the recovery variable. When p is given by

$$p(v) = v(1-v)(v-a), \qquad \text{where} \quad 0 < a < 1, \tag{1.2}$$

system (1.1) is the polynomial FitzHugh–Nagumo model [8], [32]. It has given rise to many studies and the reader is referred to the references given in [31] and [17]. There are no particular requirements with respect to the choice of p, except to have a graphical representation similar to that given by (1.2). When p is a polynomial function, it is difficult to obtain analytical results since exact solutions cannot be obtained. In order to be able to go further with the study and the understanding of

*Received by the editors August 8, 2001; accepted for publication (in revised form) March 12, 2002; published electronically November 19, 2002.
http://www.siam.org/journals/siap/63-2/39350.html

†Techniques de l'Imagerie, de la Modélisation et de la Cognition, CNRS UMR 5525, Faculté de Médecine, F-38706 La Tronche, France. Current address: Laboratory of Computational Neuroscience, EPFL, 1015 Lausanne, Switzerland (Arnaud.Tonnelier@epfl.ch).

the model, we will follow the choice originally proposed by McKean [29], considering that

$$p(v) = -v + h(v - a), \qquad \text{where} \quad 0 < a < 1, \tag{1.3}$$

and h is the Heaviside function

$$h(x) = \begin{cases} 0 \text{ if } x < 0, \\ 1 \text{ if } x > 0. \end{cases} \tag{1.4}$$

The study of model (1.1)–(1.3) with a diffusive term on v was initiated by McKean [29] and developed considerably by Rinzel and Keller [34] and Wang [39], [40]. Their analyses covered the existence and the stability of traveling wave solutions.

The use of idealized nonlinearity with the help of the Heaviside function has become a classic procedure in the modeling of threshold effects in excitable media [3], [6], [16]. This approach leads to analytical results concerning properties of the model and provides a qualitative description for a more general class of functions. As far as we know, no specific studies have been carried out on the model isolated in space (1.1)–(1.3). More generally, we are going to study the following system:

$$\begin{aligned} \frac{dv}{dt} &= -\lambda v + \mu h(v - a) - w, \\ \frac{dw}{dt} &= bv, \end{aligned} \tag{1.5}$$

where

$$\lambda > 0,\ \mu > 0,\ a > 0, \text{ and } \mu > \lambda a. \tag{1.6}$$

The latter condition shows the restriction that must be imposed upon p to obtain a shape similar to that obtained with (1.2). We are going to carry out a global analysis of equations (1.5) considering (λ, μ, a, b) as parameters. It should be noted that the change of variables

$$(\tilde{t}, \tilde{w}, \tilde{\lambda}, \tilde{\mu}) \to \frac{1}{\sqrt{b}}(t, w, \lambda, \mu) \tag{1.7}$$

enables us to consider the case $b = 1$. Nevertheless, we will not make this choice given the usefulness of the parameter b in the interpretation of the results. In addition, we are going to consider the case $b \to 0$.

Our study covers the case where a constant input I is injected into the system:

$$\begin{aligned} \frac{dv}{dt} &= p(v) - w + I, \\ \frac{dw}{dt} &= bv. \end{aligned}$$

We obtain (1.5) by putting $\tilde{w} = w - I$, which, in the phase plane, corresponds to a shift of the v-nullcline. The case of a variable current $I(t)$ will be discussed briefly and will be the subject of another paper. It should be noted that the FitzHugh–Nagumo model has an additional term in the recovery variable $\dot{w} = b(v - \gamma w)$, and the simplification $\gamma = 0$ introduces an artifact in the sense that a constant current does not change the behavior. However, since we are not interested in the bistable

regime, this limiting situation allows a qualitative description of the excitable regime and captures the bifurcations of the complete system as $\gamma \to 0$.

This article is organized as follows. In section 2, we present the context into which we put our study and introduce the elements that are useful to our analysis. In section 3, we discuss the so-called *spike solution* that corresponds to the emission of a finite number of action potentials. Particular attention is given to the study of the singular perturbed system obtained as $b \to 0$. Section 4 is devoted to an analytical study of periodic solutions, and a geometric analysis is given in section 5. We determine, in section 6, an approximation of the bigger limit cycle. Section 7 provides a mathematical link between excitability and oscillations. In the final section, we summarize our results and we discuss the problem of the number of limit cycles for the piecewise linear Liénard equations.

2. General. First, let us consider system (1.1) with p having a cubic shape similar to that given by (1.2). For a smooth reaction function, $p \in C^1$, classical results from dynamical systems theory enable us to state the following proposition.

PROPOSITION 2.1. *The single fixed point $E_0 = (0, p(0))$ is locally stable if and only if $p'(0) < 0$. If $p'(0) \geq 0$, a limit cycle, surrounding E_0, appears via a Hopf bifurcation.*

Proof. The single fixed point of (1.1) is $(0, p(0))$. Its local stability is given by the eigenvalues of the Jacobian matrix of (1.1) at $(0, p(0))$:

$$J = \begin{bmatrix} p'(0) & -1 \\ b & 0 \end{bmatrix}.$$

For $p'(0) < 0$, we obtain local stability of the fixed point. The equality $p'(0) = 0$ corresponds to the Hopf bifurcation equation. The second part of the proposition is obtained by constructing an invariant set containing E_0 and using the Poincaré–Bendixson theorem. □

The Hopf bifurcation is a mechanism that is frequently encountered in the appearance of small-amplitude oscillations [13]. It is possible to specify the behavior of the solution in the neighborhood of its Hopf bifurcation and to obtain, locally, an analytical expression for the solution of system (1.1) [20]. However, the case that we are going to look at is the so-called *excitable* one, which corresponds to $p'(0) < 0$. There is no precise mathematical definition of excitability, and we say that a system is excitable if a perturbation from its resting state leads to a large excursion for the solution in the phase plane and a return to its resting state. This phenomenon is characterized by a solution $(v(t), w(t))$ of (1.5) satisfying the following two properties:

$$\begin{aligned} &\text{(P1)} \quad \exists\, 0 < t_1 < t_2, \text{ so that } v \text{ is increasing on } [t_1, t_2], \\ &\text{(P2)} \quad \lim_{t \to +\infty} v(t) = 0. \end{aligned}$$

Such a solution will be called a *spike solution*. It should be noted that (P2) is always satisfied when the domain of attraction of $(0, p(0))$ is the whole phase plane. For our study, property (P1) is sufficient to characterize the excitability of our system. When p is the function involved in (1.5), property (P1) will be satisfied as soon as (v, w) crosses the threshold segment $[-\lambda a, -\lambda a + \mu]$ in the phase plane. These two properties are in agreement with the characterization of excitability given in [1], i.e., the existence, in the phase space of the so-called *amplifying set* and *decaying set*.

We are going to examine the two phenomena associated with the emergence of an action potential. These phenomena will be written according to the concept of *spike*

solution and periodic solution. The *spike solution* is a transient regime characterized by a finite number of action potentials. The periodic solution corresponds to the emission of an infinite number of action potentials. It is an asymptotic regime that shows the presence of a limit cycle. These two regimes represent the basic properties of neuronal excitability [33], [15].

Before proceeding with an analytical study of these regimes, we are going to give a qualitative interpretation of the dynamical behavior of system (1.1). This system can be rewritten in a convenient form usually used within the context of self-excited oscillations [30], [12]:

$$\frac{d^2v}{dt^2} - p'(v)\frac{dv}{dt} + bv = 0.$$

It is then useful to consider the energy derived from the harmonic oscillator (obtained as $p' = 0$) defined by

$$E = \frac{1}{2}\left(\frac{dv}{dt}\right)^2 + \frac{b}{2}v^2.$$

This gives a solution,

$$\frac{dE}{dt} = p'(v)\left(\frac{dv}{dt}\right)^2. \tag{2.1}$$

We find the stability of the fixed point $(0, p(0))$, provided by Proposition 2.1, when $p'(0) < 0$, which corresponds to damped oscillations in the neighborhood of this fixed point. As p' is not negative everywhere on $\mathbb{R}$, it is not possible to obtain a conclusion concerning the global stability of $(0, p(0))$. In particular, it is possible that the added energy, when $p' > 0$, is sufficient to give rise to a limit cycle. Equation (2.1) provides information concerning this cycle to the extent that it must contain at least one root of p'. Note that this result can be found using the Poincaré–Bendixson criterion. When p is the cubic polynomial proposed by FitzHugh–Nagumo (1.2), system (1.1) does not have a limit cycle [24]. Nevertheless, while keeping a similar shape for p, it is possible to obtain a limit cycle. For example, for

$$p(x) = \begin{cases} -x & \text{if } x \leq 0, \\ 10x(x - 0.3)(1 - x) & \text{if not,} \end{cases}$$

and $b = 6$, one observes, numerically, the existence of a stable limit cycle. Thus the constraint on p, said to be of cubic shape, leaves a variability in the dynamical behavior of (1.1). In the case we are going to study, where p is represented in Figure 2.1, we will see that the energy input due to the jump discontinuity of v', $\Delta v' = \mu$ when v crosses the line $v = a$, may be sufficient to give rise to a limit cycle. In this case, the limit cycle coexists with the fixed point $(0, 0)$ and this situation is termed as *hard self-excitation*.

Before beginning this study, it is necessary to specify the meaning given to a solution of (1.5) when p is discontinuous. Geometrically, a solution corresponds to a trajectory in the phase plane (v, w). If this trajectory crosses the line of discontinuity transversally, the solution is easy to define: for t such that $v(t) = a$, $v'(t)$ has a jump discontinuity $(v'(t^+) - v'(t^-) = \pm\mu)$, and elsewhere the solution is C^1 and satisfies (1.5) in the classical sense. In the case where the trajectory tangentially meets the

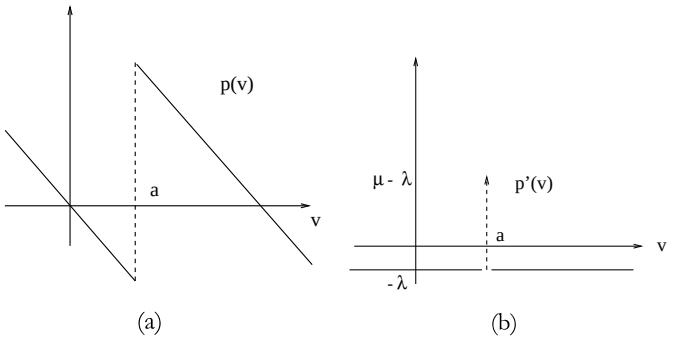


FIG. 2.1. (a) *The nonlinear function* p *and* (b) *its distributional derivative* p'.

line of discontinuity, the solution satisfies $v(t) = a$ on a nonempty set. In this case, we speak of a *generalized solution* and approach the problem from a geometrical point of view. It is not the purpose of this article to give a precise mathematical characterization of this solution, and the reader is referred to [7], [18]. It should be noted that the problem of discontinuous vector fields is covered extensively in control theory, e.g., [5], [14].

Our main results are given in the following summary. In section 3, we demonstrate that the spike solution contains only one spike when $\lambda^2 \geq 4b$ and several spikes can be emitted otherwise (depending on the initial conditions). In the former case, we derive a simple expression for the solution as $b \to 0$ and, in the latter, we give the general expression for the spike solution. The next sections focus on the case $\lambda^2 < 4b$ for which we derive, in section 4, analytical results on the existence and the expression of the periodic solutions. The periodic orbits appear via a double limit cycle bifurcation that we compute in the plane (a, b). Using a geometrical analysis (section 5), we characterize the two periodic solutions, represented in the phase plane by two concentric limit cycles. We show how in the limiting situation $\lambda \to 0$ and $\mu \to 0$ the periodic orbits can be obtained with the use of the Melnikov function. Moreover, we discuss the existence of two different types of unstable limit cycles referred to either as a classical or a generalized solution. The generalized solution is related to the discontinuity of the vector field. In section 6, the study as $\lambda \to 0$ allows us to capture and to describe the bigger limit cycle which is obtained as a Hopf bifurcation at infinity. In section 7, we show how the spike solutions and the periodic solutions are related.

3. Excitability and singular perturbation. The purpose of this section is to study the *spike solution*. In particular, we characterize this solution by the number of spikes that are part of the solution. This number corresponds to the number of times that v crosses the threshold a, where $v'^- > 0$ (where v'^- designates the left-hand derivative of v). This number includes the initial pulse corresponding to the perturbation due to the initial condition, noted as (v_0, w_0). In order to simplify the study we consider the case where $w_0 = 0$. We distinguish between several cases, according to the value of $\lambda^2 - 4b$. We prove the following proposition.

PROPOSITION 3.1. *For* $\lambda^2 \geq 4b$*, there is a spike solution when* $a < v_0 < \frac{\mu}{\lambda}$*. This solution only presents a single spike.*

Proof. First, we look at the case where $\lambda^2 > 4b$. For $v_0 < a$ and as long as $v(t) < a$, we have

$$v''(t) + \lambda v'(t) + bv(t) = 0.$$

The solution is then given by

$$v(t) = \frac{v_0}{r_+ - r_-}\left(-(\lambda - r_-)e^{r_+ t} + (\lambda + r_+)e^{r_- t}\right),$$

where

$$r_\pm = \frac{1}{2}(-\lambda \pm \sqrt{\lambda^2 - 4b}). \tag{3.1}$$

Let $\tilde{t}$ be the time defined by $\tilde{t} = 1/(r_+ - r_-)\ln(1 + \frac{\lambda}{r_+})/(1 + \frac{\lambda}{r_-})$. If $v_0 > 0$, v is decreasing on $[0, \tilde{t}\,]$ and increasing on $[\tilde{t}, +\infty[$. In addition, we have $\lim_{t\to+\infty} v(t) = 0$ and thus $\forall t > 0, v(t) < a$. When $v_0 < 0$, if $\forall t, v(t) < a$, the study is completed; conversely, if there is a time t^* so that $v(t^*) = a$, then the trajectory crosses the line $w = 0$ for a value of v greater than a, and, given a time shift, the study corresponds to the case where $v_0 > a$.

If $v_0 > a$, there is a time t^* so that $v(t^*) = 0$ and $w(t^*) = w_1 > -\lambda a + \mu$. If we put $t^* = 0$, this gives $v(t) = \frac{w_1}{r_+ - r_-}(e^{r_- t} - e^{r_+ t})$ and thus $\forall t > 0$, $v(t) < 0$ and $\lim_{t\to+\infty} v(t) = 0$.

The case where $\lambda^2 = 4b$ is dealt with in a similar way. □

For $\lambda^2 - 4b \geq 0$, the response to an input $I = I_0\delta(t - t_0)$ is a single action potential when $I_0 > a$.

We will now obtain a simple analytical expression for the potential v. The previous study showed the existence of several phases when a spike is emitted. This point can be made more specific by studying the case $b \ll 1$. This situation models the behavior of a system in which two time scales are involved; i.e., v is a fast variable and w is a slow variable. The mathematical description of the excitability is a classical one (see, for example, [17]) and is carried out using the singular perturbation theory. In our case, the relevance is to allow explicit solutions that give a simple expression for v according to the different phases of the *spike solution*.

Let there be (v_0, w_0) so that $a < v_0 < \frac{\mu}{\lambda}$ and $w_0 = 0$. In addition, let us assume that $v_0 - \frac{\mu}{\lambda}$ is of order greater than a $O(b)$. The variations of v can be separated into four phases. The first phase, which is the *excited phase*, is fast and the motion is governed approximatively by the system

$$\begin{aligned} \frac{dv}{dt} &= p(v) - w, \\ \frac{dw}{dt} &= 0, \end{aligned}$$

which gives

$$v(t) = \frac{1}{\lambda}\left(\mu + e^{-\lambda t}(\lambda v_0 - \mu)\right). \tag{3.2}$$

This approximation is valid as long as $v(t)$ is at a greater distance from the v-nullcline than a $O(b)$ value. If not, we enter the second phase where the dynamic is described using a new time scale $\tau = bt$. In this phase, v is adjusted to maintain a pseudoequilibrium at $w = p(v)$, and we have $v(\tau) = \frac{1}{\lambda}(\mu - w)$. We obtain

$$v(t) = \frac{\mu}{\lambda}e^{-\frac{\tau}{\lambda}}. \tag{3.3}$$

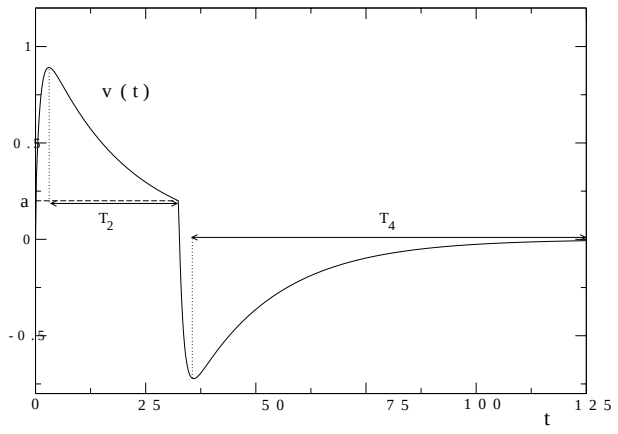


FIG. 3.1. *Solution $v(t)$ of (1.5) for $(v_0, w_0) = (0.25, 0)$, $\lambda = 1$, $a = 0.2$, $b = 0.05$, and $\mu = 1$. Intervals T_2 and T_4 designate the durations of the two slow phases.*

We enter into the third phase as v reaches a. We have a fast motion where v is given by

$$v(t) = \frac{\mu}{\lambda}(e^{-\lambda t} - 1) + a. \tag{3.4}$$

The final phase is characterized by a slow return to the equilibrium state according to

$$v(t) = \left(a - \frac{\mu}{\lambda}\right) e^{-\frac{\tau}{\lambda}}. \tag{3.5}$$

We can easily find these results by observing that the roots r_+ and r_- given by (3.1) are written $r_+ = -\frac{b}{\lambda} + O(b^2)$ and $r_- = -\lambda + O(b)$. The fast dynamic is obtained using the zero order approximation, and the slow motion by using the first order one. The different phases, (3.2)–(3.5), correspond to the charge and discharge of a capacitor, and are graphically shown in Figure 3.1. They allow precise identification of the role of each parameter. In particular, the amplitude of the potential is parameterized by $\frac{\mu}{\lambda}$ and a. In addition, it is possible to obtain an approximation of the duration of a spike T using the durations of the slow dynamics of phases two and four, written T_2 and T_4, respectively. We consider that the duration of phase four is the time for which $v(t) = O(b)$. This gives

$$T = T_2 + T_4,$$

where

$$\begin{aligned} T_2 &= \frac{\lambda}{b} \ln \frac{\mu}{\lambda a} + O\left(\frac{1}{b}\right), \\ T_4 &= O\left(-\frac{\ln b}{b}\right). \end{aligned} \tag{3.6}$$

For $b \ll 1$, it is possible to obtain a simple description of the subthreshold response to a variable input $I(t)$. This response is the one given by an RC filter, where $\lambda = \frac{1}{RC}$ and is written

$$v(t) = e^{-\lambda \cdot} * I(t).$$

If we consider a train of impulses at regular intervals, the system reacts preferentially at a high input frequency in that the higher the input frequency, the earlier a spike is emitted. More precisely, with $I(t) = I_0 \sum_{t_i} \delta(t - t_i)$, where $t_i = iT$ and $i \in \mathbb{N}$, the subthreshold response is given by

$$v(t) = I_0 e^{-\lambda t} \frac{1 - e^{\lambda(n+1)T}}{1 - e^{\lambda T}},$$

where n is the index of the final pulse of $I(t)$ before the system reaches the threshold. Thus, for an input such as $I_0 < a$ with a small frequency

$$\frac{1}{T} < \frac{\lambda}{\ln(\frac{a}{a - I_0})},$$

the system cannot emit an action potential.

Now, and for the rest of this article, unless indicated otherwise, we are going to consider the case in which $\lambda^2 - 4b < 0$. We shall see that the model presents a richer dynamic in the sense that the *spike solution* is able to present several action potentials. In addition, we will show in the following section the existence of periodic solutions. We continue the case of a solution satisfying $\lim_{t \to +\infty} v(t) = 0$ and, therefore, there exists a constant $C > 0$ and a time t^* starting from which we have

$$|v(t)| < Ce^{-\frac{\lambda}{2}t}.$$

We can then define the Laplace transform of v

$$\mathcal{L}(v)(p) = \int_0^\infty v(t) e^{-pt} dt$$

for which the region of convergence is the half plane

$$D = \left\{ p \in \mathbb{C} \mid \mathrm{Re}(p) > -\frac{\lambda}{2} \right\}.$$

We define the finite sequence of times, written $(t_i)_{0 \ldots 2n-1}$, so that $t_0 = 0$, and for $i \neq 0$, $v(t_i) = a$ and $\Delta v'(t_i) = (-1)^i \mu$. This sequence indicates the passage of potential via the line of discontinuity and corresponds to a jump of the derivative of v. An equivalent characterization of t_i is given by $v'(t_{2j}^-) > 0$ and $v'(t_{2j+1}^-) < 0$. We have, on $]t_{2i}, t_{2i+1}[$, $v(t) > a$ with $v(t_0) = v_0 > a$. The number n corresponds to the number of spikes emitted by the system. For $w_0 = 0$, we calculate

$$\mathcal{L}(v)(p) = \frac{\mu}{p^2 + \lambda p + b} \sum_{i=0}^{n-1} (e^{-pt_{2i}} - e^{-pt_{2i+1}}) + \frac{pv_0}{p^2 + \lambda p + b}.$$

We write in the following $r = \sqrt{4b - \lambda^2}$. Using inverse Laplace transforms gives

$$v(t) = v_0 \alpha(t) + \sum_{i=0}^{2n-1} (-1)^i h(t - t_i) \varphi(t - t_i), \tag{3.7}$$

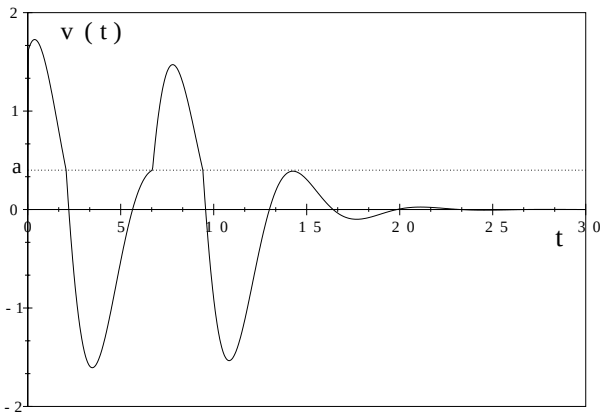


FIG. 3.2. *Solution $v(t)$ of* (1.5) *for* $(v_0, w_0) = (1.6, 0)$, $\lambda = 0.8$, $\mu = 2$, $a = 0.4$, *and* $b = 1$. *This* spike solution *presents two action potentials, or spikes.*

where

$$\varphi(t) = \frac{2\mu}{r} e^{-\frac{\lambda}{2}t} \sin \frac{r}{2} t, \tag{3.8}$$

$$\alpha(t) = e^{-\frac{\lambda}{2}t} \Big(\cos \frac{r}{2} t - \frac{\lambda}{r} \sin \frac{r}{2} t \Big). \tag{3.9}$$

The above expression characterizes the transient regime with (i) a term that depends on the initial excitation, v_0, and corresponds to damped oscillations of period $\frac{4\pi}{\sqrt{4b-\lambda^2}}$, and (ii) a sum of terms with the form $(-1)^i S_{t_i}(h\varphi)$ (where S is the shift operator) reproducing an excitation when $v'^- > 0$ (even i) or an inhibition $v'^- < 0$ (odd i). This sum shows the different crossing $v = a$ and is defined implicitly by the existence of times t_i such as $v(t_i) = a$. It is clear that t_1 exists: it is given by the smallest strictly positive solution of the equation

$$v_0 \alpha(t) + \varphi(t) = a.$$

The sequence of times (t_i) cannot be expressed with known functions and is implicitly defined using expression (3.7). Figure 3.2 illustrates the case in which the system generates two action potentials. The return to the resting state takes place via damped oscillations and induces computational properties which differ from that studied above. If we consider the case of a system that has not emitted a spike, its subthreshold response to an input $I(t)$ is given by

$$v(t) = \alpha * I,$$

corresponding to the response of an RLC filter, with $\lambda = \frac{R}{L}$ and $b = \frac{1}{LC}$, when an input I is applied. In particular, the filter response is more significant for an input signal having a resonant frequency close to $\sqrt{b - \frac{\lambda^2}{4}}$.

4. Periodic solutions. Let us assume that system (1.5) has a periodic solution. According to the expression of the vector field, this solution delimits a domain containing the origin, which is a stable fixed point. In addition, when $\lambda \neq 0$, it is possible

to construct an invariant region large enough to include this limit cycle. Thus, there are at least two limit cycles surrounding the origin, with an alternation of stable and unstable cycles, the largest being stable.

We are looking for a periodic solution $(v(t), w(t)) \in \left(L^2(0,T)\right)^2$, where T is the period of the solution. This solution can be expressed in a Fourier series

$$\begin{aligned} v(t) &= \sum_n v_n e^{2i\pi n \frac{t}{T}}, \\ w(t) &= \sum_n w_n e^{2i\pi n \frac{t}{T}}. \end{aligned} \tag{4.1}$$

The technique used, which is known as the method of the harmonic balance (see [2], for example), involves identifying (v_n, w_n) using the differential equation satisfied by (v, w).

In the phase plane, a periodic solution crosses the line $v = a$ at two points, one of which satisfies $w > 0$ and the other $w < 0$. We set t_1 and t_2 as the two successive times that satisfy $v(t_i) = a,\ i = 1, 2$, so that $\forall t \in]t_1, t_2],\ v(t) > a$. The time-translation invariance of the periodic solution allows us to define the real τ so that $t_1 = -\tau$, $t_2 = \tau$, where $0 < \tau < \frac{T}{2}$. The periodic solution looked for satisfies

$$v(t) = \begin{cases} > a & \text{on }]-\tau, \tau[, \\ < a & \text{on } [-\frac{T}{2}, -\tau[\cup]\tau, \frac{T}{2}]. \end{cases} \tag{4.2}$$

The function $t \to h(v(t) - a)$ is a T-periodic function such as

$$h(v(t) - a) = \begin{cases} 1 & \text{if } \ t \in [-\tau, \tau], \\ 0 & \text{if not,} \end{cases}$$

and we calculate that

$$h(v(t) - a) = \frac{2\tau}{T} + \sum_{n \neq 0} \frac{1}{\pi n} \sin\left(2\pi n \frac{\tau}{T}\right) e^{2i\pi n \frac{t}{T}}.$$

Therefore we obtain

$$v(t) = \sum_n c_n \sin\left(2\pi n \frac{\tau}{T}\right) e^{2i\pi n \frac{t}{T}}, \tag{4.3}$$

where

$$c_n = \frac{2\mu T i}{-4\pi^2 n^2 + bT^2 + i2\pi\lambda T n}.$$

At this stage in the study, we may remark that the mean value of v is zero (which could be seen directly with (1.5)). The mean value of w is $w_0 = \frac{2\mu\tau}{T}$. The amplitude spectrum of v is $O\left(\frac{1}{n^2}\right)$, which ensures the normal convergence of the associated Fourier series.

Let f be the function defined by

$$f(t) = \sum_n i c_n e^{i2\pi n \frac{t}{T}}.$$

With the help of trigonometric transformations, (4.3) is written

$$v(t) = \frac{1}{2}(f(t - \tau) - f(t + \tau)). \tag{4.4}$$

We calculate

$$f(t) = \frac{-2\mu}{r\left(\cosh\frac{\lambda T}{2} - \cos\frac{r}{2}T\right)} e^{\frac{\lambda}{4}(T-2t)} \left(e^{-\frac{\lambda T}{4}} \sin\frac{r}{2}(T-t) + e^{\frac{\lambda T}{4}} \sin\frac{r}{2}t\right) \tag{4.5}$$

for $0 \le t \le T$, where $r = \sqrt{4b - \lambda^2}$ and f is defined on $\mathbb{R}$ by periodicity. Therefore, f is continuous on $\mathbb{R}$ and has a derivative for $t \neq \mathbb{Z}T$. Note that f does not depend on the auxiliary variable τ. A periodic solution exists if and only if there is T and τ such as $0 < \tau < \frac{T}{2}$, solutions of

$$\begin{aligned} f(0) - f(2\tau) &= 2a, \\ f(-2\tau) - f(0) &= 2a, \end{aligned} \tag{4.6}$$

so that v, given by (4.4), satisfies (4.2). We note $x = \frac{T}{2}$ and $y = \tau$. Elementary operations show that (4.6) can be written in the form

$$\begin{aligned} F(x,y) &= 0, \\ F(x,y-x) &= 0, \\ 0 < y &< x, \end{aligned} \tag{4.7}$$

where

$$F(x,y) = \mu \sinh\lambda x \sin ry - \mu \sin rx \sinh\lambda y - ar(\cosh\lambda x - \cos rx)\cosh\lambda y. \tag{4.8}$$

The existence of periodic solutions for the differential system (1.5) is given by the existence of roots for a system of transcendental equations. We have therefore reduced the differential problem to an algebraic one that corresponds to a search for roots in $\mathbb{R}$. In contrast to perturbation methods, it is interesting to note that our analysis is a global one and gives an analytical formula for a periodic solution.

Remark. A similar study can be carried out for $\lambda^2 - 4b > 0$. We then write $r = \sqrt{\lambda^2 - 4b}$. We find that

$$f(t) = \frac{-2\mu}{r\left(\cosh\frac{\lambda T}{2} - \cosh\frac{r}{2}T\right)} e^{\frac{\lambda}{4}(T-2t)} \left(e^{-\frac{\lambda T}{4}} \sinh\frac{r}{2}(T-t) + e^{\frac{\lambda T}{4}} \sinh\frac{r}{2}t\right),$$

where T and τ are given by the resolution of (4.7) with F defined by

$$F(x,y) = \mu \sinh\lambda x \sinh ry - \mu \sinh rx \sinh\lambda y - ar(\cosh\lambda x - \cosh rx)\cosh\lambda y. \tag{4.9}$$

Using $r < \lambda$, it is easy to show that $F(x, y-x) < 0$ when $0 < y < x$, and, therefore, there cannot be solutions of (4.7) with F given by (4.9), which confirms the result of the previous section.

Starting from the study carried out above, it is possible to state several simple properties concerning a periodic solution. First of all, it is easy to see that its existence is controlled by parameters r, λ, and $\frac{a}{\mu}$. In addition, we have the following bound for the periodic solution:

$$\|v\|_{+\infty} < \frac{4\mu}{\sqrt{4b - \lambda^2}} \cdot \frac{e^{\frac{\lambda T}{2}} + 1}{e^{\frac{\lambda T}{2}} - 2},$$

which is valid for $e^{\frac{\lambda T}{2}} > 2$. In particular, we can see that the smaller the period, the larger the bound.

Based on (4.7), in the general case, it is difficult to give conditions for the existence of (T, τ). More precisely, two phenomena appear to make the study tricky: (i) the presence of solutions of (4.7) that do not correspond to a periodic solution, and (ii) the presence of a periodic solution not detected by our analysis. We will look more closely at the second point in the next section. The first point arises from the fact that the existence of exactly two solutions for the equation $v(t) = a$ on $[0, T]$, corresponding to (4.2), is not reported in system (4.7). These two situations can be illustrated by looking at the solutions of (4.7) as $a \to 0$. If we take $a = 0$, the resolution of (4.7) leads to the family of solutions $(T_k, \tau_{p_k})_k$, where $k \in \mathbb{N},\ k > 1$:

$$\begin{aligned} T_k &= \frac{2k\pi}{\sqrt{4b - \lambda^2}}, \\ \tau_{p_k} &= \frac{p_k \pi}{\sqrt{4b - \lambda^2}}, \quad \text{where } p_k = 1 \dots k-1, \end{aligned} \tag{4.10}$$

and the implicit functions theorem leads to the existence of these solutions for a sufficiently small a. In fact, only the solution obtained for $k = 2$ is admissible; other solutions do not satisfy the assumptions of our study (given by (4.2)). Numerically, this solution corresponds to a stable limit cycle. As we have already mentioned, there must be an unstable cycle separating the domain of attraction of the origin from the stable cycle one. Therefore, we are in a situation where a limit cycle has not been detected.

Before clarifying this situation, we carry out a numerical study of the specific case used in [29], [34], and [39], where $\mu = 1$ and $\lambda = 1$. The results are illustrated in Figure 4.1, where we determine in the plane (a, b) the region where a periodic solution exists. It appears that there is a value of a, noted a^*, for which there is no periodic solution for $a \geq a^*$. When $a < a^*$, the existence of a periodic solution is obtained for $b \geq b_f(a)$. The curve $b_f(a)$ is given by the resolution, in $\{(x, y) \in \mathbb{R}^2 / \ 0 < y < x\}$, of

$$\begin{aligned} F(x, y) &= 0, \\ F(x, x - y) &= 0, \\ \det F_{x,y} &= 0, \end{aligned} \tag{4.11}$$

where $F_{x,y}$ is the Jacobian matrix of the system above with respect to (x, y). Geometrically speaking, the latter condition corresponds to a tangential intersection between the two curves defined by the equations $F(x, y) = 0$ and $F(x, x-y) = 0$. For $b = b_f(a)$, there is a single unstable limit cycle. For $b > b_f(a)$, there are two concentric limit cycles. The larger one is stable, and the smaller one is unstable, separating the different domains of attraction. At $b = b_f(a)$ a fold limit cycle bifurcation (or double limit cycle bifurcation) occurs. Several limiting situations can be analytically specified. When $a \to 0$, system (4.7) always has an admissible solution (given by (4.10) with $k = 2$), and the only restriction on b is related to the existence of r. We therefore have $\lim_{a \to 0} b_f(a) = 0.25$.

We determine the value of a^* using an asymptotic expansion of (4.7) as $b \to +\infty$. More exactly, we use an asymptotic expansion as $r \to +\infty$.

We write

$$x = \frac{x_1}{r} + O\left(\frac{1}{r^2}\right)$$

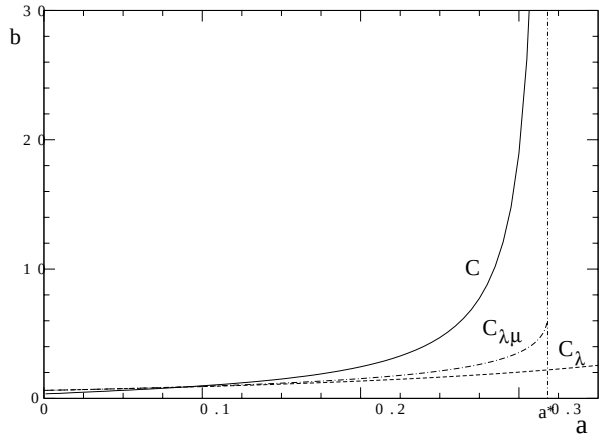


FIG. 4.1. *Curve of the fold limit cycle bifurcation* $C = \{b_f(a),\ 0 \le a < a^*\}$ *so that there are two periodic solutions if and only if* $b > b_f(a)$ *with* $a < a^*$. *Parameters are* $\lambda = 1$, $\mu = 1$. *Curves* C_λ *and* $C_{\lambda\mu}$ *correspond to the two approximations given by* (6.7).

as the expansion of x (it is easy to show that the zero order term is zero). The leading order expansion of $F(x, y)$ is

$$F(x, y) = a(\cos x_1 - 1)r + O(1),$$

which gives us the approximation

$$T = \frac{4\pi}{r} + O\left(\frac{1}{r^2}\right).$$

One notes the similarity with the expression (4.10) obtained above. In the same way, if we write

$$y = \frac{y_1}{r} + O\left(\frac{1}{r^2}\right),$$

the determination of y_1 is carried out by canceling the higher order term of the expansion of $F(x, y)$. We find

$$2\pi \sin(y_1) - \frac{a}{2}x_2^2 - a\pi^2 = 0, \tag{4.12}$$

and it should be noted that the expansion of $F(x, y-x)$ leads to the same expression. We show that $x_2 = 0$ (while remarking that x_2 is the first order term in the expansion of T as $b \to 0$ and using the symmetries of the differential equations (1.5)). We therefore find a solution of (4.12) if and only if

$$a \le \frac{1}{\pi}, \tag{4.13}$$

which enables us to obtain the value $a^* = \frac{1}{\pi}$ (Figure 4.1). When $a < a^*$ and r is large enough, we found two values of τ corresponding to exactly two limit cycles. We remark that, asymptotically, both these cycles have the same period.

We can now predict the behavior of system (1.5) for any (λ, μ). The change of variables

$$(\tilde{b}, \tilde{a}, \tilde{t}, \tilde{v}, \tilde{w}) = \left(\frac{b}{\lambda^2}, \frac{\lambda}{\mu}a, \lambda t, \frac{\lambda}{\mu}v, \frac{w}{\mu}\right)$$

enables us to find the case previously studied. Condition (4.13) is then written as

$$a \leq \frac{\mu}{\lambda\pi}, \tag{4.14}$$

and the bifurcation curve obtained from b_f is given by

$$b = \lambda^2 b_f\Big(\frac{\lambda}{\mu}a\Big). \tag{4.15}$$

Thus, for sufficiently small λ, the system has always a limit cycle. We will discuss this point in more detail in section 6. For μ large enough, the condition for existence of a periodic solution is written as $b \geq \frac{\lambda^2}{4}$.

5. Geometrical study. We are going to specify the dynamical behavior of the system in the phase plane. We also will make use of a geometrical analysis to characterize solutions of the transcendental equations system (4.7) in the sense that the search for periodic solutions should be carried out among the intersection points of the two curves $C_1 = \{(x, y), F(x, y) = 0\}$ and $C_2 = \{(x, y), F(x, y - x) = 0\}$ in the space region $0 < y < x$. Three configurations can then be distinguished.

5.1. No periodic solution. The simplest situation is obtained when the two curves C_1 and C_2 do not present any intersections. In this case, the origin is globally attractive. We already have illustrated such a configuration in Figure 3.2 and have shown that this case still appears when $\lambda^2 > 4b$ (in this case, C_1 and C_2 are defined using F given by (4.9)).

5.2. Pair of admissible solutions. We have seen that when $r \to +\infty$, it is possible to find exactly two pairs of solutions for system (4.7). When these solutions lead to an expression for the limit cycle, given by (4.4), satisfying the hypotheses (4.2), they correspond to solutions that are admissible. From numerical simulations, it appears that this situation occurs when C_1 and C_2 are two closed convex curves (in the region of the plane where $0 < y < x$). In this case, there are exactly two intersection points, which correspond to the two limit cycles (stable and unstable).

This configuration can also be found using perturbation methods. In particular, we will show a mechanism for the birth of these two limit cycles in the phase plane. We consider the following Hamiltonian system:

$$\begin{aligned} \frac{du}{dt} &= -w, \\ \frac{dw}{dt} &= bv \end{aligned} \tag{5.1}$$

for which the Hamiltonian function, written H, is given by

$$H(v, w) = v^2 + \frac{1}{b}w^2. \tag{5.2}$$

System (1.5) can be written as a perturbation of the Hamiltonian system (5.1)

$$\begin{aligned} \frac{dv}{dt} &= -w - \lambda g(v), \\ \frac{dw}{dt} &= bv, \end{aligned} \tag{5.3}$$

where $\lambda \ll 1$ and $g(v) = v - h(v-a)$. We take $\mu = \lambda$, but we have seen that the study can easily be extended to any (λ, μ) of the same order. It is easy to see that for system (5.3) the origin becomes a focus. In order to have a closer look at what becomes of the periodic trajectory of the center, we are going to use the Melnikov method [11]. This method, which arises from the averaging method, enables us to determine the periodic trajectories that are transformed into limit cycles and thus obtain an approximation of these cycles. The Melnikov function, associated with the level curve $H(v,w) = v^2 + \frac{1}{b}w^2 = l^2$, is

$$M(l) = \int_0^{2\pi} dt\ v g(v)|_{v=l\cos t}.$$

We obtain

$$M(l) = \pi l^2 - 2h(l-a)\sqrt{l^2 - a^2}.$$

Level curves of the unperturbed Hamiltonian system which transform into limit cycles are obtained as solutions of $M(l) = 0$. When $l < a$, the only solution is $l = 0$ and we find that the trajectories tend towards the origin. When $l > a$, $M(l) = 0$ is written as

$$l^4 - \frac{4}{\pi^2}l^2 + \frac{4}{\pi^2}a^2 = 0.$$

There are solutions if and only if $a \le \frac{1}{\pi}$. We then have the following result:

– If $a = \frac{1}{\pi}$, there is a single limit cycle which corresponds to the level curve defined by

$$H(v,w) = \frac{2}{\pi^2}. \tag{5.4}$$

– If $a < \frac{1}{\pi}$, there are two limit cycles which correspond to the level curves defined by

$$H(v,w) = \frac{2}{\pi^2}\left(1 \pm \sqrt{1 - \pi^2 a^2}\right). \tag{5.5}$$

These results can be added to by using system (4.7). When $\mu = \lambda \ll 1$, the second order asymptotic expansion of F gives

$$\begin{aligned} F(x,y) = {} & 2a\sqrt{b}(\cos(2\sqrt{b}x) - 1) \\ & + \Big(- a\sqrt{b}(x^2 + y^2) + \frac{a}{4\sqrt{b}}(1 - \cos(2\sqrt{b}x)) \\ & + \Big(\frac{ax}{2} - y\Big)\sin(2\sqrt{b}x) + x\sin(2\sqrt{b}y) + a\sqrt{b}y^2\cos(2\sqrt{b}y)\Big)\lambda^2 + O(\lambda^3). \end{aligned}$$

Canceling the zero order term gives

$$T = \frac{2\pi}{\sqrt{b}} + O(\lambda).$$

To find a zero order approximation of τ, written τ_0, it is necessary to use the second order expansion of F. In this case, the first order term of the expansion of T, written as T_1, is involved and the cancellation of the second order term of F (or of $F(x, y-x)$) is written as

$$-ab^2T_1^2 - a\pi^2 + \pi\sin(2\sqrt{b}\tau_0) = 0.$$

Note that the system obtained by the transformation $\lambda \to -\lambda$ and $t \to -t$ has a phase portrait that is obtained from the original system taking the symmetric with respect to the line $w = 0$. We then have $T(-\lambda) = T(\lambda)$, $\tau(-\lambda) = \tau(\lambda)$ and the expansion of T and τ have the form

$$\begin{aligned} T &= T_0 + T_2\lambda^2 + T_4\lambda^4 + \cdots, \\ \tau &= \tau_0 + \tau_2\lambda^2 + \tau_4\lambda^4 + \cdots. \end{aligned}$$

In particular, we have $T_1 = 0$ and, when $a \leq \frac{1}{\pi}$, we find two possible values for the first term of the expansion of τ corresponding to the two limit cycles obtained above:

$$\begin{aligned} {}^1\tau_0 &= \frac{1}{2\sqrt{b}}\arcsin(a\pi), \\ {}^2\tau_0 &= \frac{1}{2\sqrt{b}}(\pi - \arcsin(a\pi)). \end{aligned}$$

It is possible to obtain more refined approximations by continuing the series expansion of T and τ using (4.7). The approximation of limit cycles is then given from (4.4), (4.5). It is interesting to note the similarity of the expressions obtained here and those obtained for large r. This result is not surprising given the change of variables (1.7). In Figure 5.1 and Figure 5.2, we show a typical configuration under study. The two limit cycles that have just been characterized correspond to solutions in the classical sense in that they satisfy hypothesis (4.2) and can be obtained by our Fourier analysis.

5.3. Only one admissible solution. From numerical simulations, we observe configurations where there is only one admissible solution for system (4.7). This situation does not only appear when there is a single intersection between C_1 and C_2 since, as we have already mentioned as $a \to 0$, there can be several intersections so that only one of which is suitable. Moreover, it is possible to find exactly two intersections between C_1 and C_2 only one of which is suitable. This situation is illustrated in Figure 5.3. We have therefore detected a single limit cycle that appears to be the stable one. Naturally, the unstable cycle still exists and here we talk about a *generalized solution*, insofar as we cannot define it in the classical sense. From numerical simulations, we observe that the appearance of this *generalized solution* corresponds to a bifurcation of curves C_1 or C_2 in that at least one of these two curves no longer corresponds to a single closed curve (see Figure 5.3).

In the phase plane, the study of the vector field enables us to specify the unstable cycle, called a separatrix because it is the boundary between two domains of attraction. We write the coordinate points $(a, -\lambda a)$, (a, y_B) as A and B, respectively,

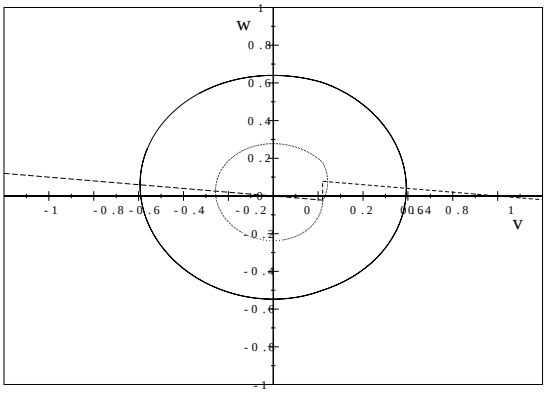


FIG. 5.1. *Unstable (dotted line) and stable (full line) limit cycles of system* (1.5). *The v-nullcline is represented. The parameters are* $\lambda = \mu = 0.1$, $a = 0.22$, $b = 1$.

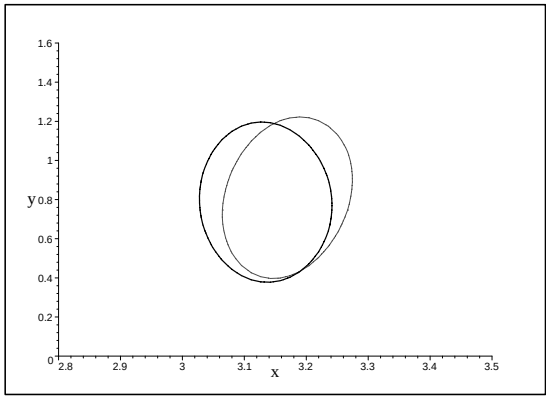


FIG. 5.2. *Curves* C_1 *(thick line) and* C_2 *(thin line). The two intersections correspond to the two limit cycles given in Figure* 5.1 *(the parameters are given in Figure* 5.1*).*

where $y_B \in \mathcal{I}$ and $\mathcal{I}$ designates the interval $[-\lambda a, -\lambda a + \mu]$. Let $\mathcal{P}$ be the parameterized curve obtained when considering the solution of system (1.5) starting from A by reversing the time. The equation for this curve is given by

$$x(t) = a(\cos rt - \tfrac{\lambda}{r} \sin rt) e^{\lambda t},$$
$$y(t) = -a \left(\lambda \cos rt + \tfrac{2}{r}(b - \tfrac{\lambda^2}{2}) \sin rt \right) e^{\lambda t}$$

as long as $x(t) < a$. Let t^* be the smallest real so that $t^* > 0$ and $x(t^*) = a$. If $y(t^*) \leq -\lambda a + \mu$, then we take $y_B = y(t^*)$ and the curve $\Gamma = [A, B] \cup \mathcal{P}$ is the boundary being looked for. This situation is displayed in Figure 5.4. If we now have $y(t^*) > -\lambda a + \mu$, we again consider the solution of system (1.5) by reversing the time but with $(x(t^*), y(t^*))$ as the initial condition. This solution crosses the segment $\mathcal{I}$ at the point B that is looked for. If this solution does not present an intersection with $\mathcal{I}$, we are in the presence of an unstable cycle that can be defined in a classical sense given by the resolution of (4.7). Nevertheless, we have not succeeded in establishing

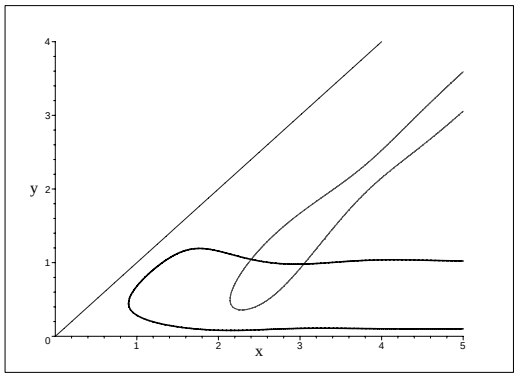


FIG. 5.3. *Curves C_1 (thick line) and C_2 (thin line). The line $y = x$ is represented. The parameters are those of Figure* 5.4.

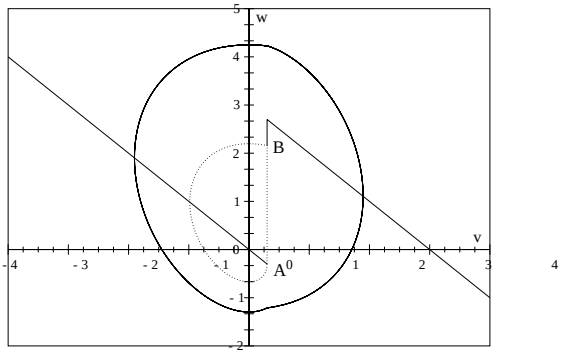


FIG. 5.4. *Stable limit cycle (full line) and unstable limit cycle (dotted line) marking the boundary with the domain of attraction of* $(0,0)$. *The parameters are* $\lambda = 1$, $a = 0.3$, $b = 2$, *and* $\mu = 3$.

precise links between the existence of the point B and the solutions of (4.7).

Another approach is to consider a family of *near* systems, the solutions of those tending towards those of (1.5). From this technique arises the mathematical difficulty of the notion of limit being considered. However, let us define the system

$$\begin{aligned} \frac{dv}{dt} &= p_\delta(v) - w, \\ \frac{dw}{dt} &= bv, \end{aligned} \tag{5.6}$$

where $p_\delta(x) = -\lambda x + \mu h_\delta(x - a)$ and h_δ is the continuous function defined by

$$h_\delta(x) = \begin{cases} 0 & \text{if } x \le 0, \\ \frac{x}{\delta} & \text{if } 0 < x < \delta, \\ 1 & \text{if } x \ge \delta. \end{cases}$$

Numerically speaking, for small values of δ, the orbits of (5.6) are a good approximation of those of system (1.5). This result requires careful study, which we have not undertaken here. The convenience of (5.6) is that they allow the application of classical theorems of existence as well as the usual numerical integration methods like the

Runge–Kutta method. In addition, it seems possible to extend the results obtained for the discontinuous system to these continuous piecewise linear systems as $\delta \to 0$.

6. Large relaxation time. In this section, we study the case of a small λ which corresponds to a system with a large time constant. When $\lambda \ll 1$, the asymptotic expansion of (4.8) is written as

$$F(x,y) = -2a\sqrt{b}(1-\cos(2\sqrt{b}x)) + \lambda\mu(x\sin(2\sqrt{b}y) - y\sin(2\sqrt{b}x)) + O(\lambda^2). \tag{6.1}$$

Therefore we have

$$T = \frac{2\pi}{\sqrt{b}} + O(\lambda),$$
$$\tau = \frac{\pi}{2\sqrt{b}} + O(\lambda).$$

The existence of a periodic solution for small λ has already been noted in section 4. We obtain a single solution for τ which is related to the big cycle. The small cycle cannot be captured by this limiting situation. Using the third order expansion of $F(x,y)$ and $F(x,y-x)$, we find that

$$\begin{aligned} T &= \frac{2\pi}{\sqrt{b}} + \frac{\pi}{4b\sqrt{b}}\lambda^2 + O(\lambda^3), \\ \tau &= \frac{\pi}{2\sqrt{b}} - \frac{a\pi}{2\mu\sqrt{b}}\lambda + \frac{\pi}{16b\sqrt{b}}\lambda^2 + O(\lambda^3). \end{aligned} \tag{6.2}$$

Using the first order expansion of T, we calculate

$$f(t) = -\frac{2\mu}{\lambda\pi}\sin\sqrt{b}t + O(1).$$

Calculation of the approximation of v, using (4.5), (4.4), (6.2), gives

$$v(t) = \frac{2\mu}{\lambda\pi}\cos\sqrt{b}t + O(1). \tag{6.3}$$

The approximation that is obtained coincides with the term carrying the fundamental frequency in the Fourier series of v. Using $w_0 = \frac{\mu}{2}$, the limit cycle approximation is given by

$$v^2 + \frac{1}{b}\left(w - \frac{\mu}{2}\right)^2 = \frac{4\mu^2}{\lambda^2\pi^2} + O(1). \tag{6.4}$$

Numerically speaking, this approximation appears to be a good one, even for large values of λ. It is possible to refine the approximation obtained by using higher order terms in the expansion (6.2). We then find

$$v(t) = \frac{2\mu}{\lambda\pi}\cos\sqrt{b}t + \frac{\mu}{\pi}\left(\frac{\pi}{\sqrt{b}} - t\right)\cos\sqrt{b}t + O(\lambda). \tag{6.5}$$

Remark 1. The terms in the expansion of v have zero mean value.

Remark 2. Approximation (6.5) must be considered for $t \in [0,T]$. This raises the problem of matching at T, a problem that we will not discuss here since we will use

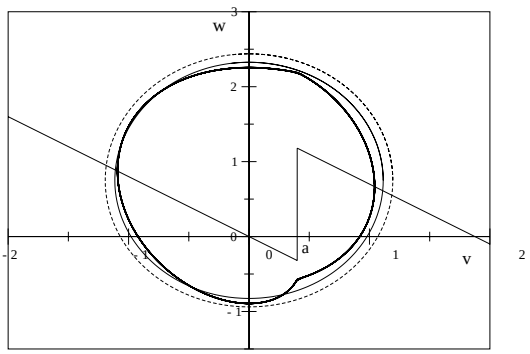


FIG. 6.1. *Stable limit cycle of system* (1.5) *(thick line) and its approximations given by* (6.6) *(thin line) and* (6.4) *(dotted line). The parameters are* $\lambda = 0.8$, $\mu = 1.5$, $a = 0.4$, *and* $b = 2$.

approximation (6.3). From a numerical point of view, this approximation appears to be better for a wide range of values of λ. This is due to the appearance of secular terms in the asymptotic expansion (6.5).

Remark 3. If the expansion of T is continued, there is no term of third order, which leads us to believe that T presents an even power series expansion.

It is interesting to compare the approximation that has just been calculated with the one previously obtained (5.5). The value found for the largest cycle, in the case of small (λ, μ), gives

$$v^2 + \frac{1}{b}\left(w - \frac{\mu}{2}\right)^2 = \frac{2\mu^2}{\lambda^2\pi^2}\left(1 + \sqrt{1 - \frac{\lambda^2\pi^2 a^2}{\mu^2}}\right), \tag{6.6}$$

which, for small values of a, corresponds to the approximation (6.4). Numerically speaking, this approximation is very precise, as shown in Figure 6.1.

Using approximations (6.4) and (6.6), we can formulate an approximate necessary condition for the existence of a periodic solution since the expression of the vector field requires that the interior of the limit cycle contains the point $(a, -\lambda a)$, which yields to

$$b > \frac{(\lambda a + \frac{\mu}{2})^2}{d^2 - a^2} \tag{6.7}$$

with $d \in \{d_\lambda, d_{\lambda\mu}\}$, where d_λ^2 and $d_{\lambda\mu}^2$ are the values of the right-hand term of equations (6.4) and (6.6), respectively. Approximation (6.4) imposes the condition $a < \frac{2\mu}{\lambda\pi}$, which is a requirement greater than that given by (6.6). Even far from its validity domain, approximation (6.7) remains useful. When $\lambda = 1$ and $\mu = 1$, Figure 4.1 shows the approximation (6.7) obtained from the study for small λ (curve C_λ) and for small (λ, μ) (curve $C_{\lambda\mu}$). For small values of a, the requirement appears to be a little too strong, in that it imposes $b > \frac{\pi^2}{16}$ when $b > \frac{1}{4}$ would do.

Let us precisely give the bifurcation giving rise to the stable limit cycle for small λ. In this case, the system under study may be considered as a perturbation of

$$\frac{dv}{dt} = \mu h(v - a) - w,$$

(6.8)
$$\frac{dw}{dt} = bv.$$

System (6.8) was previously considered as a perturbation of the Hamiltonian system obtained for $\mu = 0$. However, in this analysis, μ is not considered as a small parameter. The harmonic balance method leads to the following two cases:

- τ does not exist and we find a family of periodic solutions defined by

  (6.9)
  $$H(v, w) = c^2, \quad \text{where} \quad c < a,$$

  where H is given by (5.2).
- If we assume that τ exists, we find that the Fourier series expansion of v is divergent and therefore there is no periodic solution such as $v > a$.

For an initial condition outside the ellipse obtained with $c = a$ in (6.9), a solution of (6.8) tends towards infinity since the orbits of system (6.8) are given by

$$v^2 + \frac{1}{b}w^2 = \text{const} \quad \text{for} \quad v < a,$$
$$v^2 + \frac{1}{b}(w - \mu)^2 = \text{const} \quad \text{for} \quad v > a,$$

and, if we consider the sequence $(w_n)_{n\in\mathbb{N}}$ associated with the Poincaré section defined by $v = a$, we have

$$w_n = w_{n-1} + 2\mu.$$

Thus, the orbits spiral around the origin and move away from it. The addition of the perturbation $-\lambda v$ leads to (i) the destruction of the family of periodic solutions so that $v < a$ (the origin becomes a stable focus) and (ii) the appearance of a limit cycle towards which the orbits converge while spiraling. We have seen that the birth of the limit cycle takes place at ∞ since the diameter of the ellipse can be made arbitrarily large. We are going to specify this result in bifurcation terms.

We write (r, θ) for the polar coordinates of (v, w) and, because we are interested in the system at ∞, we introduce the variable $u = \frac{1}{r}$. Given a change of variables, we can consider the case $b = 1$. Writing (1.5) using the new variables gives

$$\frac{du}{dt} = \lambda u \cos^2\theta - u^2\mu\cos\theta h\left(\frac{\cos\theta}{u} - a\right),$$
$$\frac{d\theta}{dt} = 1 + \lambda\sin\theta\cos\theta - u\mu\sin\theta h\left(\frac{\cos\theta}{u} - a\right).$$

We are interested in the behavior of the system for $\lambda \ll 1$ and u close to 0. In this case, θ is a fast variable, the dynamic of which can be approximated by $\theta' = 1$. The averaging theorem [11] enables us to consider the approximation given by the averaged system

$$\frac{du}{dt} = \frac{1}{2\pi}\int_0^{2\pi} d\theta \quad \lambda u\cos^2\theta - u^2\mu\cos\theta h(\cos\theta),$$

where we have used the approximation $h(\frac{\cos\theta}{u} - a) \sim h(\cos\theta)$ for small $u > 0$.

We find

$$\frac{du}{dt} = \Big(\frac{\lambda}{2} - \frac{\mu}{\pi}u\Big)u,$$

which shows the appearance of a stable limit cycle. The radius of this cycle is given by $u = \frac{\lambda\pi}{2\mu}$ and is in agreement with the approximation (6.4). This is a supercritical Andronov–Hopf bifurcation which appears at ∞. As far as we know, such a bifurcation was mentioned for the first time in [37].

7. Excitability and oscillations. We may interpret the appearance of oscillations as the limit behavior of a *spike solution* when the number of action potentials becomes large. We are going to give mathematical content to this statement by showing that the periodic solution, written as $v_\gamma(t)$, can be obtained as the limit of the *spike solution*, written as $v_n(t)$, when the number of spikes n tends towards $+\infty$. Most often, the birth of oscillations is shown in terms of bifurcations using equations based on system parameters. Here, the characterization is directly obtained from the system solutions.

We consider (3.7), omitting the transient regime containing v_0, because we are interested in the asymptotic state. Using a time shift, we consider the symmetrical sum obtained from (3.7):

$$v_n(t) = \sum_{k=-n}^{n} \phi(t - t_{2k}) - \phi(t - t_{2k+1}),$$

where

$$\phi(t) = h(t)\varphi(t)$$

and φ is given by (3.8). If we assume that the spikes are produced at periodic time intervals, there exist T and τ so that $t_{2k} = kT - \tau$ and $t_{2k+1} = kT + \tau$. The existence of the pair (T, τ) is studied in section 4. We should also note that the assumption just made is linked to v_0 insofar as not all orbits converge towards a periodic solution.

We have $\phi \in L^1(\mathbb{R})$, and the Poisson formula, in the space of tempered distributions $\mathcal{S}'$, gives us

$$\lim_{n\to+\infty} v_n(t) = \frac{1}{T}\sum_{-\infty}^{+\infty} \widehat{\phi}\left(\frac{k}{T}\right) 2\ i\ \sin\left(2\pi\tau\frac{k}{T}\right) e^{2i\pi k\frac{t}{T}}.$$

As ϕ', the distributional derivative of ϕ, is in $L^1(\mathbb{R})$, equality occurs for every t, and we have the uniform convergence of the series. We calculate that

$$\widehat{\phi}(w) = \frac{2\mu}{2b - 8\pi^2 w^2 + 4i\pi\lambda w},$$

giving

$$\lim_{n\to+\infty} v_n(t) = v_\gamma(t),$$

where $v_\gamma(t)$ is the periodic solution given by (4.3), which establishes the stated result.

8. Discussion. Estimation of the maximal number and relative positions of limit cycles of a two-dimensional autonomous system is an open problem corresponding to the second part of the sixteenth Hilbert problem. Given the difficulty of the general problem, mathematicians have become interested in a particular system class, the Liénard system:

$$\begin{aligned} \frac{dv}{dt} &= p(v) - w, \\ \frac{dw}{dt} &= v. \end{aligned} \tag{8.1}$$

Most results concern the case in which p is a polynomial function. Even in this case, there are no general theoretical results and most approaches are local ones insofar as they determine only the number of limit cycles for certain parameter values. Limit cycles are obtained using perturbation methods via a Hopf bifurcation or a global bifurcation (see [28] and the references therein). Some global approaches make it possible to link the number of limit cycles to the roots of a polynomial [10], [25], but the results remain to be demonstrated.

We have studied the Liénard system where p is a piecewise linear function (linear on $]-\infty, a[$ and on $]a, +\infty[$) allowing a finite jump discontinuity at $a > 0$. We have shown that the limit cycles are characterized by the roots of a system of two transcendental equations. These roots correspond to the period of the oscillations and to an additional parameter. We have obtained an explicit expression of the limit cycles as a function of these two roots. Our results are in agreement with the local methods in that (i) the fold limit cycle bifurcation can be obtained as a perturbation of a center and (ii) the large size limit cycle can be obtained as a Hopf bifurcation at ∞. We might also consider the limit cycle obtained as $a \to 0$ as a kind of *degenerated* Hopf bifurcation. We have shown the existence of at least two limit cycles, and arguments similar to those used in [27] should enable us to demonstrate that at most two limit cycles exist. When p is a polynomial function, such a result can be obtained only for a polynomial of degree at least five [36]. It has already been observed that discontinuous dynamical systems have a richer dynamic than regular dynamical systems [9]. The obtained results, and numerical simulations that we have carried out, lead us to formulate two conjectures concerning the number of limit cycles of a piecewise linear Liénard system.

Conjecture 1. The Liénard system (8.1), with p piecewise linear on $n+1$ intervals and having n finite jump discontinuity, has up to $2n$ limit cycles.

Conjecture 2. The Liénard system (8.1), with p continuous and piecewise linear on $n+1$ intervals, has up to n limit cycles.

Conjecture 2 generalizes the result obtained in [26], [27] in which the authors proposed a continuous, and piecewise linear on $2n+1$ intervals, function p so that Liénard system (8.1) has exactly n limit cycles. The parity and periodicity of p appear to be the two properties that limit the number of limit cycles.

Beyond mathematical interest of the system under study, it is of great importance in mathematical biology where excitable systems are widely used [31], [17]. Our system is a piecewise linear version of the FitzHugh–Nagumo equations with a simplified version of the recovery process which provides an understanding of the behavior in a transparent way. First of all, we have distinguished between two dynamics according to the value of $\lambda^2 - 4b$. When $\lambda^2 - 4b \geq 0$, the system is termed *leaky integrator* and only a single spike can be emitted in response to an excitation given by the input $I = I_0\delta(t - t_0)$. When $\lambda^2 - 4b < 0$, the system is referred as being *resonator*. In this case, the response is obtained as the superposition of

$$v(t) = e^{-\frac{t}{\eta}} \sin \Omega t,$$

where $\eta = \frac{1}{\lambda}$ and $\Omega = \sqrt{b - \frac{\lambda^2}{4}}$ denote, respectively, the time constant and the natural frequency of the system. When this response is a finite sum, we obtain what we call a *spike solution*. In the case of infinite sum, we obtain a periodic solution for which an analytical expression is given by

$$v = \frac{1}{2}(S_{-\tau}f - S_{\tau}f),$$

where S is the shift operator $S_y g(x) = g(x+y)$ and f is a function that depends on the period T. In the general case, it is not possible to have an explicit expression for T and τ. However, we have obtained a set of approximate solutions which shows that the period is well approximated using

$$T = \frac{4\pi}{\sqrt{4b - \lambda^2}}.$$

We have detected two possible mechanisms for the appearance of oscillations: a fold limit cycle bifurcation and a Hopf bifurcation at infinity.

A significant biological interest is the extension of our analysis to the complete system where the recovery process is given by

$$\frac{dw}{dt} = b(v - \gamma w).$$

In this case, a change of variables allows us to rewrite the FitzHugh–Nagumo system as the generalized Liénard equation

$$\begin{aligned} \frac{dv}{dt} &= F(v) - w, \\ \frac{dw}{dt} &= G(v). \end{aligned} \tag{8.2}$$

When p is the polynomial function (1.2), the two functions F and G are third degree polynomial functions and, in contrast to the case $\gamma = 0$, three limit cycles can be obtained. We plan to explore the piecewise linear case for which an analytical study is possible but yields much more complicated expressions than those obtained in this paper. Results on such an extension will be reported elsewhere.

There remains much work to be done on our system. The simplicity of the model allows us to hope for analytical results for bursting [35]. The coexistence of a limit cycle and a stable fixed point favors the existence of such a phenomenon when an additional slow variable is added to the system. Another aspect is the study of coupled equations. In particular, we hope for promising results concerning the dynamics of coupled oscillators using the approximations obtained for the periodic solution. As a first step, we plan to explore the forced system in the context of forced piecewise linear systems [4], [21].

Acknowledgments. The author thanks J. Demongeot and J. L. Martiel for many helpful discussions.

REFERENCES

[1] J. C. Alexander, E. J. Doedel, and H. G. Othmer, *On the resonance structure in a forced excitable system*, SIAM J. Appl. Math, 50 (1990), pp. 1373–1418.

[2] D. J. Allwright, *Harmonic balance and the Hopf bifurcation*, Math. Proc. Cambridge Philos. Soc., 82 (1977), pp. 453–467.

[3] S. Amari, *Dynamics of pattern formation in lateral-inhibition type neural fields*, Biol. Cybernetics, 27 (1977), pp. 77–87.

[4] J. Belair and P. Holmes, *On linearly coupled relaxation oscillations*, Quart. Appl. Math., 42 (1984), pp. 193–219.

[5] M. S. Branicky, V. S. Borkar, and S. K. Mitter, *A unified framework for hybrid control*, in Proceedings of the 33rd IEEE Conference on Decision and Control, Lake Buena Vista, FL, 1994, pp. 4228–4234.

[6] J. W. Cahn, J. Mallet-Paret, and E. S. Van Vleck, *Traveling wave solutions for systems of ODEs on a two-dimensional spatial lattice*, SIAM J. Appl. Math., 59 (1998), pp. 455–493.
[7] A. F. Filippov, *Differential Equations with Discontinuous Righthand Sides*, Kluwer, Dordrecht, The Netherlands, 1988.
[8] R. FitzHugh, *Impulses and physiological states in theoretical models of nerve membrane*, Biophys. J., 1 (1961), pp. 445–466.
[9] A. Gasull and J. Torregrosa, *Center-focus problem for discontinuous planar differential equations*, Int. J. Bifurcation and Chaos, to appear.
[10] H. Giacomini and S. Neukirch, *Number of limit cycles of the Liénard equation*, Phys. Rev. E (3), 56 (1997) pp. 3809–3813.
[11] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer-Verlag, New York, 1983.
[12] P. Hagedorn, *Non-Linear Oscillations*, 2nd ed., Oxford Engrg. Sci. Ser. 10, Oxford University Press, New York, 1988.
[13] B. D. Hassard, N. D. Kazarinoff, and Y. H. Wan, *Theory and Application of Hopf Bifurcation*, Cambridge Univerity Press, Cambridge, UK, 1981.
[14] J. Imura and A. J. van der Schaft, *Characterization of well-posedness of piecewise linear systems*, IEEE Trans. Automat. Control, 45 (2000), pp. 1600–1619.
[15] E. M. Izhikevich, *Neural excitability, spiking and bursting*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 10 (2000), pp. 1171–1266.
[16] J. P. Keener, *Propagation of waves in an excitable medium with discrete release sites*, SIAM J. Appl. Math., 61 (2000), pp. 317–334.
[17] J. P. Keener and J. Sneyd, *Mathematical Physiology*, Springer-Verlag, New York, 1998.
[18] M. Kunze, *Non Smooth Dynamical Systems*, Springer-Verlag, Berlin, 2000.
[19] Y. Kuramoto, *Chemical Oscillations, Waves, and Turbulence*, Springer-Verlag, New York, 1984.
[20] Y. A. Kuznetsov, *Elements of Applied Bifurcation Theory*, Springer-Verlag, New York, 1994.
[21] M. Levi, *Qualitative Analysis of the Periodically Forced Relaxation Oscillators*, Mem. Amer. Math. Soc., 32 (1981), pp. 1–147.
[22] N. Levinson and D. Smith, *A general equation for relaxation oscillations*, Duke Math. J., 9 (1942), pp. 382–403.
[23] A. Liénard, *Etude des oscillations entretenues*, Rev. Gen. d'électricité, 23 (1928), p. 901.
[24] A. Lins, W. De Melo, and C. C. Pugh, *On Liénard's equations*, in *Geometry and Topology,* Lecture Notes in Math. 597, Springer-Verlag, Berlin, 1977, pp. 335–357.
[25] J. Llibre, L. Pizarro, and E. Ponce, *Limited cycles of polynomial Liénard systems. Comment on: "Number of limit cycles of the Liénard equation" [Phys. Rev. E* (3) 5b (1997) *no.* 4, 3809–3813 *by H. Giacomini and S. Neukirch],* Phys. Rev. E (3), 58 (1998), pp. 5185–5187.
[26] J. Llibre and E. Ponce, *Piecewise linear feedback systems with arbitrary number of limit cycles*, Int. J. Bifurcation and Chaos, to appear.
[27] J. Llibre, E. Ponce, and X. Zhang, *Existence of Piecewise Linear Differential Systems with Exactly n Limit Cycles for all* $n \in \mathbb{N}$, preprint, Universitat Autònoma de Barcelona, Barcelona, Spain, 2001.
[28] S. Lynch, *Liénard systems and the second part of Hilbert's sixteenth problem*, Nonlinear Anal., 30 (1997), pp. 1395–1403.
[29] M. P. McKean, *Nagumo's equation*, Adv. Math., 4 (1970), pp. 209–223.
[30] N. Minorsky, *Nonlinear Oscillations*, Van Nostrand, New York, 1962.
[31] J. D. Murray, *Mathematical Biology*, Springer-Verlag, New York, 1989.
[32] J. S. Nagumo, S. Arimoto, and S. Yoshizawa, *An active pulse transmission line simulating nerve axon*, Proc. IRE, 50 (1962), pp. 2061–2071.
[33] J. Rinzel and G. B. Ermentrout, *Axis of neural excitability and oscillations,* in Methods in Neural Modeling: From Synapses to Networks, C. Koch and I. Segev, eds., MIT Press, Cambridge, MA, 1989, pp. 135–166.
[34] J. Rinzel and J. B. Keller, *Traveling wave solutions of nerve conduction equation*, Biophys. J., 13 (1973), pp. 1313–1337.
[35] J. Rinzel and Y. S. Lee, *On different mechanisms for membrane potential bursting*, in Nonlinear Oscillations in Biology and Chemistry, Lecture Notes in Biomath. 66, H. G. Othmer, ed., Springer-Verlag, New York, Heidelberg, Berlin, 1986, pp. 19–33.
[36] G. S. Rychkov, *The maximum number of limit cycles of the system* $\dot{x} = y - a_1x^3 - a_2x^5$, $\dot{y} = -x$ *is equal to two*, Differ. Uravn., 11 (1975), pp. 380–391.
[37] M. Sabatini, *Hopf bifurcation from infinity*, Rend. Sem. Mat. Univ. Padova, 78 (1987), pp. 237–

253.
[38] B. Van der Pol, *Forced oscillations in a circuit with nonlinear resistance (receptance with reactive diode)*, London, Edinburgh and Dublin Phil. Mag., 3 (1927), pp. 65–80.
[39] W. P. Wang, *Multiple impulse solutions to McKean's caricature of nerve equation.* I—*Existence*, Comm. Pure. Appl. Math., 41 (1988), pp. 71–103.
[40] W. P. Wang, *Multiple impulse solutions to McKean's caricature of nerve equation.* II—*Stability*, Comm. Pure. Appl. Math., 41 (1988), pp. 997–1025.

SIAM J. APPL. MATH.
Vol. 63, No. 2, pp. 485–509

WAVE PROPAGATION IN SPATIALLY DISTRIBUTED EXCITABLE MEDIA*

JIANBO YANG†, SERAFIM KALLIADASIS†, JOHN H. MERKIN‡, AND STEPHEN K. SCOTT§

Abstract. Consider wave propagation through regions of no excitability (gaps) in an otherwise excitable medium. Propagation in the gaps takes place via simple diffusion. We extend the geometric method for a one-gap system developed by Lewis and Keener to the case of two and three gaps, and we obtain conditions for successful wave propagation and failure. We show that, like the one-gap system, steady-state multiplicity for the case of two gaps arises via a limit point bifurcation. We also demonstrate that in some cases the presence of a large number of gaps promotes wavefront propagation.

Key words. spatially heterogeneous excitable media, bistable, successful propagation/failure

AMS subject classifications. 34B40, 35K15, 35K20, 35K57, 35R05, 92B05, 92C30

PII. S0036139901391409

1. Introduction.

Wave propagation in spatially distributed media is relevant to a large variety of biological and chemical systems, including nerve signal transmission, population dynamics, and combustion. There are many different ways in which spatial inhomogeneity can be created. In all cases, spatial variation of the parameters in reaction-diffusion equations has been shown to have important consequences for traveling wave propagation and pattern formation. For example, Kay and Sherratt [1, 2] considered the effect of spatial variation of the demographic kinetic parameters on predator-prey systems. They were able to demonstrate that a small amount of noise has no appreciable effect on the die-out of the regular oscillations after the invasion of a front of predators with an irregular wake. However, moderate to large levels of noise could lead to the persistence of regular oscillations, but generally with a spatial frequency different from that which would normally be generated behind the invasion front. For a scalar reaction-diffusion equation with a cubic kinetic term and with a spatially varying diffusion coefficient, Xin [3] showed that, if the variation of the diffusion coefficient from its mean is sufficiently large, traveling waves no longer exist, so that a wavefront will begin to propagate from given initial conditions but will then stop advancing—a phenomenon known as "quenching" or "wave-block." Several other authors have analyzed propagation with spatially varying diffusion coefficients including wave-blocking phenomena in bistable reaction-diffusion systems [4, 5] and the excitable FitzHugh–Nagumo equation [6].

In this paper we study wave propagation through localized regions of no excitability in an otherwise excitable medium. Following Sneyd and Sherratt [7] and Lewis and Keener [8], we will refer to these regions of no excitability as the "gaps." Such a gap model has relevance in several areas. For instance, the role of the gaps is analogous to

*Received by the editors June 22, 2001; accepted for publication (in revised form) April 10, 2002; published electronically November 19, 2002. This research was supported by the ESF Research Programme Reactor.

http://www.siam.org/journals/siap/63-2/39140.html

†Department of Chemical Engineering, University of Leeds, Leeds LS2 9JT, UK (S.Kalliadasis@leeds.ac.uk). The research of the second author was supported by the University of Leeds and by an ORS award.

‡Department of Applied Mathematics, University of Leeds, Leeds LS2 9JT, UK.

§School of Chemistry, University of Leeds, Leeds LS2 9JT, UK.

that of fire breaks in flame propagation [9]. In the context of Ca^{2+} wave propagation between inositol 1,4,5-trisphosphate (IP_3) receptors [7], the gaps correspond to regions of low IP_3 concentration. The propagation of electrical excitation in cardiac tissue also encounters regions of depressed excitability. For instance, the atrioventricular node, the normal electrical pathway between the atria and ventricular myocardium, is a localized region of low excitability and conductivity [8]. Finally, gaps can be used to model demyelination of nerve fibers [10, 11] resulting from diseases of the central nervous system [12]—in this case, the loss of myelin slows or even stops the transmission of action potentials through the nerve cells.

Our excitable medium model is the bistable reaction-diffusion equation

$$u_t = u_{xx} + g(u,x) \tag{1a}$$

on $-\infty < x < +\infty$, $t > 0$, where the function $g(u,x)$ is defined piecewise as follows:

$$g(u,x) = \begin{cases} f(u), & -\infty < x \le x_1,\ x_2 < x \le x_3,\ \dots,\ x_{2n} < x < +\infty, \\ 0, & \text{otherwise.} \end{cases} \tag{1b}$$

The kinetic term f is taken to be the cubic function

$$f(u) = u(1-u)(u-\alpha), \qquad 0 < \alpha < \frac{1}{2}, \tag{1c}$$

that describes excitation outside the n gaps, $[x_i, x_{i+1}]$ for $i = 1, 3, \dots, 2n-1$. Equation (1a) is subject to the boundary conditions

$$u \to 1, \qquad x \to -\infty,$$

$$u \to 0, \qquad x \to +\infty, \tag{1d}$$

with all the x derivatives tending to zero as $x \to \pm\infty$. The bistable equation is essentially a version of the nonlinear cable equation that has been used to describe the flow of electricity along nerve axons [12]. Within the context of nerve impulse transmission in particular, the bistable equation can be used as a model for myelinated nerve axons. The same equation was also adopted by Lewis and Keener [8] as a model system for the atrioventricular node in the heart. These authors studied the existence, stability, and bifurcation properties of the steady states of the bistable equation (1a) in the presence of a single gap. They developed a phase plane/geometric method that allows the derivation of criteria for wave-block: their analysis indicates that wave-block is associated with a limit point bifurcation; i.e., there exists a minimal gap width above which wave-block occurs. Furthermore, they considered different gap dynamics, including a leaky gap with a linear decay of u in addition to diffusion, and a gap with small but nonzero excitability. A gap with linear decay was also considered by Grindrod and Sleeman [10] and Grindrod [11] to describe leakage of ionic transport from the axoplasm into small pockets of plasma held within the myelin sheath, while a general analysis of the steady states of the bistable equation with a spatially varying reaction term has recently been performed by Salazar and Solà-Morales [13].

The mechanism for successful propagation and failure through an active medium with localized regions of no excitability was studied in detail by Poptsova and Guria [14]. These authors performed initial-value computations of the bistable equation as well as the two-variable excitable FitzHugh–Nagumo model, to demonstrate the

existence of a critical gap width for a single-gap system above which wave-block (referred to by the authors as "locking-up") occurs. They also performed initial-value computations for a periodic sequence of gaps and spacings between the gaps, and they demonstrated the existence of a critical separation distance below which wave-block occurs.

The problem of wave propagation in the presence of several gaps separated by different spacings is considered here. This is a substantially more complex situation than propagation in a heterogeneous medium with a single gap. Our analysis parallels the work by Lewis and Keener [8] and extends their geometric method to a hybrid geometric-algebraic method for the case of multiple gaps. The method allows us to obtain criteria for propagation failure in the bistable equation (1), where gap dynamics is governed purely by diffusion, and to gain insight into the underlying dynamical structure of the problem. Although the geometric method facilitates the visualization of the steady states in the phase plane, setting up the equations governing these steady states and solving for the critical quantities directly does give a systematic way to construct these steady states. For two gaps with lengths less than the critical length for a medium with a single gap, our analysis reveals the existence of a critical spacing between the two gaps, below which the system approaches a steady state and propagation is suppressed. However, if the first gap length is larger than the critical length, with the second gap still being less than the critical length, there are two critical values for the spacing: for values of the spacing less than the smaller critical value and greater than the larger critical value, the system approaches a steady state. The same phenomenon occurs in a medium with three gaps, provided that the final gap length is less than the critical length for a medium with a single gap. For a given value of the spacing between the first two gaps, there are *two* critical values for the second spacing: for values of the second spacing less than the smaller critical value and greater than the larger critical value, the system approaches a steady state. This somewhat surprising result implies that the final spacing has a profound impact on wave propagation and, in fact, promotes propagation across purely diffusive regions only when its value is between the two critical values.

Hence, the bistable equation behaves in a dramatically different fashion than the system we studied previously [15], in which a heterogeneous medium was considered using a cubic autocatalysis model with autocatalyst decay in the gaps, which were defined as the regions where the reactant concentration was zero. In this two-variable model, the autocatalyst was taken to diffuse and react with a reactant loaded at a constant initial concentration throughout a reaction domain except in the gap regions. One of our main findings was that if any gap length is larger than a critical value, wave propagation will be suppressed; unlike the case studied here, for three gaps in the reaction domain there was only one critical value for the second spacing. Finally, our predictions from the hybrid geometric-algebraic method are in excellent agreement with numerical solutions of the system in (1) as an initial-value problem.

2. Wave propagation: Success and failure in a single-gap domain. In the absence of heterogeneities, the bistable equation in (1) admits traveling wave solutions that connect the two stable rest states, $u = 0$ and $u = 1$ (see, for example, Keener and Sneyd [12]). The solution is of the form

$$u(x,t) = u(\xi) = \frac{1}{2}\left[1 - \tanh(\sqrt{2}\xi/4)\right], \qquad c = \frac{1-2\alpha}{\sqrt{2}}, \tag{2}$$

where $\xi = x - ct$. Note that the direction of propagation changes at $\alpha = 1/2$. This traveling wave solution is a heteroclinic trajectory that connects the two saddle-points

$(0,0)$ and $(1,0)$ in the (u, u_ξ) phase plane. In addition, the bistable equation exhibits threshold phenomena (see [16]). Specifically, if the initial data is sufficiently small, the solution of the bistable equation approaches zero uniformly in the limit $t \to \infty$. However, there are initial conditions with compact support lying between 0 and 1 for which the solution approaches 1 uniformly for large times and, as a consequence of the comparison theorem for scalar parabolic operators, any two solutions of the bistable equation that are ordered at some time remain ordered for all subsequent times. Hence, initial conditions larger than the threshold will initiate a solution that approaches 1 for large times. Finally, the traveling wave solution of the bistable equation has been shown to be stable by Fife and McLeod [17], and in fact, starting from any initial data that lies between 0 and α as $x \to +\infty$ and between α and 1 as $x \to -\infty$, the solution will become arbitrarily close to some phase shift of the traveling wave solution (2) for sufficiently large times.

The impact of a single gap, of width $W = x_2 - x_1$, on the propagation of an established permanent-form traveling wave was considered in detail by Poptsova and Guria [14] and Lewis and Keener [8]. For W small but nonzero, the dynamics is similar to the spatially uniform case with $W = 0$. Poptsova and Guria [14] and Lewis and Keener [8] demonstrated that, as the front of the wave approaches the gap, it slows down because there is no excitability within the gap. For W sufficiently small, the front is able to supply sufficient u across the gap to excite the upstream side of the gap. Thus, after a delay, the wavefront can propagate through the gap and, after a sufficiently large distance, the front is able to recover fully and reestablish its permanent waveform. As W is increased, the delay increases, but the wavefront is still able to penetrate across the heterogeneity. Eventually, for W larger than some critical value W_c, the solution approaches a spatially inhomogeneous steady state of system (1), with $u_t = 0$ and without any wave development beyond the gap. A similar wave-block phenomenon was observed in the cubic autocatalytic system with decay studied in [15].

This behavior is confirmed from a full numerical solution of the bistable equation (1) as an initial-value problem in an extended domain, with long intervals before the first gap and after the last gap, to establish domain independence and to ensure that u approaches the two rest states 1 and 0 on the space spanned by the eigenvectors obtained from the linearized version of (1) at the infinities. We utilize a standard Crank–Nicolson-type implicit scheme for solving parabolic equations, with the x-derivatives approximated by central differences. (The advantages of an implicit scheme over an explicit scheme are obvious, as an explicit scheme would require very short time steps for a reasonable spatial accuracy.) The translational invariance of the system in x allows us to take $x = 0$ as the starting point of the computations. We start the integrations with $u = 0$ everywhere except in the first 100 grid points, where we set $u = 1.0$. The time-discretization uses a two-level scheme. In advancing from time t to time $t + \Delta t$, we replace the time derivative terms by first-order differences involving the solution at the old time level and the as-yet-unknown solution at the new time level. We evaluate the other terms using a weighted average of the solution at the two time levels. At each time level, the fully discrete system is a set of nonlinear algebraic equations, which we solve using Newton–Raphson iteration. The accuracy of the numerical simulations was determined by careful convergence tests under mesh refinement and time-step sizing. In all cases, grid sizes and time steps were kept smaller than 10^{-1}, while in some cases grid sizes and time steps as small as 10^{-3} were employed to accurately resolve critical gaps and spacings.

Our results confirm the value $W_c \simeq 6.5$ for $\alpha = 0.3$ obtained by Lewis and

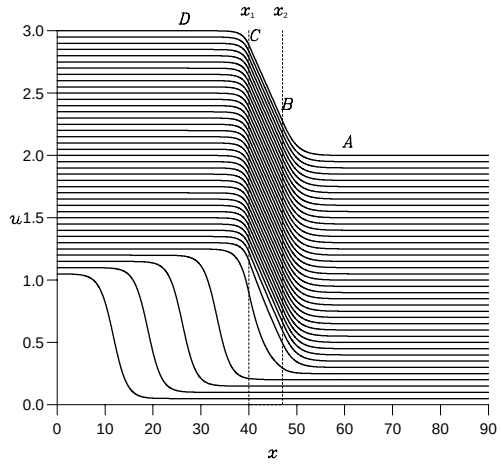


FIG. 2.1. *Numerical solution of* (1) *for* $\alpha = 0.3$ *as an initial-value problem on a domain with* $L = 90.0$ *and a gap of width* $W = 7.0 > W_c$. *Here* u *is plotted at equal times starting at* $t = 25.5$. *The time lapse between any two successive curves is* 22.5 *time units.*

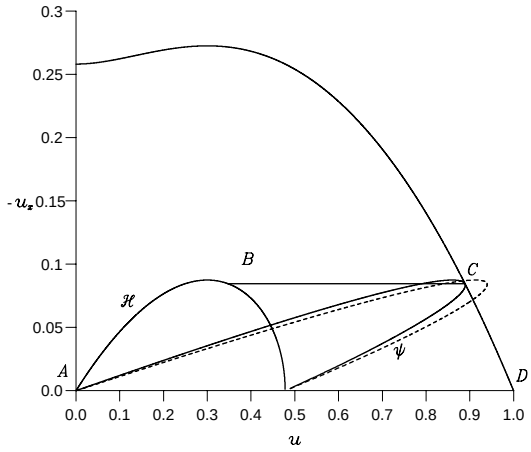


FIG. 2.2. *Phase-portrait of* $u_{xx} + f(u) = 0$ *in the* (u, u_x) *phase plane for the steady-state in Figure* 2.1. $\mathcal{H}$ *indicates the homoclinic orbit emanating from the saddle point* $(0, 0)$, *and* ψ *the transfer map of* $\mathcal{H}$. *Solid curve: curve of* ψ *tangential to the stable manifold of the saddle* $(1, 0)$ *occurring at* $W = W_c = 6.402$. *Dashed curve: curve of* ψ *with* $W = 7.0$.

Keener [8]. Figure 2.1 depicts the evolution of the wavefront towards a gap with width $W = 7.0$. The wavefront slows down and eventually stops. The final result of the evolution is a steady-state solution of the bistable equation. Lewis and Keener developed a phase plane method to construct the steady-state solutions and to predict the critical width W_c. Their method is essentially based on piecing together in the phase plane the invariant manifolds that constitute the steady-state solutions. Below we offer an algebraic version of the geometric method of Lewis and Keener.

Figure 2.2 shows the steady state of Figure 2.1 in the (u, u_x) phase plane. Three

curves in the phase portrait are of particular interest: (i) the curve $\mathcal{H}$, which is a portion of the homoclinic orbit emanating from the saddle point at (0,0) and is described by $u_x = -\sqrt{-2\int_0^u f(v)dv}$; (ii) the transfer map of the homoclinic orbit by the flow in the gap which maps points on $\mathcal{H}$ onto a new curve ψ_W defined by $\psi_W : [u, u_x] \mapsto [-u_x W + u, u_x]$; (iii) the stable manifold of the saddle point $(u, u_x) = (1, 0)$ described by $u_x = -\sqrt{2\int_u^1 f(v)dv}$. Hence, the equations describing the steady state are

$$u_x^B = -\sqrt{-2\int_0^{u^B} f(v)dv}, \tag{3a}$$

$$u^C = -u_x^B W + u^B, \tag{3b}$$

$$u_x^C = -\sqrt{2\int_{u^C}^1 f(v)dv}, \tag{3c}$$

$$u_x^B = u_x^C, \tag{3d}$$

where the superscripts B and C refer to points B and C, respectively, in the phase plane and in Figure 2.1. The above are four equations for the five unknowns u_x^B, u^B, u_x^C, u^C, and W. The existence of real solutions to these equations ensures that the transfer map and the stable manifold of (1,0) intersect. However, the intersection can be either tangent or transversal. To ensure a tangent intersection, and hence intersection at only one point, the slopes of the transfer map and the stable manifold must coincide at point C. For this purpose we must express ψ as a function of u. However, it is not possible to obtain an explicit expression for $\psi(u)$, and therefore we resort to the parametric representation

$$u = -W\sqrt{-2\int_0^t f(v)dv} + t,$$

$$\psi \equiv u_x = -\sqrt{-2\int_0^t f(v)dv},$$

where $0 \le t \le t_{\max}$ with $t_{\max} = (2/3)(1+\alpha)+(1/3)\sqrt{4-10\alpha+4\alpha^2}$. Hence, the slope ψ_u can be easily obtained: $\psi_u = (du_x/dt)/(du/dt)$. At point C, u_x evaluated from ψ, $u_x = -\sqrt{-2\int_0^{t^C} f(v)dv}$, must be equal to u_x evaluated from the stable manifold, $u_x = -\sqrt{2\int_{u^C}^1 f(v)dv}$. A comparison of $\sqrt{-2\int_0^{t^C} f(v)dv} = \sqrt{2\int_{u^C}^1 f(v)dv}$ with (3a), (3c), and (3d) indicates that $t^C = u^B$. Hence, the condition of equal slopes at C,

$$\psi_u(t^C) = \frac{f(u^C)}{\sqrt{-2\int_1^{u^C} f(v)dv}},$$

can be expressed in terms of u^B and u^C and, after being combined with (3a)–(3d), gives a system of three nonlinear algebraic equations:

$$\int_0^{u^B} f(\upsilon)d\upsilon = \int_1^{u^C} f(\upsilon)d\upsilon, \tag{4a}$$

$$f(u^B)f(u^C) = 2\frac{f(u^B) - f(u^C)}{u^B - u^C}\int_0^{u^B} f(\upsilon)d\upsilon, \tag{4b}$$

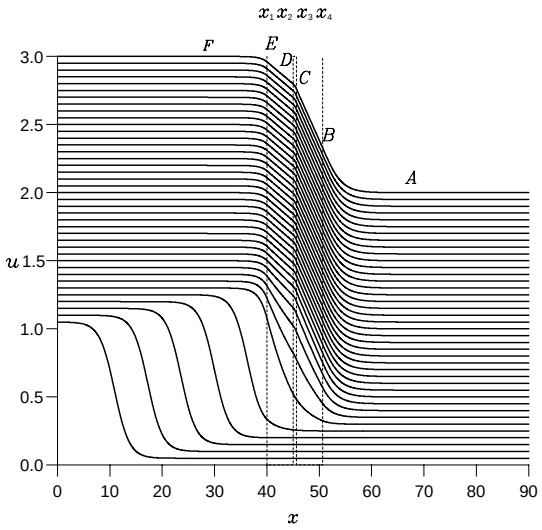
$$W_c = \frac{u^C - u^B}{\sqrt{-2\int_0^{u^B} f(\upsilon)d\upsilon}}. \tag{4c}$$

The first two equations can be solved with a simple trial-and-error procedure to obtain u^B and u^C. The critical gap width W_c can then be obtained explicitly from (4c). For $\alpha = 0.3$, $W_c = 6.402$. An asymptotic analysis of these equations as $\alpha \to 1/2$ shows that $W_c \sim 2\sqrt{2}[(1-2\alpha)/12]^{1/3}$ with $u^{B,C} \sim 1 - [(1-2\alpha)/12]^{1/3}$.

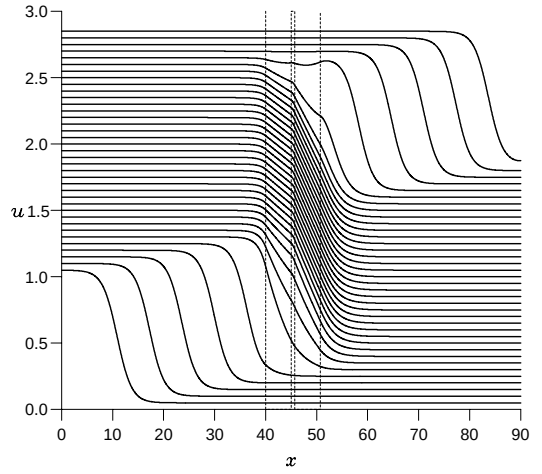
Finally, two points of interest in the phase portrait are u_1, the maximum u value on the homoclinic orbit $\mathcal{H}$ (see Figure 2.2), and u_2, the u value on the stable manifold of (1,0) with a slope the maximum of $-u_x$ on the homoclinic orbit. Lewis and Keener [8] pointed out that different kinetic terms f that are the same on the intervals $0 < u < u_1$ and $u_2 < u < 1$ would give the same W_c and the same steady states for $W > W_c$. This is simply due to that fact that f does not play a role in the purely diffusive gap. The kinetic term f, however, is critical in determining the shape of the homoclinic orbit and the stable manifold. The same argument can be readily extended to the case of two- and three-gap systems studied in the next two sections.

3. Propagation in a domain with two gaps. We consider the propagation of a wavefront in a domain with two gaps of widths W_1 and W_2. The region of excitability of length S between the two gaps will boost any wave passing through the first gap before it encounters the second gap. Hence, we seek to determine the minimum separation distance S_c between the two gaps that allows the wavefront to propagate successfully through the whole domain. Numerical integration of (1) for $\alpha = 0.3$ suggests that $S_c \simeq 0.68$ for $W_1 = W_2 = 5$. Figure 3.1(a) shows wave-block for $W_1 = W_2 = 5.0$ (less than the critical value $W_c = 6.402$ for a domain with a single gap) with a spacing $S = 0.66$. Figure 3.1(b) shows successful propagation of the wavefront for $S = 0.69$ (for the same values of W_1 and W_2).

We now extend the geometric method developed by Lewis and Keener to the case with two gaps. Figure 3.2 shows the phase portrait of the steady state in Figure 3.1(a). As with the one-gap case, we can compute the transfer map $\psi_{W_2} : [u, u_x] \mapsto [-u_x W_2 + u, u_x]$ to obtain the mapped curve ψ and the stable manifold of the saddle-point $(1, 0)$. Notice that the transfer map of the homoclinic orbit is now defined by the last gap W_2, while the first gap W_1 corresponds to curve DE emanating from the stable manifold of (1,0) in the phase plane. Another curve of interest in Figure 3.2 is the phase orbit ϕ tangent to ψ. The monotonicity of ψ_{W_2} in W_2 and the continuity of ψ and any orbit in the phase plane ensures that, for a given W_2, there will be only one orbit tangent to the mapped curve ψ. This phase orbit is given by $u_x = -\sqrt{-2\int_0^u f(\upsilon)d\upsilon + \beta}$, where $\beta = u_x^2$ at $u = 0$ for the phase orbit ϕ. We shall demonstrate that this phase orbit determines the critical spacing for propagation failure.



(a)



(b)

FIG. 3.1. (a) *Failure of wavefront propagation in a domain with two gaps,* $W_1 = W_2 = 5.0 < W_c$ *for a spacing between the two gaps of* $S = 0.66$ *with* $\alpha = 0.3$*;* (b) *successful propagation for* $W_1 = W_2 = 5.0$ *and* $S = 0.69$. *Here* u *is plotted at equal times from time* $t = 22.5$. *The time lapse between successive curves is* 22.5 *time units.*

The condition of equal slopes for ψ and ϕ, along with the fact that u_x^C evaluated from ψ must be equal to u_x^C evaluated from ϕ, gives

$$\frac{f(t^C)}{W_2 f(t^C) + \sqrt{-2\int_0^{t^C} f(v)dv}} = \frac{f(u^C)}{\sqrt{-2\int_0^{t^C} f(v)dv}}, \tag{5a}$$

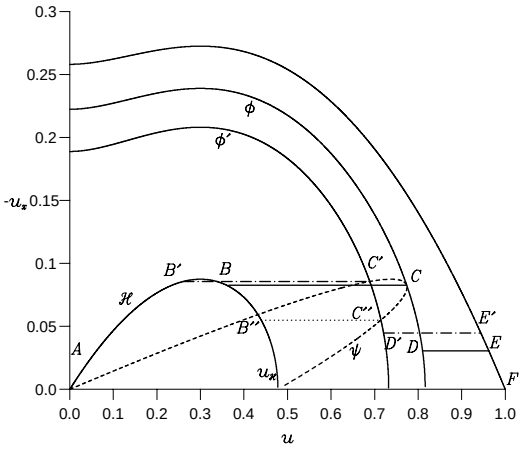


FIG. 3.2. *Phase portrait of $u_{xx} + f(u) = 0$. The phase orbit ϕ determines the critical spacing between the two gaps. The tangent intersection between ϕ and ψ occurring at $\beta = 0.04949$ corresponds to the steady state in Figure* 3.1(a). *The phase orbit through C' and C'' corresponds to $\beta = 0.0357$. The curve $AB'C'D'E'F$ is a stable steady state for $S = 0.49 < S_c$ and $W_1 = W_2 = 5.0$. The curve $AB''C''D'E'F$ is an unstable steady state for $W_1 = W_2 = 5.0$ and $S = 0.13$. Here $u_{\mathcal{H}}$ corresponds to the intersection of the homoclinic orbit with the u-axis.*

where

$$u^C = -W_2\sqrt{-2\int_0^{t^C} f(v)dv} + t^C. \tag{5b}$$

The above form a system of two equations for the two unknowns u^C and t^C, which can be easily solved numerically. The constant β for the phase orbit ϕ can then be obtained from

$$\beta = 2\left\{\int_0^{u^C} f(v)dv - \int_0^{t^C} f(v)dv\right\}, \tag{5c}$$

which fully determines point C and the orbit ϕ. Point B can now be determined from the second gap, which is purely diffusive: $u^C = u^B - u_x^B W_2$, where $u_x^B = u_x^C = -\sqrt{-2\int_0^{t^C} f(v)dv}$, or

$$u^B = u^C - W_2\sqrt{-2\int_0^{t^C} f(v)dv}, \tag{5d}$$

which fully determines point B. We can now locate points D and E from the requirement that u_x^E evaluated from the stable manifold must be equal to u_x^D evaluated from ϕ, which when combined with (5c) yields

$$\int_0^{u^C} f(v)dv - \int_0^{t^C} f(v)dv = \int_0^{u^D} f(v)dv - \int_{u^E}^1 f(v)dv. \tag{6a}$$

TABLE 3.1
Lower critical spacing between two gaps of width W_1 and W_2 obtained from the initial-value problem and the geometric method for $\alpha = 0.3$.

W_1	W_2	Equation (7)
4.0	4.0	0.32
5.0	5.0	0.65
6.0	6.0	1.36

The final equation is then obtained from the first gap: $u^E = u^D - W_1 u_x^E$, with $u_x^E = -\sqrt{2\int_{u^E}^1 f(\upsilon)d\upsilon}$ as obtained from the stable manifold. We therefore have

$$u^E - u^D = W_1\sqrt{2\int_{u^E}^1 f(\upsilon)d\upsilon}. \tag{6b}$$

A comparison of $u_x^B = -\sqrt{-2\int_0^{u^B} f(\upsilon)d\upsilon}$ with $u_x^C = -\sqrt{-2\int_0^{t^C} f(\upsilon)d\upsilon}$ shows that $u^B \equiv t^C$. We then have four equations, (5a), (5d), (6a), and (6b), for the four unknowns u^B, u^C, u^E, and u^D. We can now utilize the phase orbit ϕ to obtain an explicit expression for the critical spacing S_c between the two gaps:

$$S_c = \int_{u^C}^{u^D} \frac{du}{\sqrt{-2\int_0^u f(\upsilon)d\upsilon + \beta}}. \tag{7}$$

For $S < S_c$, wave-block occurs. The dashed-dot line in Figure 3.2 depicts one such steady state for $S = 0.49$; as in the single-gap case, there are actually two steady states, one of which stable (associated with C') as we shall demonstrate later on.

The steady state in Figure 3.1(a) can also be constructed with an alternative method. The phase orbit ϕ in Figure 3.2 can be mapped by the flow in the first gap W_1. The mapped curve, say ψ', can be easily obtained as follows: simple diffusion in the first gap implies $u^E - u^D = -W_1 u_x^D$, which, with u_x^D evaluated from ϕ, gives $u^E - u^D = W_1\sqrt{-2\int_0^{u^D} f(\upsilon)d\upsilon + \beta}$, the equation for ψ'. This curve, not shown in Figure 3.2, will intersect the stable manifold transversally exactly at point E. (It can be shown that the point at which ϕ intersects the $-u_x$ axis will be mapped into the region to the right of the stable manifold of (1,0).) We can then easily locate point D (DE is parallel to the u-axis), and hence a simple integration on ϕ from C to D will give S_c.

Table 3.1 shows the critical spacing obtained from (7) (obviously one must first use (5a), (5c), (5d) and (6a), (6b) to obtain u^B, u^C, u^D, u^E, and β) and the initial-value problem for different values of $W_1 = W_2$. We notice that S_c increases as the gap width increases. The variation of S_c as a function of W_2 for given W_1 ($< W_c$) as obtained from (7) is given in Figure 3.3. Clearly, for a domain with two equal gaps, wave propagation will fail at the final gap if $W_1 = W_2 > W_c$, where W_c is the critical gap width for the single-gap domain. On the other hand, if $W_1 = W_2 = W_c/2$, the wavefront will propagate even if both gaps are brought together, so $S_c \to 0$ in this limit. All curves in Figure 3.3 blow-up to infinity as W_1 approaches W_c from below: from Figure 3.2, as W_2 approaches W_c, point C moves close to the stable manifold of (1,0), while the segment DE moves closer to the u-axis.

For a domain with two unequal gaps, with $W_2 < W_c$ and $W_1 < W_c$, it is clear that, with W_1 fixed, $S_c \to \infty$ as $W_2 \to W_c$. Also if $W_2 < W_c - W_1$, $S_c \to 0$ and

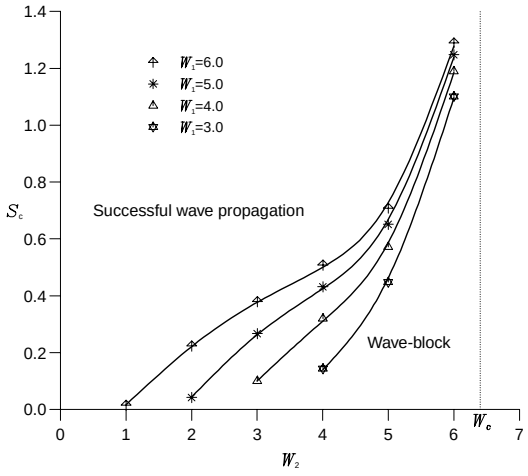
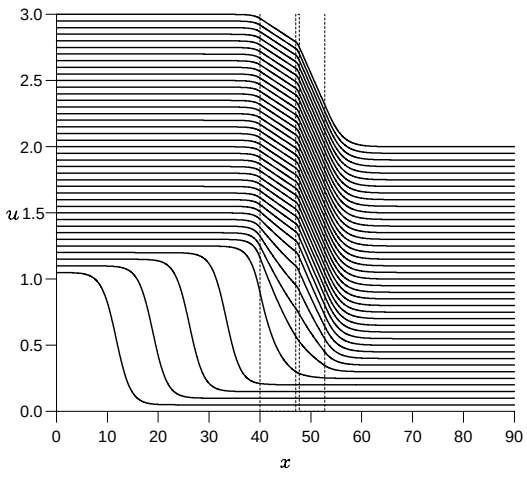


FIG. 3.3. *Critical spacing S_c for a two-gap system as a function of W_2 for different W_1 values with $W_1 < W_c$, obtained from* (7).

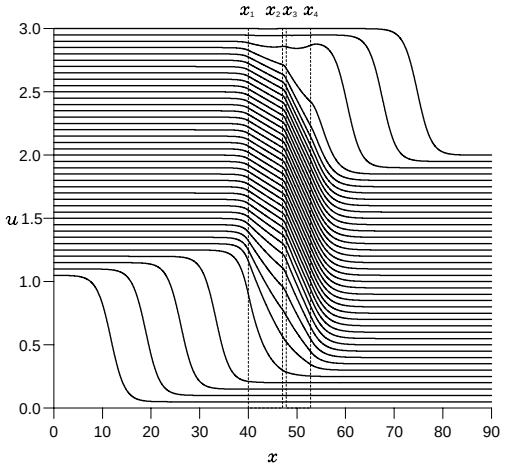
the wavefront will propagate successfully in this case even if the two gaps merge. However, for $W_1 > W_c$ the situation can be dramatically different: from Figures 3.4(a) and 3.4(b), where $W_1 = 7.0 > W_c$ and $W_2 = 5.0 < W_c$, we see that there is a critical spacing length between the two gaps for successful propagation. From the initial-value problem, we obtain $S_c \simeq 0.765$ in this case. This phenomenon is quite different from what we found in [15], where we showed that, if any gap length is larger than the critical length for a domain with a single gap, wave propagation is always blocked. Of course, the fundamental differences between the two systems are that the system in [15] is a two-variable model and the dynamics in the gap is characterized by a linear decay of the autocatalytic intermediate species in addition to pure diffusion. Furthermore, we anticipate wave-block when $S \to \infty$ as Figure 2.1 shows. This then implies that there exists a *second* critical length for S above 0.765. From the initial-value problem, we find that this second critical spacing is approximately 5.13; see Figure 3.5 for a numerical experiment with $W_1 = 7.0$, $W_2 = 5.0$, and $S = 5.13$. This is also different from what we met in [15], where a sufficiently large spacing promotes propagation: for the present problem with two gaps, a sufficiently large spacing can inhibit propagation across the purely diffusive region. These observations do not necessarily imply that, whenever there is decay in the gap (as in [15]), facilitation of propagation will not occur. Indeed, we anticipate that facilitation could certainly occur if the decay were sufficiently small with respect to the negative portion of the kinetic term f.

Interestingly, in some circumstances, successful propagation can occur in one direction but not in the other direction; i.e., the assymmetry due to multiple gaps can induce unidirectional block. For instance, consider the case $W_1 = 7$ and $W_2 = 5$. For a spacing S between the two critical spacings, the wave will propagate successfully; however, reversing the direction of propagation, i.e., $W_1 = 5$ and $W_2 = 7$, will block the wavefront for all S as $W_2 > W_c$ for a single-gap system.

Figure 3.6 is a phase plane representation of the steady state in Figure 3.5. ψ_2 is the curve obtained from mapping the homoclinic orbit $\mathcal{H}$ for $W_2 = 5$. However, unlike



(a)



(b)

FIG. 3.4. (a) *Failure of wavefront propagation in a domain with two gaps,* $W_1 = 7.0 > W_c$, $W_2 = 5.0 < W_c$, *for a spacing between the two gaps* $S = 0.75$ *with* $\alpha = 0.3$; (b) *successful propagation for* $W_1 = 7.0 > W_c$, $W_2 = 5.0 < W_c$, *and* $S = 0.78$. *Here* u *is plotted at equal times starting at* $t = 25.5$. *The time lapse between successive curves is* 25.5 *time units.*

the phase portrait in Figure 3.2, where point B is located in the region $u > \alpha$, point B in Figure 3.6 is located to the left of the maximum value of $-u_x$ for the homoclinic orbit. Point B is then mapped into point C, with BC parallel to the u axis. There is a periodic orbit ϕ within $\mathcal{H}$ passing through C, and one can use this periodic orbit to define CD in the spacing between the two gaps. Point D is then connected to point E on the stable manifold of (1,0), with point E defined from the curve ψ_1 resulting from the mapping of the periodic orbit ϕ for $W_1 = 7$. We shall demonstrate that this periodic orbit determines the second critical spacing for a two-gap system.

An algebraic representation of the phase portrait can be obtained as follows: the

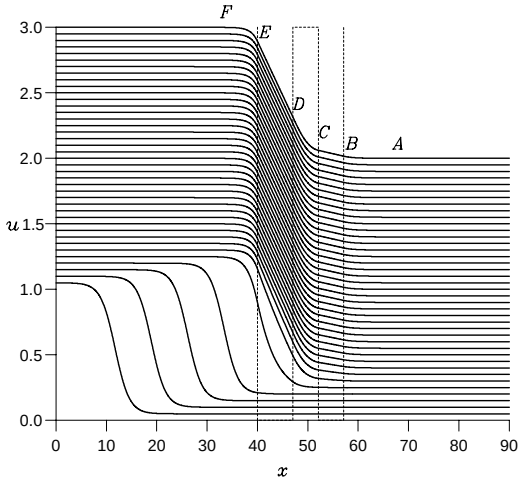


FIG. 3.5. *Wave propagation in a two-gap domain with* $W_1 = 7.0 > W_c$, $W_2 = 5.0 < W_c$, *and* $S = 5.13$, *leading eventually to wave-block. Contrast with Figure* 3.4(b), *where wave propagation is successful. Here* u *is plotted at equal times from time* $t = 25.5$, *with a time lapse between successive curves of* 25.5 *time units.*

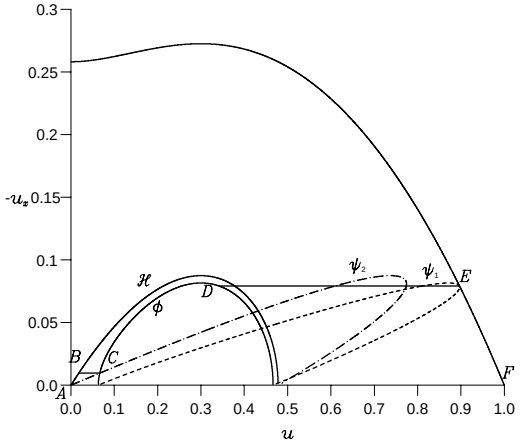


FIG. 3.6. *Phase portrait of the steady state in Figure* 3.5. *Curve* ψ_2 *results from mapping the homoclinic orbit* $\mathcal{H}$, *and* ψ_1 *from mapping the periodic orbit* ϕ *within* $\mathcal{H}$.

equation of any periodic orbit within $\mathcal{H}$ is given by

$$u_x = -\sqrt{\frac{1}{2}u^4 - \frac{2(\alpha+1)}{3}u^3 + \alpha u^2 - \beta},$$

where $\beta \in [0, \beta_{\max} = -(1/6)\alpha^4 + (1/3)\alpha^3]$, a constant parameterizing the periodic orbits within $\mathcal{H}$, with $\beta = \beta_{\max}$ corresponding to point $(\alpha, 0)$. We now utilize the

parametric representation of ψ_1,

$$u = -W_1\sqrt{\frac{1}{2}t^4 - \frac{2(\alpha+1)}{3}t^3 + \alpha t^2 - \beta} + t \equiv f_1(t,\beta),$$

$$\psi_1 = u_x = -\sqrt{\frac{1}{2}t^4 - \frac{2(\alpha+1)}{3}t^3 + \alpha t^2 - \beta} \equiv f_2(t,\beta),$$

from which the slope of ψ_1 with respect to u can be easily obtained as $du_x/du = (du_x/dt)/(du/dt) = f_1(t,\beta)/f_2(t,\beta)$. To ensure a tangent intersection at E, the slope of ψ_1 with respect to u must be equal to the slope of the stable manifold at E:

$$\frac{f(u^E)}{\sqrt{-2\int_1^{u^E} f(\upsilon)d\upsilon}} = \frac{f_2(t^E,\beta)}{f_1(t^E,\beta)}. \tag{8}$$

The requirement that u_x^E evaluated from ψ_1 equal u_x^E evaluated from the stable manifold gives

$$\sqrt{-2\int_1^{u^E} f(\upsilon)d\upsilon} = \sqrt{\frac{1}{2}(t^E)^4 - \frac{2(\alpha+1)}{3}(t^E)^3 + \alpha(t^E)^2 - \beta}, \tag{9}$$

and, as ED represents a purely diffusive process,

$$u^D = u^E - W_1\sqrt{-2\int_1^{u^E} f(\upsilon)d\upsilon}. \tag{10}$$

With $t^E \equiv u^D$, equations (8), (9), and (10) are a system of three equations for the three unknowns u^E, u^D, and β. With BC now a purely diffusive process and $u_x^B = u_x^C$ with u_x evaluated from the homoclinic orbit, we obtain

$$u^C = u^B + W_2\sqrt{-2\int_0^{u^B} f(\upsilon)d\upsilon}. \tag{11}$$

The final equation originates from the requirement that u_x evaluated from the homoclinic orbit at B equal u_x evaluated from the periodic orbit at C:

$$-\sqrt{\frac{1}{2}(u^C)^4 - \frac{2(\alpha+1)}{3}(u^C)^3 + \alpha(u^C)^2 - \beta} = \sqrt{-2\int_0^{u^B} f(\upsilon)d\upsilon}. \tag{12}$$

Thus (11) and (12) form a system of two equations for the two unknowns u^C and u^B. The periodic orbit ϕ can then be used to obtain an explicit expression for the second critical spacing $S_c^{'}$,

$$S_c^{'} = \int_{u^C}^{u^D} \frac{du}{\sqrt{\frac{1}{2}u^4 - \frac{2(\alpha+1)}{3}u^3 + \alpha u^2 - \beta}}, \tag{13}$$

where $\beta = 9.897 \times 10^{-4}$ for $W_1 = 7$ and $W_2 = 5$. The second critical spacing obtained from (13) is $S_c^{'} \simeq 5.15$, in excellent agreement with the 5.13 value obtained from the initial-value problem in Figure 3.5. The agreement between the expression in (13)

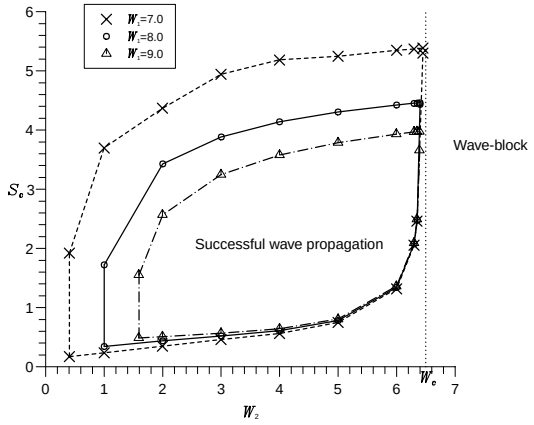


FIG. 3.7. *Critical spacings for a two-gap system as a function of W_2 for different W_1 values with $W_1 > W_c$.*

and the initial-value problem improves as we increase the domain size and decrease the space and time steps in our numerical scheme.

Figure 3.7 shows the variation of S_c with W_2 for given W_1 ($> W_c$). Notice the existence of a closed region in the $S - W_2$ parameter space, outside of which there is block and inside of which there is successful propagation. As W_1 increases, the region for successful propagation shrinks, while as W_1 approaches W_c from above, the turning point close to the line $W_2 = W_c$ moves towards infinity.

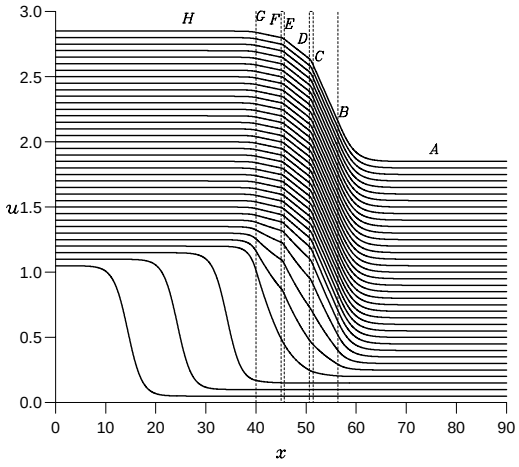
4. Propagation with three gaps. Figure 4.1(a) shows wave-block for $W_1 = W_2 = W_3 = 5 < W_c$ with $\alpha = 0.3$. The first spacing is $S_1 = 0.66$, and the second spacing $S_2 = 0.69$. Figure 4.1(b) depicts the successful propagation of a wavefront for $S_1 = 0.66$, $S_2 = 0.72$. The numerical solution of the bistable equation as an initial-value problem indicates that for $S_1 = 0.66$, $S_{2,c} \simeq 0.71$ for $W_1 = W_2 = W_3 = 5$.

We now construct the steady-state wavefront for a domain with three gaps. Figure 4.2 shows the phase plane portrait of the steady state in Figure 4.1(a). Point C can be found from the solution of (5a), (5b), with W_2 replaced by W_3. Once u^C and t^C are known, β_2 for the phase orbit ϕ_2 can be evaluated from (5c). The remaining equations are

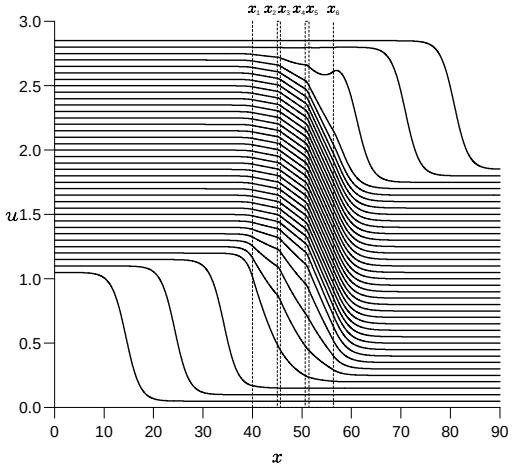
$$u^F = u^G - W_1\sqrt{2\int_{u^G}^{1} f(v)dv}, \tag{14a}$$

$$S_1 = \int_{u^E}^{u^F} \frac{du}{\sqrt{-2\int_0^u f(v)dv + \beta_1}}, \tag{14b}$$

$$W_2\sqrt{-2\int_0^{u^D} f(v)dv + \beta_2} + u^D = u^E, \tag{14c}$$



(a)



(b)

FIG. 4.1. (a) *Propagation failure through a three-gap domain with* $W_1 = W_2 = W_3 = 5.0 < W_c$. *The spacings between the three gaps are* $S_1 = 0.66$ *and* $S_2 = 0.69$*;* (b) *successful propagation with* $W_1 = W_2 = W_3$ *for* $S_1 = 0.66$ *and* $S_2 = 0.72$. *Here* u *is plotted at equal time intervals from time* $t = 35$ *with a time lapse between successive curves of* 35 *time units.*

$$\beta_2 - \beta_1 = 2\left\{\int_0^{u^D} f(v)dv - \int_0^{u^E} f(v)dv\right\}, \tag{14d}$$

$$\beta_1 = 2\left\{\int_0^{u^F} f(v)dv - \int_{u^G}^1 f(v)dv\right\}, \tag{14e}$$

where β_1 is the constant for the phase orbit ϕ_1. Hence, we have a system of five equations with five unknowns, u^G, u^F, u^E, u^D, and β_1, which can be solved numerically to

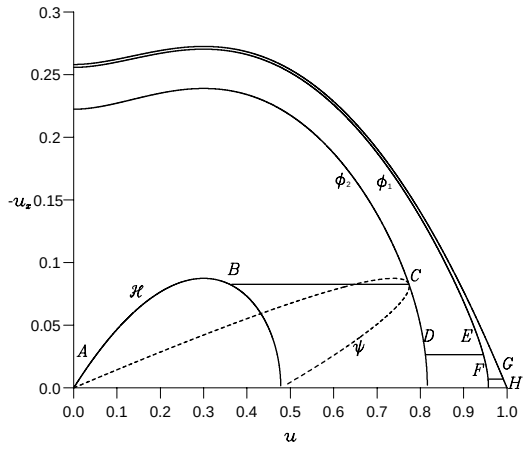


FIG. 4.2. *Phase-portrait of $u_{xx} + f(u) = 0$ for the three-gap steady state in Figure* 4.1(a). *The phase orbits $\phi_{1,2}$ determine the smallest critical size of the second spacing.*

TABLE 4.1
Critical second spacing as a function of the first spacing S_1 for a domain with three gaps $W_1 = W_2 = W_3 = 5.0$ for $\alpha = 0.3$.

S_1	Equation (14)
0.06	0.81
0.21	0.77
0.36	0.74
0.51	0.71
0.66	0.70

determine the points G, F, E, D and the phase orbit ϕ_1. The second critical spacing, $S_{2,c}$, can then be obtained explicitly from

$$S_{2,c} = \int_{u^C}^{u^D} \frac{du}{\sqrt{-2\int_0^u f(v)dv + \beta_2}}. \tag{15}$$

Equations (14a), (14c)–(14e), and (15) form a system of five equations for the five unknowns, u^G, u^F, u^E, u^D, and β_1. Alternatively, for a given S_2, the first critical spacing, $S_{1,c}$, can be obtained from (14b).

Table 4.1 gives the critical length of the second spacing obtained from (15). All gap lengths were taken as equal to 5. Notice the weak dependence of $S_{2,c}$ on S_1. For $S_2 \to \infty$ and $S_1 < S_c$ we anticipate wave-block, as in this case the third gap has no influence on the first two gaps and the system behaves essentially as a two-gap system with the resulting steady state being exactly the same as the steady state associated with a two-gap domain. This then means that there exists a *second* critical length, $S_{2,c}^{'}$, above 0.71 and such that the wave fails to propagate for $S_2 < S_{2,c}$, $S_2 > S_{2,c}^{'}$ and propagates successfully for $S_{2,c} < S_2 < S_{2,c}^{'}$. The analytical construction of $S_{2,c}^{'}$ follows a procedure similar to that for the two-gap system in the previous section. From the initial-value problem, we find $S_{2,c}^{'} \simeq 9.0$—see Figure 4.3 for a numerical experiment with $S_1 = 0.66$ and $S_2 = 9.0$. Hence, like the one-spacing case analyzed

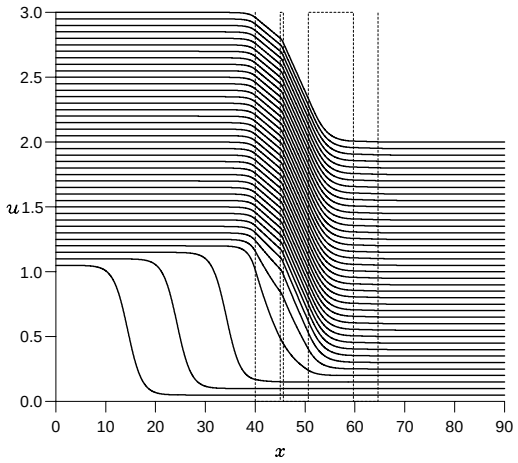


FIG. 4.3. *Wave propagation in a three-gap domain with* $W_1 = W_2 = W_3 = 5.0$ *and* $S_1 = 0.66$, $S_2 = 9.0$, *leading eventually to wave-block. Contrast with Figure* 3.1(b), *where wave propagation is successful. Here* u *is plotted at equal times from time* $t = 35$, *with a time lapse between successive curves of* 35 *time units.*

in section 3, successful propagation is observed for values of S_2 between the two critical values.

It is now useful to summarize the critical conditions for propagation across two or three gaps. From these conditions, a series of "rules" can be established to allow us to predict whether a wave will pass through an array of two or three gaps:

(i) Two-gap system.

1. $W_1 > W_c \simeq 6.5$ and $W_2 < W_c$. There exist two critical spacings S_c, $S_c^{'}$, with $S_c < S_c^{'}$. We have three cases:
   – for $S < S_c$, the wave is blocked.
   – for $S_c < S < S_c^{'}$, wave propagation is successful.
   – for $S > S_c^{'}$, the wave is blocked.
2. $W_1 < W_c$ and $W_2 < W_c$. There exists only one critical spacing, S_c. Here we have two cases:
   – for $S < S_c$, the wave is blocked.
   – for $S > S_c$, wave propagation is successful.
3. $W_2 > W_c$. The wave is blocked independently of W_1.

(ii) Three-gap system. The situation here is much more complicated, and the development of critical conditions for successful wave propagation and failure requires a case-by-case analysis. Here, we focus on the case $W_1 = W_2 = W_3 = 5.0$. S_c is the critical spacing for a two-gap system with $W_1 = W_2 = 5.0$. $S_{1,2}$ are the spacings between the first two and last two gaps, respectively.

1. $S_1 < S_c$. There exist two critical spacings for S_2, $S_{2,c}$, and $S_{2,c}^{'}$, with $S_{2,c} < S_{2,c}^{'}$. Here we have three cases:
   – for $S_2 < S_{2,c}$, the wave is blocked.
   – for $S_{2,c} < S_2 < S_{2,c}^{'}$, wave propagation is successful.
   – for $S_2 > S_{2,c}^{'}$, the wave is blocked.
2. $S_1 > S_c$. There exists only one critical spacing for S_2, $S_{2,c}$. We now have two

cases:
– for $S_2 < S_{2,c}$, the wave is blocked.
– for $S_2 > S_{2,c}$, wave propagation is successful.

5. Propagation with several gaps. The procedure outlined in the previous sections can be generalized to an N-gap domain. In this case we have a system of $3N-3$ equations for the $3N-3$ unknowns, $\beta_i\,(i=1,2,\dots,N-2)$, $u_i\,(i=1,2,\dots,2N-2)$, and one of $S_1, S_2, \dots, S_{N-1}$ (for $W_i\,(i=1,2,\dots,N)$ fixed):

$$u^i = u^{i+1} + W_{\frac{i+1}{2}}\sqrt{-2\int_0^{u^i} f(v)dv + \beta_{i-2}}, \qquad i=3,5,\dots,2N-1,$$

$$\beta_{i+1}-\beta_i = 2\left\{\int_0^{u^{2i+2}} f(v)dv - \int_0^{u^{2i+1}} f(v)dv\right\}, \qquad i=2,3,\dots,N-2,$$

$$S_i = \int_{u^{2i+2}}^{u^{2i+1}} \frac{du}{\sqrt{-2\int_0^u f(v)dv + \beta_i}}, \qquad i=1,2,\dots,N-1,$$

where $u^1 = u^2 + W_1\sqrt{2\int_{u^1}^1 f(v)dv}$ and $\beta_1 = 2\{\int_0^{u^2} f(v)dv - \int_{u^1}^1 f(v)dv\}$.

We must therefore fix $N-2$ of the spacings S_i or, alternatively, fix $2N-2$ spacings and gaps from the total number of $2N-1$ spacings and gaps. Obviously, the spacings have to be chosen such that $-u_x$ always decays as x increases, and hence $u_{x_i}(i=2N-1,2N-2,\dots,1)$ should be a decreasing sequence of $x_i\,(i=2N-1,2N-2,\dots,1)$. In other words, the phase plane of an N-gap system should form a "staircase" from the point where the mapped curve intersects the phase orbit ϕ_{N-1} until the point x_1. We notice that the point u^{2N-1} is always determined from the transfer map of the homoclinic orbit through the *final* gap W_N, and hence only points $u^i\,(i=2N-2,2N-3,\dots,1)$ need to be determined.

The existence of two critical values for S in a two-gap domain, and two critical values for S_2 in a three-gap domain, renders the development of rules for prediction of wave propagation and failure in an N-gap system almost a prohibitive process. Indeed, even for the relatively simple case of four gaps and three spacings, we anticipate, by analogy with the three-gap system, that there will be two critical values for S_1, S_2, and S_3. (Clearly, for a sufficiently large S_2 the four-gap system should approach the steady state of the system $\{W_1, S_1, W_2\}$ in Figure 3.1(a).) We have to emphasize here that we always require the final gap $W_N < W_{cr}$. (W_{cr} is the critical gap length for a medium with a single gap.) At the same time as we increase the number of gaps, the critical values of the spacings between the gaps should decrease: the number of phase orbits, which represent the various invariant manifolds of the steady state, increases, and the "staircase" that links all these orbits gets closer to a continuous curve. (Notice, for example, that the value of S_c in (7) decreases as the distance between u^C and u^D decreases.) Therefore, a large number of spacings/gaps can actually promote wavefront propagation across purely diffusive regions. Figure 5.1 shows successful propagation in an extended domain with a random distribution of gaps occupying 80% of the domain [100, 200]. Notice that the wavefront propagates successfully and without significant slowing through the spacings of length 0.51 and 0.60, which are smaller than 0.68, the smallest critical spacing between a two-gap system with $W_1 = W_2 = 5.0$ considered in section 3.

The question of successful propagation/failure for a large number of gaps (larger than the three considered in section 4) can be addressed by generating a large number

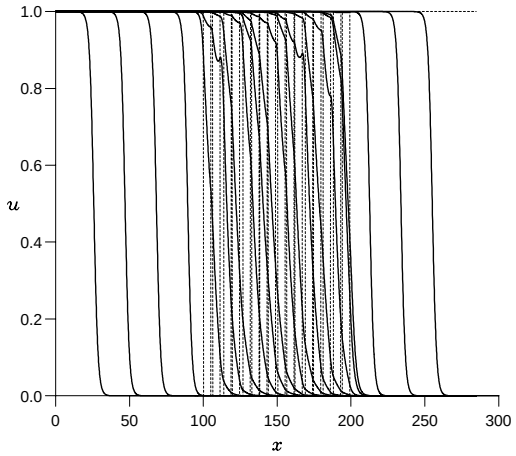


FIG. 5.1. *Successful wave propagation in an extended domain with a random distribution of gaps occupying* 80% *of the domain* $[100, 200]$. *All gaps are taken to be equal to* 5. *The sequence of spacings is the following:* 1.33, 2.58, 0.72, 2.40, 0.99, 0.51, 0.72, 1.48, 1.11, 0.60, 2.18, 0.69, 1.57, 1.94, *and* 1.18. *The time lapse between successive curves is* 75 *time units.*

of random distributions of gaps with a given "void fraction" (fraction of the domain occupied by the gaps). We then have to numerically determine from the initial-value problem the probability (percentage "pass rate") that a wave will successfully propagate along the whole length of the domain. For this purpose, we performed numerical simulations with 20 different random distributions of gaps with length 5 occupying 80% of the domain [100, 200] and a total of 15 spacings in this domain. In all cases, the wave propagated successfully through the medium. However, there might be a particular combination of spacings/gaps which blocks the wavefront. Although an accurate estimate of the probability for successful propagation might require a number of distributions much larger than 20 (which would be computationally a very intensive process), the fact that 20 different random distributions lead to successful propagation indicates a high probability for success.

The above observations depend on the value of the excitability threshold α. In general, as α decreases, the homoclinic orbit that connects to $(0, 0)$ shrinks and, although there is more available space in the phase plane for the phase orbits associated with the steady state, the critical spacings decrease as the intersection between the mapped curve and the phase orbit ϕ_{N-1} gets closer to the u-axis. (At the same time, a more excitable system should require smaller critical spacings for successful propagation.) However, the critical gaps for successful propagation should increase as the distance between the homoclinic orbit and the stable manifold increases.

In addition, we also computed the average speed of the wavefront, defined as the speed of the point at which $u = 1/2$ and based on the time it takes for the front to travel from the first to the last gap. A very interesting conclusion is that, for the system in Figure 5.1 with 16 gaps in the domain [100, 200], the front travels faster than in the cases of 1, 2, or 3 gaps with a void fraction 80% in the same domain. This result along with the main conclusions of the previous sections should have significant implications for a number of areas where the bistable equation in (1) is used as a model system to study wave propagation in heterogeneous media. For example, as Poptsova

and Guria [14] point out, in the context of biophysics, malfunction of the propagation of an impulse in cardiac tissue can cause different cardiac diseases such as ischemia and arrhythmia, which might be explained by the presence of local nonexcitable areas in the cardiac fibre.

6. Multiplicity and stability of solutions. Lewis and Keener [8] proved that, for the one-gap case, if the gap length is larger than the critical length, there exist two steady-state solutions arising via a limit point bifurcation. In this study, we take the two-gap case as an example and show that, if the spacing length is less than (for $W_{1,2} < W_c$ fixed) or the first gap length is larger than (for $W_2 < W_c$ and S fixed) the corresponding critical length, there will also be two steady-state solutions arising via a limit point bifurcation (two such steady states, $AB'C'D'E'F$ and $AB''C''D'E'F$, are depicted in Figure 3.2). From the phase plane analysis in the two-gap case, we have shown that either the spacing length or the gap length is related to the phase orbit constant β, and so we take β as our bifurcation parameter. A bifurcation diagram for this system is shown in Figure 6.1, with $\beta^* = 0.04949$ and $u^*(x_4) = 0.3464$. To acquire further information about the steady solutions, we consider the bifurcation structure of the system $u_{xx} + f(u) = 0$ in more detail.

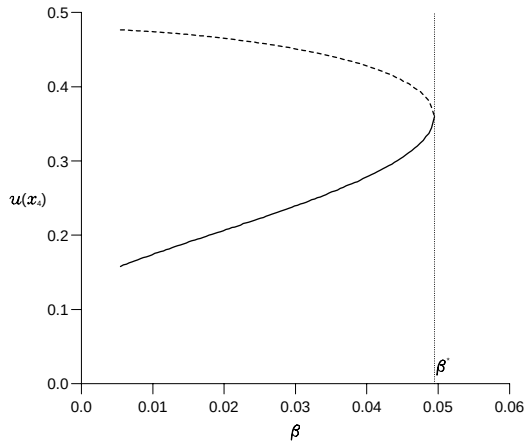


FIG. 6.1. *Bifurcation diagram for $u(x_4)$ of the steady-state solution vs. phase orbit constant β in the two-gap case. The solid line represents stable solutions, and the dashed line represents unstable solutions. The critical β is $\beta^* = 0.04949$. The parameters are $\alpha = 0.3$ and $W_2 = 5.0$.*

Let $v = u_x$ and $v = G_{\mathcal{H}}(u) = \sqrt{-2\int_0^u f(u)du}$ describe the homoclinic orbit $\mathcal{H}$ for $v > 0$ on $u \in (0, u_{\mathcal{H}})$. The function $G_{\mathcal{H}}(u)$ is strictly decreasing for $u \in (\alpha, \overline{u}_{\mathcal{H}})$, with $\overline{u}_{\mathcal{H}}$ the maximum value of $\bar{u}$ on the homoclinic orbit, and thus can be inverted on this region, $u = G_{\mathcal{H}}^{-1}(v) \equiv U_{\mathcal{H}}(v)$. Similarly, $v = G_C(u) = \sqrt{-2\int_0^u f(v)dv + \beta}$ describes the phase orbit through points C' and C'' of Figure 3.2, and therefore we can use $u = G_C^{-1}(v;\beta) = U_C(v;\beta)$ to describe the decreasing portion of the phase orbit with constant β. We can now express the map ψ_{W_2} as

$$\psi_{W_2} : [U_{\mathcal{H}}(v), v] \to [vW_2 + U_{\mathcal{H}}(v), v]$$

and look for a solution to

$$vW_2 + U_{\mathcal{H}}(v) = U_C(v;\beta),$$

where $v \equiv u_x(x_4)$ is a solution of the steady-state equation. Rearranging this equation, we obtain

$$F(v;\beta) = vW_2 + U_{\mathcal{H}}(v) - U_C(v;\beta) = 0.$$

Assume now that we know a solution of the steady-state equation for $\beta = \beta_0$ with $u_x(x_4;\beta_0) = v_0$; then

$$F(v_0;\beta_0) = v_0W_2 + U_{\mathcal{H}}(v_0) - U_C(v_0;\beta_0) = 0. \tag{16a}$$

By expanding F in a Taylor series about (v_0,β_0), we get

$$\begin{aligned} F(v;\beta) \sim\ & [W_2 + U^v_{\mathcal{H}}(v_0) - U^v_C(v_0;\beta_0)](v-v_0) - U^\beta_C(v_0;\beta_0)(\beta-\beta_0) \\ & + \frac{1}{2}[U^{vv}_{\mathcal{H}}(v_0) - U^{vv}_C(v_0;\beta_0)](v-v_0)^2 - \frac{1}{2}U^{\beta\beta}_C(v_0;\beta_0)(\beta-\beta_0)^2 \\ & - U^{v\beta}_C(v-v_0)(\beta-\beta_0) + \{\text{h.o.t.}\}, \end{aligned} \tag{16b}$$

where

$$U^v_{\mathcal{H}}(v_0) = \frac{dU_{\mathcal{H}}}{dv}(v_0), \quad U^v_C(v_0,\beta_0) = \frac{\partial U_C}{\partial v}(v_0,\beta_0), \quad U^\beta_C(v_0,\beta_0) = \frac{\partial U_C}{\partial \beta}(v_0,\beta_0),$$

$$U^{vv}_{\mathcal{H}}(v_0) = \frac{d^2U_{\mathcal{H}}}{dv^2}(v_0), \quad U^{vv}_C(v_0,\beta_0) = \frac{\partial^2 U_C}{\partial v^2}(v_0,\beta_0), \quad U^{\beta\beta}_C(v_0,\beta_0) = \frac{\partial^2 U_C}{\partial \beta^2}(v_0,\beta_0),$$

$$U^{v\beta}_C(v_0,\beta_0) = \frac{\partial^2 U_C}{\partial v \partial \beta}(v_0,\beta_0).$$

Thus, bifurcations can occur when

$$W_2 + U^v_{\mathcal{H}}(v_0) - U^v_C(v_0;\beta_0) = 0. \tag{16c}$$

Note that $U^v_{\mathcal{H}}(v_0) < 0$ is necessary for obtaining solutions to this equation, because W_2 and $-U^v_C(v_0;\beta_0)$ are both > 0. This is always the case for the branch of the homoclinic orbit that we have chosen to work with. However, this is not the case on the $u \in (0,\alpha)$ portion of the homoclinic orbit $\mathcal{H}$, which also means that all bifurcations occur when $u > \alpha$. By solving the two equations (16a) and (16c), we can get v_0 and β_0 as a function of v_0.

We now let

$$\beta \sim \beta_0 + \epsilon\beta_1 + \epsilon^2\beta_2 + \cdots, \qquad v \sim v_0 + \epsilon v_1 + \epsilon^2 v_2 + \cdots,$$

where $0 < \epsilon \ll 1$. When these series are substituted into (16b), one gets the $O(\epsilon)$ term

$$-U^\beta_C(v_0;\beta_0)\beta_1 = 0,$$

which implies $\beta_1 = 0$. At $O(\epsilon^2)$ we find

$$-U^\beta_C(v_0;\beta_0)\beta_2 + \frac{1}{2}[U^{vv}_{\mathcal{H}}(v_0) - U^{vv}_C(v_0;\beta_0)]v_1^2 = 0,$$

and hence β_2 is arbitrary. Choosing $\beta_2 = -1$ yields

$$v_1 = \pm\sqrt{\frac{-2U_C^{\beta}(v_0;\beta_0)}{U_{\mathcal{H}}^{vv}(v_0) - U_C^{vv}(v_0;\beta_0)}},$$

which is real if the quantity in the root is > 0, i.e.,

$$-\frac{U_C^{\beta}(v_0;\beta_0)}{U_{\mathcal{H}}^{vv}(v_0) - U_C^{vv}(v_0;\beta_0)} > 0.$$

We can easily confirm that this always the case, and hence we have two solutions which coalesce at $\beta = \beta_0$ and vanish as β increases above β_0. From the phase portrait in Figure 3.2 and from (5d), (6b), and (7), we can see that the two steady-state solutions correspond to different first gap and spacing lengths. That is, if we fix W_2 and S, there are two solutions with different first gaps. Alternatively, if we fix W_1 and W_2, there two solutions with different spacings. Finally, the solution branch with $v < v_0$ immediately following the bifurcation point has

$$W_2 + U_{\mathcal{H}}^{v}(v) - U_C^{v}(v;\beta) > 0,$$

while the solution with $v > v_0$ has

$$W_2 + U_{\mathcal{H}}^{v}(v) - U_C^{v}(v;\beta) < 0.$$

Whenever there are two solutions, the upper solution (as in Figure 6.1) is unstable and the lower is stable. Lewis and Keener [8] proved the stability of the steady states for the one-gap case by constructing time-independent lower and upper solutions of (1). We follow the same method here to study the stability of the two branches in Figure 6.1. Although our discussion is confined to the two-gap case, it can be extended to any number of gaps.

Following Pauwelussen [18], we call χ a lower solution of the equation $Nu \equiv u_t - u_{xx} - g(u,x) = 0$ if χ satisfies $N\chi \leq 0$ on differentiable segments of χ and $\chi_x(x^+,t) \geq \chi_x(x^-,t)$ on any point x at which χ_x has discontinuities. Similarly, ω is an upper solution when it satisfies the same conditions with all the inequality signs reversed. Using upper and lower solutions, the following lemma from Pauwelussen [18], which is an extension of a theorem by Aronson and Weinberger [16] based on the comparison principle, allows us to examine the asymptotic behavior of the solutions as $t \to \infty$: if $\omega(x)$ is an upper solution of (1), then $u(x,t;\omega)$ is a nonincreasing function of t, with $\lim_{t\to\infty} u(x,t;\omega) = q(x)$, where $q(x)$ the largest stationary solution of (1) satisfying the inequality $q(x) \leq \omega(x)$. Similarly, if $\chi(x)$ is a lower solution of (1), then $u(x,t;\chi)$ is a nondecreasing function of t, and $\lim_{t\to\infty} u(x,t;\chi) = \tau(x)$, where $\tau(x)$ is the smallest possible stationary solution of (1) such that $\tau(x) \geq \chi(x)$.

Therefore, a steady state $\bar{u}(x)$ with $\chi < \bar{u} < \omega$, where χ and ω are lower and upper solutions, respectively, and arbitrarily close to $\bar{u}$, is stable since, as $t \to \infty$, $u(x,t;\chi)$ will tend to the smallest steady state that is greater than or equal to χ, and $u(x,t;\omega)$ will tend to the largest steady state that is smaller than or equal to ω. Because χ and ω are arbitrarily close to $\bar{u}$ (by construction), the smallest steady state and the largest steady state is $\bar{u}$. Hence, $\lim_{t\to\infty} u(x,t;\chi) = \lim_{t\to\infty} u(x,t;\omega) = \bar{u}$, which proves the stability of $\bar{u}$. On the other hand, a steady state $\bar{u}(x)$ with $\omega < \bar{u} < \chi$, where χ and ω are lower and upper solutions, respectively, and arbitrarily close to $\bar{u}$, is unstable since $u(x,t;\omega)$ approaches the largest steady state that is smaller than ω. This solution is the smaller of the two steady states and is associated with the lower branch in Figure

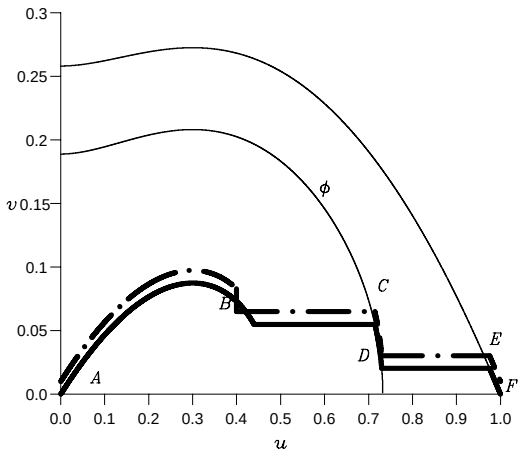


FIG. 6.2. *Steady-state solution (solid line) and time-independent upper solution (dot-dashed line) for a two-gap system.*

6.1, indicating that the smaller solution is stable and the largest solution unstable. At the same time, $u(x,t,\chi)$ will not approach a steady state as there is no steady-state solution larger than χ.

We now construct an upper solution ω, with $\omega < \bar{u}$. From $x = x_D$ to x_C in Figure 6.2, we take ω to closely follow curve χ with $\omega_x(x_C) = \bar{u}_x(x_C) + \epsilon$, with ϵ arbitrarily small. Because curve χ is monotonic with $d\bar{u}/d\bar{u}_x < 0$ in the region of interest, we require that $\epsilon > 0$ such that $\omega < \bar{u}$. For $x_C < x < x_B$, we take ω to be a straight line such that ω and ω_x are continuous at $x = x_C$. The equation for $\omega(x_B)$ is then $\omega(x_B) = -W_2(\bar{u}_x(x_C) + \epsilon) + U_C(\bar{u}_x(x_C) + \epsilon, \beta)$. The remainder of ω is taken to be the homoclinic orbit with ω continuous. Thus, ω is a solution everywhere except $x = x_B$, where ω_x is not continuous for $\epsilon \neq 0$. Finally, for ω to be an upper solution, the jump condition $\omega_x(x_B^+) < \omega_x(x_B^-)$ must be satisfied. We therefore require that $\omega(x_B) < U_{\mathcal{H}}(\bar{u}_x(x_C) + \epsilon)$ for points with $\bar{u} \in (\alpha, \bar{u}_{\mathcal{H}})$, where $U'_{\mathcal{H}} < 0$. This gives the inequality

$$-W_2(\bar{u}_x(x_C) + \epsilon) + U_C(\bar{u}_x(x_C) + \epsilon, \beta) - U_{\mathcal{H}}(\bar{u}_x(x_C) + \epsilon) < 0,$$

which, when expanded in a Taylor series at $\epsilon = 0$, yields

$$\begin{aligned} &-W_2(\bar{u}_x(x_C)) + U_C(\bar{u}_x(x_C), \beta) - U_{\mathcal{H}}(\bar{u}_x(x_C)) \\ &- (W_2 + U'_C(\bar{u}_x(x_C), \beta) - U'_{\mathcal{H}}(\bar{u}_x(x_C)))\epsilon + \cdots < 0. \end{aligned}$$

Because $\bar{u}(x)$ is the steady-state solution, we have that

$$-W_2(\bar{u}_x(x_C)) + U_C(\bar{u}_x(x_C), \beta) - U_{\mathcal{H}}(\bar{u}_x(x_C)) = 0,$$

and since ϵ is positive and arbitrarily small, for ω to exist we require that

$$W_2 + U'_C(\bar{u}_x(x_C), \beta) - U'_{\mathcal{H}}(\bar{u}_x(x_C)) > 0.$$

Similar arguments can be used to construct the lower solutions, as well as lower and upper solutions with $\chi < \bar{u} < \omega$. Notice that in all cases and independently of the number of gaps, the lower and upper solutions require only *one* discontinuity at the downstream edge of the last gap (point B in Figure 6.2 for a two-gap system).

7. Conclusion. We have considered wave propagation in an excitable medium through localized regions of no excitability (the "gaps") by using the bistable equation as a model system. We extended the geometric method of Lewis and Keener for a single gap domain to a hybrid geometric-algebraic method for the case of multiple gaps. The method allowed us to obtain criteria for successful wave propagation and failure ("wave-block") in the bistable equation. For a system with two and three gaps, we found that there are two critical values for the last spacing when all other spacings are fixed: for values of the second spacing that are less than the smaller critical value and greater than the larger critical value, wave-block occurs. Our findings can be readily extended to a larger number of gaps. We also demonstrated that much like the one-gap case, a two-gap system exhibits multiplicity and is such that there are two steady states with different values of the first gap for a fixed second gap and spacing, and two steady states with different spacings for fixed first and second gaps. Finally, in some cases increasing the number of gaps/spacings was found to promote wavefront propagation across purely diffusive regions as the critical gaps for successful propagation increase and the critical spacings decrease.

Acknowledgment. We thank the anonymous referee for useful comments and suggestions.

REFERENCES

[1] A.L. Kay and J.A. Sherratt, *Spatial noise stabilizes periodic wave patterns in oscillatory systems on finite domains*, SIAM J. Appl. Math., 61 (2000), pp. 1013–1041.
[2] A.L. Kay and J.A. Sherratt, *On the persistence of spatiotemporal oscillations generated by invasion*, IMA J. Appl. Math., 63 (1999), pp. 199–216.
[3] J.X. Xin, *Existence and nonexistence of traveling waves and reaction-diffusion front propagation in periodic media*, J. Statist. Phys., 73 (1996), pp. 893–926.
[4] H. Ikeda and M. Mimura, *Wave-blocking phenomena in bistable reaction-diffusion systems*, SIAM J. Appl. Math., 49 (1989), pp. 515–538.
[5] J.P. Keener, *Homogenization and propagation in the bistable equation*, Phys. D, 136 (2000), pp. 1–17.
[6] G. Rousseau and R. Kapral, *Asynchronous algorithm for integration of reaction-diffusion equations for inhomogeneous excitable media*, Chaos, 10 (2000), pp. 812–825.
[7] J. Sneyd and J. Sherratt, *On the propagation of calcium waves in an inhomogeneous medium*, SIAM J. Appl. Math., 57 (1997), pp. 73–94.
[8] T.J. Lewis and J.P. Keener, *Wave-block in excitable media due to regions of depressed excitability*, SIAM J. Appl. Math., 61 (2000), pp. 293–316.
[9] J.X. Xin and J. Zhu, *Quenching and propagation of bistable reaction-diffusion fronts in multidimensional periodic media*, Phys. D, 81 (1995), pp. 94–110.
[10] P. Grindrod and B.D. Sleeman, *A model of a myelinated nerve axon: Threshold behaviour and propagation*, J. Math. Biol., 23 (1985), pp. 119–135.
[11] P. Grindrod, *A model for a myelinated nerve axon*, in Ordinary and Partial Differential Equations, B.D. Sleeman and R.J. Jarvis, eds., Lecture Notes in Math. 1151, Springer-Verlag, New York, 1985, pp. 183–191.
[12] J.P. Keener and J. Sneyd, *Mathematical Physiology*, Springer-Verlag, New York, 1998.
[13] D. Salazar and J. Solà-Morales, *On the number and indices of equilibria in a space-dependent bistable parabolic equation*, Nonlinearity, 14 (2001), pp. 121–132.
[14] M.S. Poptsova and G.T. Guria, *Autowave tunneling through a non-excitable area of active media*, Gen. Physiol. Biophys., 16 (1997), pp. 241–261.
[15] J. Yang, S. Kalliadasis, J.H. Merkin, and S.K. Scott, *Wave propagation in distributed media*, Chaos, 11 (2001), pp. 479–486.
[16] D.G. Aronson and H.F. Weinberger, *Nonlinear diffusion in population genetics, combustion and nerve pulse propagation*, in Partial Differential Equations and Related Topics, Lecture Notes in Math. 446, Springer-Verlag, New York, 1975, pp. 5–49.
[17] P.C. Fife and J.B. McLeod, *The approach of solutions of nonlinear diffusion equations to travelling wave solutions*, Arch. Rat. Mech. Anal., 65 (1977), pp. 335–361.
[18] J.P. Pauwelussen, *Nerve impulse propagation in a branching nerve system*, Phys. D, 4 (1981), pp. 67–88.

SIAM J. APPL. MATH.
Vol. 63, No. 2, pp. 510–537

A THERMOMECHANICAL MODEL FOR ENERGETIC MATERIALS WITH PHASE TRANSFORMATIONS*

GREGORY A. RUDERMAN†, D. SCOTT STEWART‡, AND JACK JAI-ICK YOH‡

Abstract. A model is developed to describe energetic materials with phase transformations from solid to liquid to gas with an exothermic chemical reaction. The model uses a phase variable and a reaction progress variable as thermodynamically independent state variables. A configurational force balance is used to derive an evolution law for the phase variable. The evolution equation for the reaction progress variable is posed as a basic law. In various limits the material is a classical elastic solid, a Newtonian viscous liquid, and a compressible gas. The model is examined in relation to classical equilibrium thermodynamics in a quasi-static limit. The model formulation is specialized to simple motions which are analyzed in a companion paper.

Key words. combustion, phase transformations, energetic materials

AMS subject classifications. 74A50, 74F10, 74F25, 74A15, 80A22, 80A25

PII. S0036139901390258

1. Introduction.

This paper presents a thermodynamically self-consistent model that can describe a material that undergoes phase transitions from solid to liquid to gas with an exothermic chemical reaction. The model development is quite basic and is likely to have wider applications, but the motivation for the study is to describe the behavior and properties of energetic materials such as those used in pyrotechnic materials such as condensed explosives and solid propellants.

Condensed phase energetic materials (EMs) are most typically room temperature organic solids that bind substantial chemical energy in molecular bonds. Upon initiation of chemical reaction between submolecular constituents within the solid, energy is released that is subsequently available to do work or is converted into heat. The advantage of the condensed phase explosive is that the energy per unit volume is approximately a thousand times higher than its premixed, gaseous counterpart.

For the purposes of illustration and to help us develop a conceptual framework, we will consider the energetic material HMX, $[CH_2 - N(N0_2)]_4$, [1] (a solid explosive compound) to be a base-line energetic material. HMX is solid at room temperature and pressure, and when fully chemically decomposed, its gaseous products are simple gases like water vapor, carbon dioxide, and molecular nitrogen. There are thousands of known energetic (explosive) compounds, so our choice of HMX is both practical (because of its wide use) and representative, in that nearly all of the modeling issues considered here apply to similar materials. Fundamental scientific questions surround the phenomena of ignition and release of energy in these materials (EMs) subsequent to impact with a piston or due to a rapid shearing motion. At high impact speeds, (typically on the order of 1000 m/sec), simple hydrodynamic models give an adequate

*Received by the editors June 4, 2001; accepted for publication (in revised form) April 29, 2002; published electronically November 19, 2002. This work was sponsored by the U.S. Air Force Research Laboratory. This work was carried out with resources from the U.S. Air Force Research Laboratory, Armament Directorate, Eglin AFB, Florida, F08630-95-004 and the U.S. Air Force Office of Scientific Research, Physical Mathematics Directorate, F49620-96-1-0260.

http://www.siam.org/journals/siap/63-2/39025.html

†Air Force Research Laboratory, Edwards Air Force Base, Edwards AFB, CA 93524 (Gregory.Ruderman@edwards.af.mil).

‡Department of Theoretical and Applied Mechanics, University of Illinois, Urbana-Champaign, Urbana, IL 61801 (dss@uiuc.edu, yoh1@llnl.gov).

description for both ignition and transition to detonation. Hydrodynamic models are expressed in the form of the Euler equations for reactive gas dynamics [2], which balance kinetic energy, elastic potential energy, and the chemical energy released by the reaction. By virtue of the speed of collision and the short duration of the ignition event, one can justify the neglect of other types of energy and their transfer. However, at lower impact speeds (typically below 1000 m/sec) one must fully take into account the solid nature of the material. In contrast, models for lower-speed impact must reflect a large number of types of energy and mechanisms by which energy in the condensed phases can be transformed, localized, and dissipated. A successful model must be able to describe three-dimensional stress distributions, heat conduction, phase transformations, and chemical reaction as the material changes from solid to liquid to gas.

Thus, accounting for the change in phase and chemical reaction are essential parts of modeling the ignition of energetic solids. In order to do this in a continuum modeling framework, one must add additional thermodynamic state variables that reflect the internal degrees of freedom that measure the extent of reaction and phase change in the material. Necessarily, one must posit additional balance laws and provide the required constitutive theory to complete a model formulation. One does this by using physical considerations (that may lie in the proposed model's subscale physics) to pose the required additional balance laws. For example, in the case of classical combustion theory (see Williams [3] for a representative discussion of the derivation of the commonly used equations of combustion theory), the additional state variables that correspond to the internal degrees of freedom are the mass fractions of all the independent chemical species. The additional balance laws are literal statements of molecular mass balance for each independent species. Other constitutive forms required to describe the evolution of the mass fraction variable are based on well-known laws of collisional reaction (in the case of gaseous chemical reaction), Fickean diffusion, and so on. Importantly, the added balance laws themselves have an identifiable, *molecular origin* and are directly related to physically unambiguous statements of mass balance. However, while the physics at the molecular subscale is clear, the continuum-scale formulation embraces the added (partial) mass conservation statements as primitive, physical laws that must be given by ansatz.

When modeling the phase changes from solid to liquid to gas it is also important to have a physical understanding of the molecular origins of state variables and constitutive forms that describe the phase change. On the molecular scale, a typical EM solid like HMX is comprised of nitrated hydrocarbon molecules that reside in a highly ordered crystal lattice. Large quantities of energy are released only if there is a chemical reaction between smaller pieces of the molecule, juxtaposed or dislodged by deformation, which subsequently release their chemical energy through elementary exothermic reactions typical of those for the gas-phase chemistry. For example, the liquid phase of HMX is known to be very reactive and short-lived compared to the solid phase; likewise HMX vapor is extremely reactive [4], [5]. The liquid phase is molecularly less well ordered than the solid, with larger average intermolecular distances than the solid. If correlated to the average intermolecular spacing (say), the gas phase is less ordered than the liquid. Thus a state variable (sometimes called an order parameter or a phase field variable) can be introduced to reflect a continuum measure of molecular order of the condensed phases (solid crystalline and liquid phases) and the gaseous phase. We will call the order parameter, or phase field variable, simply the phase variable ϕ and assume that it is normalized in such a way so that $\phi = 0$

corresponds to a solid, $\phi = 1$ a liquid, and $\phi = 2$ a gas.

In this formulation, the precise relationship of noninteger values of a phase variable like ϕ to the molecular subscale structure of the material is somewhat ambiguous in contrast to the unambiguous meaning of reactant mass fractions in combustion theory. In a more advanced theory it is anticipated that ϕ will be assigned to specific molecular coordinates. Advances in molecular dynamics of condensed phase systems do promise to eventually provide a more substantial basis for physical assignment of the phase variables, possibly based on the average molecular spacing (say) or other molecularly based kinematic variables [6], [7].

Despite possible ambiguity in its precise physical interpretation, if a phase variable is to be used in a model to represent an independent degree of freedom, it should be constrained by standard principles found in the theory of continuum mechanics. In the regions where the phase is pure (i.e., $\phi = 0$, 1, or 2) the material properties and the constitutive relations must describe the pure material with the properties of that phase. We require that the formulation has a sense in which it is thermodynamically and tensorially consistent. This allows further developments in a rational and systematic manner in three dimensions. We consider a simplified model of an EM (HMX, say) which we suppose has three relevant phases: a solid phase, a liquid phase, and a gas phase. We assume that the path from solid to gas goes through the successive phase transformations, solid $\rightarrow$ liquid $\rightarrow$ gas. Phase boundaries are to be represented by (typically thin) regions across which the value of the phase variable changes from one constant to another. Also, we will use a single (lumped chemistry) progress variable λ, to describe the extent of exothermic chemical reaction λ with value $\lambda = 0$ when no reaction has occurred and $\lambda = 1$ when the reaction is completed. The model allows chemical reactions in any phase.

A key aspect of the model is explicit partitioning of the energy associated with specific internal (thermal) energy, chemical reaction energy, elastic potential (deformational) energy, and energies associated with phase change, such as the enthalpies associated with melting of the solid and evaporation of the liquid, and potential energies stored at phase boundaries. The partitioning of the energy is represented by a decomposition of the Helmholtz free energy ψ into the various parts associated with the energies listed above, such that $\psi = \psi_{thermal} + \psi_{elastic} + \psi_{reaction} + \psi_{phase} + \psi_{grad(phase)}$. The constitutive forms used for $\psi_{thermal}$ and $\psi_{elastic}$ are found in discussions of thermo-elastic materials. The constitutive forms for ψ_{phase} and $\psi_{grad(phase)}$ contain the energies of phase change and energies stored near phase change interfaces. The constitutive form for $\psi_{reaction}$ can be found in a discussion of premixed combustible materials. The free energies and other constitutive variables are allowed to depend on both the phase variable ϕ and the reaction progress variable λ as well as the temperature T and the deformation gradient $\boldsymbol{F}$ and the gradient of ϕ, $\vec{\nabla}\phi$. The governing equations are formulated by statements of conservation of mass, momentum, energy, evolution equations for the change in phase, and the progress of the chemical reaction.

The treatment we use to describe the evolution of the phase variable follows classical treatments that arose in the discussion of solidification (for example, see [8]) but specifically follows a consistent formulation pioneered by Gurtin [9]. Gurtin has argued for a separate continuum balance of configurational forces acting near the boundaries separating pure phases in the volumetric bulk. The arguments for including these additional forces may be justified by consideration of short-range van der Waals forces that typically are generated near phase boundaries due to local changes in the intermolecular distances. The arguments for such configurational forces

are similar to those used to explain classical surface tension forces. The hypothesis is that if the configurational forces act in the vicinity of the boundary near the change in phase and in the bulk, they can be in balance, and if so, they must not effect the overall (conventional) momentum balances. Hence the force balance is posited as a basic law. However, with the postulate of a balance of configurational forces comes the consequence that those forces do work. The working rate is accounted for explicitly in the overall energy balance.

The second law of thermodynamics (the Clausius–Duhem inequality) restricts the form of the constitutive theory so that the rate processes are dissipative and entropy is increasing. An important outcome of these arguments is the derivation of an evolution equation for the phase variable ϕ that is essentially a Ginzburg–Landau equation with additional forcing terms. The evolution equation for ϕ is a time-dependent, reaction-diffusion equation which is amply capable of describing the pattern formation associated with phase transformation. The richness of the resulting theory becomes evident in the energy equation. Due to the decomposition of the Helmholtz free energy, the energy equation contains contributions from all the different terms in the partition and reflects the fact that in the energetic material, energy is converted and distributed to many different forms such as elastic, kinetic, internal, and phase gradient energy (stored in interfaces).

In the sections that follow, the development of the model is given, based on the continuum-thermodynamic formulation described above. In section 2 we review the continuum-thermodynamic formulation consistent with conventional combustion theory [3], [10], [11], [12] that specifically includes a reaction progress variable. A (nonstandard) presentation of the Helmholtz energy decomposition is given and the attendant standard arguments for restrictions placed by the second law are given. In section 3 we present a model for a material that changes from solid to liquid to gas and present a Helmholtz free-energy decomposition that is suitable to describe such a material, subject to second law restrictions. In section 4, the combined model for an EM (with both phase change and chemical reaction) is then presented. In section 5 we discuss various limiting cases of the model. We discuss the relationship of the model to classical quasi-static thermodynamics and illustrate examples based on fits to HMX properties to illustrate the dynamics of a phase change that would be calculated in the classical theory. Section 6 presents special formulations of the model equation for three important simple motions. These cases are (i) constant volume evolution (which is a generalization of the classical constant volume explosion formulation found in combustion theory), (ii) one-dimensional, time-dependent longitudinal compression (expansion), and (iii) one-dimensional, time-dependent shear motion. The solution of the equations for these three important cases for an HMX-like material is the subject of the companion paper [13].

In what follows, a "c" subscript denotes a condensed phase, either solid or liquid, an "f" subscript denotes fluid, either liquid or gas, an "s" subscript denotes solid, an "l" subscript denotes liquid, and a "g" subscript denotes gas. The spelled out subscripts "solid," "liquid," and "gas" refer to constant values for that pure phase. The notation is kept as simple as possible in an attempt make the paper easier to read. Bold face quantities can either be vectors or tensors. If obvious, the constant arguments during differentiation are dropped. Our notation is standard, insofar as is possible and follows a well-known text like Bowen [12].

1.1. Kinematics. Let the Eulerian (spatial) coordinates of position in the lab-frame be given by $\boldsymbol{x}$ and the Lagrangian (material) coordinates (or particle coordi-

nates) be given by $\boldsymbol{X}$. For simplicity we will assume that $\boldsymbol{X}$ represent the initial position of material particles. Then the mapping of the deformations that define the particle trajectory paths is given by

$$\boldsymbol{x} = \boldsymbol{x}(\boldsymbol{X}, t)\,. \tag{1.1}$$

The deformation gradient $\boldsymbol{F}$ is defined by the derivative

$$\boldsymbol{F} = \frac{\partial \boldsymbol{x}}{\partial \boldsymbol{X}}\,, \tag{1.2}$$

and the velocity of particles $\boldsymbol{v}$ is defined by the time derivative of the particle trajectories $\boldsymbol{v} = (\partial \boldsymbol{x}/\partial t)_{\boldsymbol{X}}$. The velocity gradient is $\boldsymbol{L} = \vec{\nabla}\boldsymbol{v}$. Let the dot notation, $\dot{()}$, refer to the material derivative. A standard identity that can be verified by the previous definitions and the chain rule gives the material (particle-fixed) time derivative of the deformation gradient as $\dot{\boldsymbol{F}} = \boldsymbol{L}\boldsymbol{F}$. A statement of conservation of mass in the material frame is that the ratio of the instantaneous density, ρ, of the particle to a reference (ambient) density of the solid, ρ_0, is equal to the determinant of the deformation gradient

$$\det(\boldsymbol{F}) = \frac{\rho_0}{\rho}\,. \tag{1.3}$$

2. Review of the thermomechanics for a simple model of a reactive flow. The standard combustion model, for a premixed mixture that can explode or burn, can be derived from a simple mixture theory; see references [3], [10], [11], [12]. The combined model that we introduce later incorporates the features of the standard combustion model, so we review its derivation. Importantly, the reaction progress variable λ represents a product mass flux. Hence λ is treated differently from the phase variable ϕ, which is introduced later to describe the change in phase from solid to liquid to gas.

For the purpose of discussion, one assumes that there are only two distinct species, fuel and product (say). The corresponding chemical reaction is written as $F \rightarrow P + Q_{hc}\,(heat)$. All physical properties of the two species such as the molecular weights, specific heats, conductivities, etc. are assumed to be identical, save the heats of formation, the weighted difference of which is the heat of combustion.

We start with the balance laws for conservation of mass, linear momentum (without body forces), and energy:

$$\dot{\rho} + \rho\,(\vec{\nabla}\cdot\boldsymbol{v}) = 0\,, \tag{2.1}$$

$$\rho\,\dot{\boldsymbol{v}} = \vec{\nabla}\cdot\boldsymbol{\sigma} + \rho\boldsymbol{f}\,, \tag{2.2}$$

$$\rho\dot{e} = \boldsymbol{\sigma} : \vec{\nabla}\boldsymbol{v} - \vec{\nabla}\cdot\boldsymbol{q} + \rho\, r\,. \tag{2.3}$$

In the energy equation, r is a volumetric energy production term that typically represents radiation or volumetric heating (or cooling) in combustion theory. The body force is given by $\boldsymbol{f}$. In addition, we invoke a primitive evolution law for the reaction progress variable λ:

$$\rho\dot{\lambda} = \vec{\nabla}\cdot\boldsymbol{s} + \rho\,\Omega\,. \tag{2.4}$$

The vector $\boldsymbol{s}$ is the flux of mass of reacted species per unit area per unit time and $\rho\,\Omega$ is the instantaneous rate of creation of mass of the reacted species per unit volume.

Then λ is recognized as the mass fraction of the product species. Further, by direct correspondence with the standard combustion equations, one can interpret $\boldsymbol{s} = \rho\lambda\boldsymbol{V}$, where $\boldsymbol{V}$ is the diffusion velocity of the product species (say), and where $\rho\lambda$ is the partial density fraction of the same product.

To these basic laws we add the second law of thermodynamics, the Clausius–Duhem inequality

$$\rho\,\dot{\eta} \geq -\vec{\nabla}\cdot\left(\frac{\boldsymbol{q}}{T}\right) + \vec{\nabla}\cdot\left(\frac{Q_{hc}\boldsymbol{s}}{T}\right) + \frac{\rho r}{T}\,, \tag{2.5}$$

where Q_{hc}, the heat of combustion, is the exothermic energy release per unit mass and the term $\vec{\nabla}\cdot(Q_{hc}\,\boldsymbol{s}/T)$ represents the gradient of the entropy flux associated with the chemical reaction.

2.1. Constitutive forms and restrictions. Next consider the classical forms and assumptions that lead to the combustion equations of premixed materials found in texts like [3] or [10]. The formulation uses the Helmholtz free energy, which is defined in terms of the internal energy and entropy as $\psi = e - T\eta$. We start with the assumption that ψ is specified by

$$\psi = \psi(\boldsymbol{F}, T, \lambda), \tag{2.6}$$

and we assume similar dependencies for e, η, and all other thermodynamic variables. Next we consider the implication of the entropy inequality and deduce various restriction imposed by it on the constitutive formulation.

If we use the definition of the Helmholtz free energy to get an expression for the entropy, as $\eta = (e - \psi)/T$, and take the material derivative, we obtain $\dot{\eta} = (\dot{e} - \dot{\psi} - \eta\dot{T})/T$. In particular the derivative $\dot{\psi}$ appears and, using the form assumed above, it is calculated as

$$\dot{\psi} = \frac{\partial\psi}{\partial\boldsymbol{F}}\boldsymbol{F}^T : \vec{\nabla}\boldsymbol{v} + \frac{\partial\psi}{\partial T}\dot{T} + \frac{\partial\psi}{\partial\lambda}\dot{\lambda}\,. \tag{2.7}$$

Using this expression for $\dot{\psi}$ and using the energy equation to replace $\dot{e}$ in the entropy inequality leads to an intermediate result:

$$\left(\boldsymbol{\sigma} - \rho\frac{\partial\psi}{\partial\boldsymbol{F}}\boldsymbol{F}^T\right) : \vec{\nabla}\boldsymbol{v} - \rho\left(\eta + \frac{\partial\psi}{\partial T}\right)\dot{T} - (\boldsymbol{q} - Q_{hc}\boldsymbol{s})\cdot\frac{\vec{\nabla}T}{T} - \rho\frac{\partial\psi}{\partial\lambda}\dot{\lambda} - Q_{hc}\vec{\nabla}\cdot\boldsymbol{s} \geq 0\,. \tag{2.8}$$

We restrict our choice in constitutive theory to forms that will automatically satisfy this dissipation inequality as the physical processes in the material range over all admissible deformations and temperature fields. For example, since $\vec{\nabla}\boldsymbol{v}$ can be regarded as an independent field, then in the standard way we restrict the form of the stress tensor such that

$$\boldsymbol{\sigma} = \rho\frac{\partial\psi}{\partial\boldsymbol{F}}\boldsymbol{F}^T + \boldsymbol{\sigma}^{diss}\,, \tag{2.9}$$

where the dissipative stress $\boldsymbol{\sigma}^{diss}$ satisfies $\boldsymbol{\sigma}^{diss} : \vec{\nabla}\boldsymbol{v} \geq 0$. This last requirement is clearly satisfied by the classical choice for a viscous fluid,

$$\boldsymbol{\sigma}^{diss} = \nu_g(\vec{\nabla}\cdot\boldsymbol{v})\boldsymbol{I} + 2\mu_g\boldsymbol{D}\,, \tag{2.10}$$

where $\boldsymbol{D} = (\vec{\nabla}\boldsymbol{v} + \vec{\nabla}\boldsymbol{v}^T)/2$ and ν_g, μ_g are positive and are identified as the gas-phase bulk and shear viscosities. The assumed form of the stress becomes

$$\boldsymbol{\sigma} = \rho \frac{\partial \psi}{\partial \boldsymbol{F}} \boldsymbol{F}^T + \nu_g (\vec{\nabla} \cdot \boldsymbol{v}) \boldsymbol{I} + 2\mu_g \boldsymbol{D} \,. \tag{2.11}$$

In a similar fashion, since $\dot{T}$ is independent, we require that the Helmholtz free energy must satisfy Gibbs' relation

$$\frac{\partial \psi}{\partial T} = -\eta \,. \tag{2.12}$$

The entropy inequality is now satisfied if the following reduced inequality is satisfied:

$$-(\boldsymbol{q} - Q_{hc}\boldsymbol{s}) \cdot \frac{\vec{\nabla} T}{T} - \rho \frac{\partial \psi}{\partial \lambda} \dot{\lambda} - Q_{hc} \vec{\nabla} \cdot \boldsymbol{s} \geq 0 \,. \tag{2.13}$$

If we assume that the change in the Helmholtz free energy with respect to the progress variable is related to the heat of combustion (which also can be verified and put into direct correspondence with forms derived in mixture theory of reacting gases; see [10], [11], [12]),

$$\frac{\partial \psi}{\partial \lambda} = -Q_{hc} \,, \tag{2.14}$$

and we use the evolution equation for the progress variable $\rho\dot{\lambda} - \vec{\nabla} \cdot \boldsymbol{s} = \rho\Omega$, then the reduced inequality can be recast as

$$-(\boldsymbol{q} - Q_{hc}\boldsymbol{s}) \cdot \frac{\vec{\nabla} T}{T} + \rho \, Q_{hc} \, \Omega \geq 0 \,. \tag{2.15}$$

Finally we make the choice that the energy flux vector is the sum of a Fourier heat conductive flux and an energy flux associated with the diffusion of the product species,

$$\boldsymbol{q} = -k\vec{\nabla} T + Q_{hc} \, \boldsymbol{s} \,, \tag{2.16}$$

and we require that for an exothermic chemical reaction with $Q_{hc} > 0$, the reaction rate must be positive with $\Omega \geq 0$. With these restrictions the second law is automatically satisfied. Recall that $\boldsymbol{s}$ represented the mass flux vector of the product species, $\boldsymbol{s} = \rho\lambda\boldsymbol{V}$, where $\boldsymbol{V}$ is the diffusion velocity of that species. Without further restriction we can make a standard assumption that the diffusion velocity is related to the gradient of the species concentration through a Fick's law relation,

$$\boldsymbol{s} = \rho\lambda\boldsymbol{V} = d \, \vec{\nabla}\lambda \,, \tag{2.17}$$

where $d \geq 0$ is a diffusion coefficient.

2.2. Temperature form of the energy equation. We present the temperature form of the energy equation in terms of a specification of the Helmholtz free energy, in order to set the stage for later discussions. We use the definition of the specific internal energy in terms of the temperature and the entropy, $e = \psi + T\eta$, to obtain $\dot{e} = \dot{\psi} + \eta\dot{T} + T\dot{\eta}$. Next we use the form of the Helmholtz energy $\psi(\boldsymbol{F}, T, \lambda)$ and Gibbs' relation $\eta = -\partial\psi/\partial T$ to generate expressions for $\dot{\psi}$ and $\dot{\eta}$ as

$$\dot{e} = \frac{\partial \psi}{\partial \boldsymbol{F}} \boldsymbol{F}^T : \vec{\nabla}\boldsymbol{v} + \frac{\partial \psi}{\partial T} \dot{T} \,, \quad \dot{\eta} = -\frac{\partial^2 \psi}{\partial T \partial \boldsymbol{F}} \boldsymbol{F}^T : \vec{\nabla}\boldsymbol{v} - \frac{\partial^2 \psi}{\partial T^2} \dot{T} \,. \tag{2.18}$$

We then insert these expression into (2.3) and make some further simplifications. A collection of terms appears that is associated with the stress-related dissipation

$$\left(\boldsymbol{\sigma} - \rho\frac{\partial\psi}{\partial\boldsymbol{F}}\boldsymbol{F}^T\right) : \vec{\nabla}\boldsymbol{v} = \boldsymbol{\sigma}^{diss} : \vec{\nabla}\boldsymbol{v}\,.$$

Using the classical definition of the specific heat at constant deformation (volume),

$$c_v \equiv T\frac{\partial\eta}{\partial T}\bigg|_{\boldsymbol{F}} = -T\left(\frac{\partial^2\psi}{\partial T^2}\right)\bigg|_{\boldsymbol{F}}, \tag{2.19}$$

the energy equation can be rewritten as follows:

$$\rho c_v\dot{T} = -\vec{\nabla}\cdot\boldsymbol{q} + \boldsymbol{\sigma}^{diss} : \vec{\nabla}\boldsymbol{v} + \rho T\frac{\partial^2\psi}{\partial T\partial\boldsymbol{F}}\boldsymbol{F}^T : \vec{\nabla}\boldsymbol{v}\,. \tag{2.20}$$

The term $\rho T(\partial^2\psi/\partial T\partial\boldsymbol{F})\boldsymbol{F}^T : \vec{\nabla}\boldsymbol{v}$ is a stress work term classically associated with thermal stresses. As we will see below in the case of gaseous combustion for ideal gases, this term is proportional to the pressure work term $-p\,(\vec{\nabla}\cdot\boldsymbol{v})$, where $p = \rho R_g T$ and R_g is the ideal gas constant.

2.2.1. The form of the Helmholtz free energy from classical combustion theory. To complete the classical formulation for premixed combustion, one must specify the form of the Helmholtz free energy. The forms can be extracted from the correct forms found in the binary mixture theory of premixed gases; see [12], [3], [10], [11]. Let $\boldsymbol{B} = \boldsymbol{F}\boldsymbol{F}^T$ be the left Cauchy–Green tensor and let $III_{\boldsymbol{B}} = (\rho_0/\rho)^2$ be the third invariant of $\boldsymbol{B}$. Then the form of the Helmholtz free energy for a thermally ideal material, with the additional term required for the change in enthalpy associated with combustion, is comprised of three parts: a thermal energy density $\psi_1 = c_v[(T - T_0) - T\ln(T/T_0)]$, a strain energy density associated with the temperature (that defines the pressure in terms of the density and temperature) $\psi_2 = -1/2\,R_g T\ln(III_{\boldsymbol{B}})$, and the chemical enthalpy $\psi_3 = -Q_{hc}\lambda$. Thus the total free energy $\psi = \psi_1 + \psi_2 + \psi_3$ is

$$\psi = c_v(T - T_0) - c_v T\ln\left(\frac{T}{T_0}\right) - \frac{1}{2}R_g T\,\ln(III_{\boldsymbol{B}}) - Q_{hc}\lambda\,. \tag{2.21}$$

It follows that the elastic part of the stress can be computed from this form of the free energy and identifies the classical thermodynamic pressure p. In particular, we have that $\rho(\partial\psi/\partial\boldsymbol{F})\boldsymbol{F}^T = 2\rho(\partial\psi/\partial\boldsymbol{B})\boldsymbol{B} = -\rho R_g T\boldsymbol{I} \equiv -p\boldsymbol{I}$, which leads to the identification of the pressure p by the ideal gas law, $p = \rho R_g T$. Also, the thermal stress work term is rewritten $\rho\,(\partial\psi/\partial\boldsymbol{F})\boldsymbol{F}^T : \vec{\nabla}\boldsymbol{v} = -p\,(\vec{\nabla}\cdot\boldsymbol{v})$. The corresponding form of the entropy and the internal energy (obtained from the definition of the Helmholtz free energy and Gibbs' relation) are given by

$$e = c_v(T - T_0) - Q_{hc}\lambda\,, \qquad \eta = c_v\ln\left(\frac{T}{T_0}\right) + \frac{R}{2}\,\ln(III_{\boldsymbol{B}})\,. \tag{2.22}$$

2.3. Summary of the governing equations for a premixed reactive fluid. Here we summarize the results of the last section that reduce to the classical form of the combustion equations for a premixed combustible fluid. These equations incorporate the various restrictions and constitutive forms that we assumed and are suitable for solving initial value problems ordinarily associated with the theory of premixed

combustion. The entropy (dissipation) inequality is not included in our list since it is automatically satisfied by construction of the model. The equations for $\rho, \boldsymbol{v}, T$, and λ are

$$\dot{\rho} + \rho\vec{\nabla}\cdot\boldsymbol{v} = 0\,, \tag{2.23}$$

$$\rho\,\dot{\boldsymbol{v}} = \vec{\nabla}\cdot\boldsymbol{\sigma} + \rho\boldsymbol{f}\,, \tag{2.24}$$

$$\rho c_v \dot{T} = \vec{\nabla}\cdot(k\vec{\nabla}T) + \boldsymbol{\sigma}^{diss} : \vec{\nabla}\boldsymbol{v} - p(\vec{\nabla}\cdot\boldsymbol{v}) + \rho\, Q_{hc}\,\Omega\,, \tag{2.25}$$

$$\rho\dot{\lambda} = \vec{\nabla}\cdot(d\,\vec{\nabla}\lambda) + \rho\Omega\,, \tag{2.26}$$

with the constitutive relation for the stress given by $\boldsymbol{\sigma} = -\rho R_g T\boldsymbol{I} + \nu_g\vec{\nabla}\cdot\boldsymbol{v}\boldsymbol{I} + 2\mu_g\boldsymbol{D}$, with $\boldsymbol{\sigma}^{diss} = \nu_g\vec{\nabla}\cdot\boldsymbol{v}\boldsymbol{I} + 2\mu_g\boldsymbol{D}$, and with $\boldsymbol{D} = (\vec{\nabla}\boldsymbol{v} + \vec{\nabla}\boldsymbol{v}^T)/2$.

3. Thermomechanics of a model of a material with phase changes from solid to liquid to gas. Here we develop a model for a material that can undergo a phase change from solid to liquid to gas in preparation for the development of the combined model, which includes chemical reaction and exothermic energy release. The important difference in the development in this section from that in section 2 is the introduction of a phase variable that is used to describe and delineate the separate phases. In order to describe the phase transitions we introduce the (normalized) variable ϕ so that $\phi = 0$ corresponds to the solid phase, $\phi = 1$ to the liquid phase, and $\phi = 2$ to the vapor phase. In its pure phases, solid, liquid, and gas, the material is prescribed by classical models for that pure phase, i.e., a compressible elastic solid and a compressible Newtonian liquid and gas.

A consistent thermodynamic formulation for the model is developed through an extension of a formulation proposed by Gurtin [9]. Energy expended by the system during a phase change is associated with configurational forces of two types—a configurational stress that acts at or near the boundaries between phases which is balanced by a configurational force distributed in the bulk. The power expenditure of these forces must be accounted for in the energy balance. If one assumes that the configurational forces in the material are balanced separately (this is a posited balance), then the evolution of the phase field ϕ is constrained by the entropy inequality to be dissipative and further considerations lead to the derivation of an evolution law for ϕ. This is in contrast to the formulation of the last section, which considered the evolution law for the progress variable λ as posited. Presumably (and we have considered this in some detail that is not presented here), an alternative to deriving the equation for ϕ is to pose an evolution equation as fundamental and then derive the consequence of local balance for the configurational forces. Either way, one comes to similar physical conclusions. The consequences of this choice, in absence of better, physically based arguments, need to be judged against the forms of the equations that result that allow us to solve interesting initial value problems.

The starting point is the form of the general laws. The differential form of the general law for mass, (2.1), and momentum, (2.2), are unchanged from the previous section. We turn to the more unfamiliar considerations of the force balance law associated with the phase change and corresponding changes in the energy balance next.

3.0.1. Force balances associated with the change in phase. Associated with the evolution of the phase variable ϕ, we introduce a balance of configurational stress $\boldsymbol{\xi}$, a configurational internal force density π_ϕ. The integral form of the balance

law for a body in region $\mathcal{B}$ with boundary $\partial\mathcal{B}$ is

$$\int_{\partial\mathcal{B}} \boldsymbol{\xi}\cdot\boldsymbol{n}\, dA + \int_{\mathcal{B}} (\pi_\phi)\, dV = 0, \tag{3.1}$$

and with the use of the divergence theorem, the corresponding differential form of the balance law is

$$\vec{\nabla}\cdot\boldsymbol{\xi} + \pi_\phi = 0\,. \tag{3.2}$$

3.0.2. Rate of work. The rate of work expended on $\mathcal{B}$ is due to the external forces acting on the surface and within the volume of $\mathcal{B}$. Gurtin [9] shows that the correct form for the rate of work due to all stresses is

$$\mathcal{W} \equiv \int_{\partial\mathcal{B}} \big(\boldsymbol{\sigma}\boldsymbol{n}\cdot\boldsymbol{v} + \boldsymbol{\xi}\cdot(\dot{\phi}\boldsymbol{n})\big)\, dA + \int_{\mathcal{B}} \boldsymbol{b}\cdot\boldsymbol{v}\, dV. \tag{3.3}$$

The integral form of the energy balance can be written in the standard way as the material derivative of the total energy (internal and kinetic) equated to the rate of work minus the energy flux out of the body plus the rate of heating by any other source; thus

$$\frac{D}{Dt}\int_{\mathcal{B}} \rho(e + \tfrac{1}{2}|\boldsymbol{v}|^2)\, dV = \mathcal{W} - \int_{\partial\mathcal{B}} \boldsymbol{q}\cdot\boldsymbol{n}\, dA + \int_{\mathcal{B}} \rho r\, dV\,. \tag{3.4}$$

To obtain the differential form we convert the surface integrals into volume integrals and use the divergence theorem. The resulting integral must hold everywhere for all subvolumes, so the resulting integrand is set to zero, which leads to an intermediate differential form (not shown). We then use the momentum equation and take its dot product with the velocity $\boldsymbol{v}$ to get the standard work-energy statement on a material path and subtract that result from the above-mentioned intermediate form to get the following form of the energy equation:

$$\rho\dot{e} = -\vec{\nabla}\cdot\boldsymbol{q} + \boldsymbol{\sigma} : \vec{\nabla}\boldsymbol{v} + \boldsymbol{\xi}\cdot\vec{\nabla}(\dot{\phi}) - \pi_\phi\dot{\phi} + \rho r\,. \tag{3.5}$$

The main difference from the classical form is the appearance of the two work terms $\boldsymbol{\xi}\cdot\vec{\nabla}(\dot{\phi})$ and $-\pi_\phi\dot{\phi}$ that derive from the configurational forces. For upcoming considerations of the entropy inequality, it is useful to use identities (which can be verified easily in Cartesian index notation) to rewrite the term $\boldsymbol{\xi}\cdot\vec{\nabla}(\dot{\phi})$ as

$$\boldsymbol{\xi}\cdot\vec{\nabla}(\dot{\phi}) = \dot{\overline{\vec{\nabla}\phi}}\cdot\boldsymbol{\xi} + \vec{\nabla}\phi\otimes\boldsymbol{\xi} : \boldsymbol{L}\,, \tag{3.6}$$

so that the revised energy equation reads as

$$\rho\dot{e} = -\vec{\nabla}\cdot\boldsymbol{q} + \boldsymbol{\sigma} : \vec{\nabla}\boldsymbol{v} + \dot{\overline{\vec{\nabla}\phi}}\cdot\boldsymbol{\xi} + \vec{\nabla}\phi\otimes\boldsymbol{\xi} : \vec{\nabla}\boldsymbol{v} - \pi_\phi\dot{\phi} + \rho r\,. \tag{3.7}$$

3.0.3. The entropy inequality. Finally, to these basic laws we must add the second law of thermodynamics, the Clausius–Duhem inequality

$$\rho\,\dot{\eta} \geq -\vec{\nabla}\cdot\Big(\frac{\boldsymbol{q}}{T}\Big) + \frac{\rho r}{T}\,. \tag{3.8}$$

Note that since ϕ is not assumed to be related to a partial mass density of material, there is no entropy flux term like $\vec{\nabla}\cdot(Q_{hc}\,\boldsymbol{s}/T)$ that appears in the combustion-based entropy inequality (2.5). Equation (3.8) is the classical (inert) form of the entropy inequality.

3.0.4. Constitutive forms and restrictions from the entropy inequality. We restrict our attention to a general class of constitutive equations and start with a very general assumption that the free energy density ψ, the Cauchy stress $\boldsymbol{\sigma}$, the configurational stresses $\boldsymbol{\xi}$, and the internal configurational force π_ϕ, the entropy density η, and the heat flux $\boldsymbol{q}$ at any point $(\boldsymbol{x}, t)$ are dependent on the deformation gradient $\boldsymbol{F}$, the temperature T, the phase field ϕ, the gradients $\vec{\nabla}T$, $\vec{\nabla}\phi$, and the velocity gradient $\boldsymbol{L}$, such that we can write

$$\psi = \psi(\boldsymbol{F}, T, \phi, \vec{\nabla}T, \vec{\nabla}\phi, \boldsymbol{L}). \tag{3.9}$$

We assume that $\boldsymbol{\sigma}, \boldsymbol{\xi}, \pi_\phi, \eta$, and $\boldsymbol{q}$ all depend on the same argument list, $(\boldsymbol{F}, T, \phi, \vec{\nabla}T, \vec{\nabla}\phi, \boldsymbol{L})$. We use the definition of the Helmholtz free energy to get an expression for the entropy, $\eta = (e - \psi)/T$, take the material derivative, and then use the energy equation to replace $\dot{e}$ and use the chain rule to replace $\dot{\psi}$. These substitutions into the entropy inequality lead to the intermediate result:

$$\left(\boldsymbol{\sigma} - \rho\frac{\partial\psi}{\partial\boldsymbol{F}}\boldsymbol{F}^T + \vec{\nabla}\phi\otimes\boldsymbol{\xi}\right) : \vec{\nabla}\boldsymbol{v} - \rho\left(\eta + \frac{\partial\psi}{\partial T}\right)\dot{T} - \left(\pi_\phi + \rho\frac{\partial\psi}{\partial\phi}\right)\dot{\phi}$$

$$-\rho\,\frac{\partial\psi}{\partial\vec{\nabla}T}\cdot\dot{\vec{\nabla}T} - \left(\rho\frac{\partial\psi}{\partial\vec{\nabla}\phi} - \boldsymbol{\xi}\right)\cdot\dot{\vec{\nabla}\phi} - \rho\left(\frac{\partial\psi}{\partial\boldsymbol{L}}\right) : \dot{\boldsymbol{L}} - \boldsymbol{q}\cdot\frac{\vec{\nabla}T}{T} \geq 0\,. \tag{3.10}$$

Again we restrict our choice of constitutive forms to those that automatically satisfy this dissipation inequality as the physical process in the material ranges over all admissible deformations and temperature and phase fields. We restrict the form of the stress tensor such that

$$\boldsymbol{\sigma} = \rho\frac{\partial\psi}{\partial\boldsymbol{F}}\boldsymbol{F}^T - \vec{\nabla}\phi\otimes\boldsymbol{\xi} + \boldsymbol{\sigma}^{diss}\,, \tag{3.11}$$

where again $\boldsymbol{\sigma}^{diss}$ must be chosen to satisfy $\boldsymbol{\sigma}^{diss} : \vec{\nabla}\boldsymbol{v} \geq 0$. Later we will take $\boldsymbol{\sigma}^{diss}$ to be given by (2.10), where the shear and bulk viscosities are taken to be functions of the phase field variable ϕ. We require that Gibbs' relation be satisfied and that the configurational force $\boldsymbol{\xi}$ be defined by the derivative of the Helmholtz free energy with respect to the gradient of ϕ such that

$$\eta = -\frac{\partial\psi}{\partial T} \quad \text{and} \quad \boldsymbol{\xi} = \rho\frac{\partial\psi}{\partial\vec{\nabla}\phi}\,. \tag{3.12}$$

We also assume that the Helmholtz free energy is independent of $\boldsymbol{L} = \vec{\nabla}\boldsymbol{v}$ and the temperature gradient $\vec{\nabla}T$ so that

$$\frac{\partial\psi}{\partial\boldsymbol{L}} = 0 \quad \text{and} \quad \frac{\partial\psi}{\partial\vec{\nabla}T} = 0 \tag{3.13}$$

hold. We also suppose that the energy flux vector is described by a Fourier heat conduction law, $\boldsymbol{q} = -k\vec{\nabla}T$, and insist that k is a positive constant that can be a function of the temperature and the order parameter, i.e., $k(\phi, T) \geq 0$. Then the reduced dissipation inequality now has the form

$$-\left(\pi_\phi + \rho\frac{\partial\psi}{\partial\phi}\right)\dot{\phi} \geq 0\,. \tag{3.14}$$

The final form of the reduced dissipation inequality is satisfied if we require that the phase changes be dissipative and if we allow π_ϕ to take the form

$$-\left(\pi_\phi + \rho\frac{\partial\psi}{\partial\phi}\right) = B\dot{\phi}, \tag{3.15}$$

where $B \geq 0$. Equation (3.15) is an evolution equation for the phase variable ϕ. Note that the configurational force balance (3.2) defines $\pi_\phi = -\vec{\nabla}\cdot\boldsymbol{\xi}$ and with the configurational force identified by $\boldsymbol{\xi} = \rho(\partial\psi/\partial\vec{\nabla}\phi)$ leads to $\pi_\phi = -\vec{\nabla}\cdot(\rho\partial\psi/\partial\vec{\nabla}\phi)$. Thus (3.15) can be re-expressed as

$$B\dot{\phi} = \vec{\nabla}\cdot\left(\rho\frac{\partial\psi}{\partial\vec{\nabla}\phi}\right) - \rho\frac{\partial\psi}{\partial\phi}\,. \tag{3.16}$$

Given appropriate forms for ψ (such as quadratic dependence of ψ on $\vec{\nabla}\phi$), (3.16) is recognized as an advection, reaction-diffusion equation, which, given an assumed form for ψ, can generate a Ginzburg–Landau equation. The coefficient B^{-1} is then recognized as a kinetic rate constant for the phase transformation.

3.1. Temperature form of the energy equation. In order to show the coupling between the thermal (temperature) field, the stress field, and the phase field, we present an alternative form of the energy equation. Starting with the energy balance (3.7) we use the definition of the specific internal energy in terms of the temperature and the entropy, $e = \psi + T\eta$, to obtain $\dot{e} = \dot{\psi} + \eta\dot{T} + T\dot{\eta}$. Next we use the form of the Helmholtz energy $\psi(\phi, T, \vec{\nabla}\phi, \boldsymbol{F})$ and Gibbs' relation, $\eta = -\partial\psi/\partial T$, to generate expressions for $\dot{\psi}$ and $\dot{\eta}$:

$$\dot{\psi} = \frac{\partial\psi}{\partial\boldsymbol{F}}\boldsymbol{F}^T : \vec{\nabla}\boldsymbol{v} + \frac{\partial\psi}{\partial T}\dot{T} + \frac{\partial\psi}{\partial\phi}\dot{\phi} + \frac{\partial\psi}{\partial\vec{\nabla}\phi}\cdot\dot{\vec{\nabla}\phi}\,, \tag{3.17}$$

$$\dot{\eta} = -\frac{\partial^2\psi}{\partial T\partial\boldsymbol{F}}\boldsymbol{F}^T : \vec{\nabla}\boldsymbol{v} - \frac{\partial^2\psi}{\partial T^2}\dot{T} - \frac{\partial^2\psi}{\partial T\partial\phi}\dot{\phi} - \frac{\partial^2\psi}{\partial T\partial\vec{\nabla}\phi}\cdot\dot{\vec{\nabla}\phi}\,. \tag{3.18}$$

We then insert these expressions into (3.7) and make some further simplifications. In the resulting collection, terms proportional to $\dot{T}$ drop out because of Gibb's relation $\eta = -\partial\psi/\partial T$. Likewise, terms proportional to $\dot{\vec{\nabla}\phi}$ drop out because of the relation for the configurational stress $\boldsymbol{\xi} = \rho(\partial\psi/\partial\vec{\nabla}\phi)$. A collection of terms appear that is associated with the stress-related dissipation

$$\left(\boldsymbol{\sigma} - \rho\frac{\partial\psi}{\partial\boldsymbol{F}}\boldsymbol{F}^T + \vec{\nabla}\phi\otimes\boldsymbol{\xi}\right) : \vec{\nabla}\boldsymbol{v} = \boldsymbol{\sigma}^{diss} : \vec{\nabla}\boldsymbol{v},$$

and a collection of terms associated with the dissipation induced by the phase transformation appears,

$$-\left(\rho\frac{\partial\psi}{\partial\phi} + \pi_\phi\right)\dot{\phi} = B\dot{\phi}^2\,.$$

Using the classical definition of the specific heat at constant deformation (volume), $c_v \equiv T(\partial\eta/\partial T)_{\boldsymbol{F}} = -T(\partial^2\psi/\partial T^2)_{\boldsymbol{F}}$, the energy equation can be rewritten as follows:

$$\begin{aligned}\rho c_v\dot{T} = -\vec{\nabla}\cdot\boldsymbol{q} + \boldsymbol{\sigma}^{diss} : \vec{\nabla}\boldsymbol{v} + B\dot{\phi}^2 + \rho T\frac{\partial^2\psi}{\partial T\partial\boldsymbol{F}}\boldsymbol{F}^T : \vec{\nabla}\boldsymbol{v} \\ + \ \rho T\frac{\partial^2\psi}{\partial T\partial\phi}\dot{\phi} + \rho T\frac{\partial^2\psi}{\partial T\partial\vec{\nabla}\phi}\cdot\dot{\vec{\nabla}\phi} + \rho r\,.\end{aligned} \tag{3.19}$$

Some straightforward physical interpretations can be made for the various terms. The term $\boldsymbol{\sigma}^{diss} : \vec{\nabla}\boldsymbol{v}$ is the viscous dissipation associated with the stress. The term $B\dot{\phi}^2$ is a dissipation associated with the phase change. The term $\rho T(\partial^2\psi/\partial T\partial\phi)\,\dot{\phi}$ is an energy source term that is associated with enthalpic changes in phase (similar to those associated with the heat of combustion for reacting flows). The term $\rho T(\partial^2\psi/\partial T\partial\boldsymbol{F})\boldsymbol{F}^T : \vec{\nabla}\boldsymbol{v}$ is (again) a stress work term classically associated with thermal stresses. Similarly, the term $\rho T(\partial^2\psi/\partial T\partial\vec{\nabla}\partial\phi)\cdot\dot{\vec{\nabla}}\phi$ is a thermal stress work term associated with the configurational stress of the phase change.

3.1.1. Invariance requirements and isotropy. Most energetic solids are encountered as fine-grained polycrystalline aggregates and are often modeled with conventional isotropic liquid and gaseous forms. We now restrict our attention to isotropic materials, and we ignore possible anisotropic properties in this model. As is conventional we require that the material response is invariant under superposed rigid changes of observer. It can be shown in a standard way that the constitutive dependence on the deformation gradient $\boldsymbol{F}$ can be replaced by the left Cauchy–Green tensor $\boldsymbol{B} = \boldsymbol{F}\boldsymbol{F}^T$ and that the dependence on the velocity gradient is replaced by the symmetric stretching tensor $\boldsymbol{D} = (\boldsymbol{L} + \boldsymbol{L}^T)/2$. Furthermore, isotropy requires that the dependence on $\boldsymbol{B}$ appears through its principal scalar invariants $I_{\boldsymbol{B}} = \text{trace}\boldsymbol{B}$, $II_{\boldsymbol{B}} = \frac{1}{2}\big((\text{trace}\boldsymbol{B})^2 - \text{trace}(\boldsymbol{B}^2)\big)$, and $III_{\boldsymbol{B}} = \det\boldsymbol{B}$.

3.1.2. Constitutive specification of the Helmholtz free energy. Having made arguments that constrain the general form of the constitutive description, we next specialize the forms to extend the phase field constitutive forms and to capture commonly used classical forms for the pure solid, liquid, and gas phases. Without regard to exothermic chemical reaction, we will assume that the Helmholtz free energy is composed of four parts, such that we can write

$$\psi = \psi_1 + \psi_2 + \psi_3 + \psi_4\,. \tag{3.20}$$

The first two, ψ_1, ψ_2, are to be associated with the formulation of the phase transformations—the phase gradient energy density and the enthalpies associated with the phase transition. The latter two, ψ_3, ψ_4, are of classical origins—the thermal energy density and the strain energy density.

We assume that the Helmholtz free energy depends on $\vec{\nabla}\phi$ only through ψ_1 and that the phase gradient energy density is specified with the explicit quadratic dependence by

$$\psi_1 = \tfrac{1}{2}\gamma_\phi|\vec{\nabla}\phi|^2\,. \tag{3.21}$$

It follows from (3.12) that the configurational force $\boldsymbol{\xi}$ is determined by the formula

$$\boldsymbol{\xi} = \rho\frac{\partial\psi}{\partial\vec{\nabla}\phi} = \rho\gamma_\phi\vec{\nabla}\phi\,. \tag{3.22}$$

The physical interpretation of the phase-configurational stress $\boldsymbol{\xi}$ is as a traction that acts near or in the phase transition region in the direction of the gradient of $\vec{\nabla}\phi$, i.e., perpendicular to contours of constant ϕ.

Next we consider the contribution ψ_2, the phase transition energy density which reflects enthalpy changes during phase transition and is specified as

$$\psi_2 = \frac{1}{2}\Psi^{well}\mathcal{F}(\phi) + \beta_m(\phi)Q_m\Big(\frac{T}{T_m} - 1\Big) + \beta_v(\phi)Q_v\Big(\frac{T}{T_v} - 1\Big)\,. \tag{3.23}$$

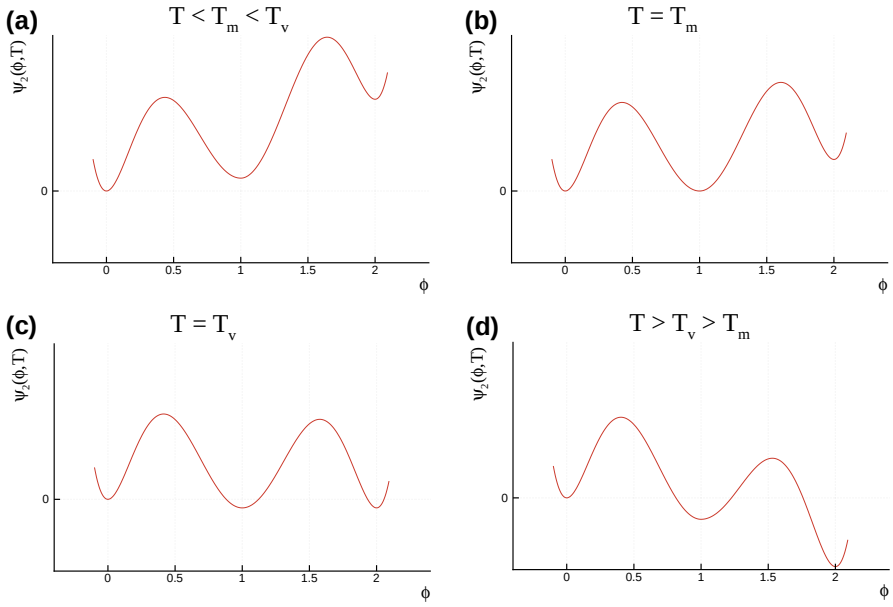


FIG. 1. *Plot of ψ_2 as a function of ϕ with T variation.*

The constants $\Psi^{well} > 0$, $Q_m < 0$, and $Q_v < 0$ represent a potential well depth and the heats of melting and vaporization. The constants $T_m > 0$ and $T_v > 0$ represent temperatures of melting and vaporization. The triple-well potential $\mathcal{F}(\phi)$ can be described by a smooth positive definite function whose isolated zeros are at $\phi = 0, 1$, and 2, representing three local minima. In addition, $\mathcal{F}(\phi)$ is assumed to be locally quadratic near the zeros at $\phi = 0, 1$, and 2, i.e., near $\phi = 0$, $\mathcal{F} \sim \phi^2$, near $\phi = 1$, $\mathcal{F} \sim (\phi - 1)^2$, and near $\phi = 2$, $\mathcal{F} \sim (\phi - 2)^2$. As an illustration, $\mathcal{F} = [\phi(\phi - 1)(\phi - 2)]^2$ has this property. The function $\beta_m(\phi)$ is assumed to be smooth and monotonically increasing and has values from 0 to 1 on the range $0 \le \phi \le 1$ with zero derivative elsewhere. The function $\beta_v(\phi)$ is similarly assumed to be monotonically increasing with values from 0 to 1 on the range $1 \le \phi \le 2$. Note that the derivative of transition energy density $\partial\psi_2/\partial\phi$ generates source terms in both the energy and phase equations represented as

$$\frac{\partial \psi_2}{\partial \phi} = \frac{1}{2}\Psi^{well}\frac{\partial \mathcal{F}}{\partial \phi}(\phi) + \beta_m'(\phi)Q_m\Big(\frac{T}{T_m} - 1\Big) + \beta_v'(\phi)Q_v\Big(\frac{T}{T_v} - 1\Big). \tag{3.24}$$

Figure 1 illustrates the assumed dependence of $\psi(\phi, T)$ on ϕ and T. Starting from (a) through (d), temperature T is raised from below T_m to above T_v, representing a standard melting-evaporation process. The transition energy density in case (a) has its minimum at $\phi = 0$. As T is increased through T_m and then T_v, we see a shift in the global minima from pure solid to solid-liquid and to liquid-vapor. As T eventually exceeds T_v as shown in (d), the energy minimizing well shifts to a vapor state at $\phi = 2$. The coefficients and functions Ψ^{well}, β_m, β_v can be adjusted (if needed) to reflect more accurately the physical properties observed in accordance with the phase transformation. Here we have chosen very simple forms.

We again assume the classical form for the thermal energy density and choose ψ_3

(which has the same form as ψ_1 in section 2) to be

$$\psi_3 = c_v \left[(T - T_0) - T \ln(T/T_0) \right], \tag{3.25}$$

where c_v is the specific heat at constant deformation. This is consistent with simple ideal models for solids, liquids, and gases.

Finally we choose a form for ψ_4, the strain energy density. We assume that it is composed of three subparts. The first part is associated with the thermal expansion stresses commonly identified in the condensed phase:

$$\psi_{4a} = -\frac{\alpha_c(\phi) K}{2\rho_0} (T - T_0) \ln(III_{\boldsymbol{B}}), \tag{3.26}$$

where K is the solid bulk modulus and α_c is the linear coefficient of thermal expansion. We again take $\alpha_c(\phi)$ to be a smooth, nonzero function in the condensed phases, solid and liquid, and zero in the gas phase. For example, $\alpha_c(0) = \alpha_{solid}$, $\alpha_c(1) = \alpha_{liquid}$, and $\alpha_c(2) = 0$. The second part of ψ_4 is associated with the pressure commonly identified in an ideal gas that we encountered in the previous section on gaseous combustion:

$$\psi_{4b} = -\frac{1}{2} R_g(\phi) T \ln(III_{\boldsymbol{B}}). \tag{3.27}$$

Here $R_g(\phi)$ plays the role of the ideal gas constant except that it is assumed to be nonzero in the gas phase and at or near zero in the solid and liquid condensed phase such that $R_g(0) = 0$, $R_g(1) = 0$, $R_g(2) = R_{gas}$.

The third part, ψ_{4c}, is based on properties of a compressible neo-Hookean, Blatz–Ko solid [15] which is given as

$$\psi_{BK} = \frac{\mu}{2\rho_0} (I_{\boldsymbol{B}} - 3) + \frac{\mu(1 - 2\nu)}{2\rho_0 \nu} \left(III_{\boldsymbol{B}}^{-\nu/(1-2\nu)} - 1 \right). \tag{3.28}$$

The constants ν and μ here represent the Poisson ratio of the material and the elastic Lamé parameter, μ. The contribution to the stress associated with this potential is

$$\boldsymbol{\sigma}_{BK} = 2\rho \frac{\partial \psi_{BK}}{\partial \boldsymbol{B}} \boldsymbol{B} = \mu_s \frac{\rho}{\rho_0} \boldsymbol{B} - \mu \frac{\rho}{\rho_0} III_{\boldsymbol{B}}^{-\nu/(1-2\nu)} \boldsymbol{I}. \tag{3.29}$$

We use this to model the elastic deformation of the solid, but for the liquid we pose a slightly altered form of this potential based on purely isotropic deformations. Consider the isotropic (either uniform contraction or expansion) given by $\boldsymbol{x} = s\boldsymbol{X}$, where s is the stretch ratio of material line segments. It follows simply that $\boldsymbol{F} = s\boldsymbol{I}$, $\boldsymbol{B} = s^2 \boldsymbol{I}$, $III_{\boldsymbol{B}} = \det(\boldsymbol{B}) = (\rho_0/\rho)^2 = s^6$, $s = (\rho_0/\rho)^{1/3}$, $\boldsymbol{B} = (\rho_0/\rho)^{1/3} \boldsymbol{I}$, and $(\rho_0/\rho)^{1/3} = III_{\boldsymbol{B}}^{1/6}$. For the Blatz–Ko solid, the isotropic stress is related to the volume ratio by

$$\boldsymbol{\sigma} = -\mu \frac{\rho}{\rho_0} \left[\left(\frac{\rho}{\rho_0} \right)^{-\frac{2\nu}{1-2\nu}} - \left(\frac{\rho}{\rho_0} \right)^{-\frac{1}{3}} \right] \boldsymbol{I}. \tag{3.30}$$

We choose our model for the strain energy of the liquid to have the same functional form for the isotropic stress dependence on the density ratio as that for the solid, and

merely note that we replace the dependence on ρ_0/ρ by $III_{\boldsymbol{B}}^{1/2}$ and work backwards. The corresponding Helmholtz free energy for the liquid would take the form

$$\psi_{BK(liquid)} = \frac{3}{2}\frac{\mu}{\rho_0} III_{\boldsymbol{B}}^{1/3} + \frac{\mu(1-2\nu)}{2\rho_0\nu}\left(III_{\boldsymbol{B}}^{-\nu/(1-2\nu)} - 1\right). \tag{3.31}$$

We can combine the two potentials for the solid and the liquid in the following way. Let $\mu_s(\phi)$ be a coefficient such that $\mu_s(0) = \mu_{solid}$ and it is zero for $\phi \geq 1$. Let $\mu_l(\phi)$ be a smooth function such that $\mu_l(1) = \mu_{liquid}$ with $\mu_l(0) = \mu_l(2) = 0$. One makes similar definitions for ν_s and ν_l. Let μ_c be defined as the sum $\mu_c = \mu_l + \mu_s$, and $\nu_c = \nu_l + \nu_s$. Then the combined solid, liquid, elastic potential can be written as

$$\psi_{4c} = \frac{\mu_s}{2\rho_0}(I_{\boldsymbol{B}} - 3) + \frac{3}{2}\frac{\mu_l}{\rho_0} III_{\boldsymbol{B}}^{1/3} + \frac{\mu_c(1-2\nu_c)}{2\rho_0\nu_c}\left(III_{\boldsymbol{B}}^{-\nu_c/(1-2\nu_c)} - 1\right). \tag{3.32}$$

Note that other functional forms for the strain energy density could have been chosen for ψ_{4c}, but we chose the Blatz–Ko form since it has a simple reduction to compressible linear elasticity in the limit of small strain, which is deemed convenient for our purposes. We anticipate that as the solid become significantly nonlinearly elastic, we expect that a phase transformation will occur so that the specific choice of Blatz–Ko is not a sensitive one for the properties of the model. The deformational portion of stress associated with this strain energy is ψ_{4c},

$$\boldsymbol{\sigma}^{def} \equiv 2\rho\frac{\partial\psi_{4c}}{\partial\boldsymbol{B}}\boldsymbol{B} = \mu_s\frac{\rho}{\rho_o}\boldsymbol{B} - \mu_c\frac{\rho}{\rho_o} III_{\boldsymbol{B}}^{-\nu_s/(1-2\nu_s)}\boldsymbol{I} + \mu_l\frac{\rho}{\rho_o} III_{\boldsymbol{B}}^{1/3}\boldsymbol{I}. \tag{3.33}$$

3.2. Total free-energy density and summary of constitutive forms. The form of $\psi = \psi_1 + \psi_2 + \psi_3 + \psi_{4a} + \psi_{4b} + \psi_{4c}$ is written as

$$\begin{aligned}
\psi = {} & \frac{\mu_s(\phi)}{2\rho_0}(I_{\boldsymbol{B}} - 3) + \frac{\mu_c(\phi)(1-2\nu_s)}{2\rho_0\nu_s}\left(III_{\boldsymbol{B}}^{-\nu_s/(1-2\nu_s)} - 1\right) + \frac{3\mu_l(\phi)}{2\rho_o} III_{\boldsymbol{B}}^{1/3} \\
& - \frac{\alpha_c(\phi)K}{2\rho_0}(T - T_0)\ln(III_{\boldsymbol{B}}) - \frac{1}{2}R_g(\phi)T\ln(III_{\boldsymbol{B}}) && \text{strain energy density} \\
& - c_v(\phi)\left[T\ln\left(\frac{T}{T_0}\right) - (T - T_0)\right] && \text{thermal energy density} \\
& + \frac{1}{2}\Psi^{well}\mathcal{F}(\phi) + \beta_m(\phi)\left(\frac{T}{T_m} - 1\right)Q_m + \beta_v(\phi)\left(\frac{T}{T_v} - 1\right)Q_v && \text{phase transition energy density} \\
& + \frac{1}{2}\gamma_\phi|\vec{\nabla}\phi|^2. && \text{gradient energy density}
\end{aligned} \tag{3.34}$$

The constitutive theory is essentially complete. The stress is given by the general expression

$$\boldsymbol{\sigma} = \rho\frac{\partial\psi}{\partial\boldsymbol{B}}\boldsymbol{B} - \vec{\nabla}\phi\otimes\boldsymbol{\xi} + \boldsymbol{\sigma}^{diss}, \tag{3.35}$$

with $\boldsymbol{\xi}$ given by $\boldsymbol{\xi} = \rho\gamma_\phi\vec{\nabla}\phi$ and $\boldsymbol{\sigma}^{diss}$ given by $\boldsymbol{\sigma}^{diss} = \mu_f(\vec{\nabla}\cdot\boldsymbol{v})\boldsymbol{I} + 2\mu_f\boldsymbol{D}$. The stress formula becomes

$$\begin{aligned}
\boldsymbol{\sigma} = {} & \mu_s\frac{\rho}{\rho_0}\boldsymbol{B} - \mu_c\frac{\rho}{\rho_0} III_{\boldsymbol{B}}^{-(\nu_s/1-2\nu_s)}\boldsymbol{I} + \mu_l\frac{\rho}{\rho_o} III_{\boldsymbol{B}}^{1/3}\boldsymbol{I} \\
& - \alpha_c(\phi)K\frac{\rho}{\rho_0}(T - T_0)\boldsymbol{I} - \rho R_g(\phi)T\boldsymbol{I} \\
& - \rho\gamma_\phi\vec{\nabla}\phi\otimes\vec{\nabla}\phi + \nu_f(\vec{\nabla}\cdot\boldsymbol{v})\boldsymbol{I} + 2\mu_f\boldsymbol{D}.
\end{aligned} \tag{3.36}$$

The energy flux vector remains $\boldsymbol{q} = -k\vec{\nabla}T$. The various source terms in the energy and phase equations can be computed from the forms given in (3.34).

We can now summarize the governing equations for the phase change model as

$$\dot{\rho} + \rho\vec{\nabla}\cdot\boldsymbol{v} = 0\,, \tag{3.37}$$

$$\rho\,\dot{\boldsymbol{v}} = \vec{\nabla}\cdot\boldsymbol{\sigma} + \rho\boldsymbol{f}\,, \tag{3.38}$$

$$\begin{aligned}\rho c_v\dot{T} = {}& \vec{\nabla}\cdot(k\vec{\nabla}T) + \boldsymbol{\sigma}^{diss} : \vec{\nabla}\boldsymbol{v} + B\dot{\phi}^2 + \rho T\frac{\partial^2\psi}{\partial T\partial\boldsymbol{F}}\boldsymbol{F}^T : \vec{\nabla}\boldsymbol{v} \\ & + \rho T\frac{\partial^2\psi}{\partial T\partial\phi}\dot{\phi} + \rho T\frac{\partial^2\psi}{\partial T\partial\vec{\nabla}\phi}\cdot\dot{\vec{\nabla}}\phi + \rho r\,,\end{aligned} \tag{3.39}$$

$$B\dot{\phi} = \vec{\nabla}\cdot(\rho\gamma_\phi\vec{\nabla}\phi) - \rho\frac{\partial\psi}{\partial\phi}\,, \tag{3.40}$$

$$\dot{\boldsymbol{F}} = \boldsymbol{L}\boldsymbol{F}, \tag{3.41}$$

where B, c_v, γ_ϕ, k, etc. are constitutive scalars which could be regarded as functions of both ϕ and T. We have added the kinematic identity (3.41) in order to compute the evolution of the displacement gradients.

4. The combined model: Modifications to include chemical reaction. Here we list the modifications required to combine both models into one. First we take the phase change model as the starting point and we retain all the assumptions and assumed forms of the previous section, specifically in regards to the appearance of ϕ. The configurational force balance (3.2) is retained as a fundamental balance law (the consequence of which leads to the derivation of the evolution equation for ϕ, equation (3.40)).

Next we assume that, in addition to ϕ, which measures the molecular order of the phase, the mass fraction λ simultaneously measures the amount of exothermic chemical reaction that has taken place. So λ is added to all the argument lists; in particular, in the expression for ψ we assume the dependence

$$\psi = \psi(\boldsymbol{F}, T, \phi, \lambda, \vec{\nabla}\phi, \boldsymbol{L})\,. \tag{4.1}$$

A statement of conservation of λ is added in the form of (2.4), which reflects a molecularly based conservation of species. The second law must be modified to include the entropy flux associated with the heat of combustion (so it takes the same form as (2.5)),

$$\rho\,\dot{\eta} \geq -\vec{\nabla}\cdot\left(\frac{\boldsymbol{q}}{T}\right) + \vec{\nabla}\cdot\left(\frac{Q_{hc}\boldsymbol{s}}{T}\right) + \frac{\rho r}{T}\,. \tag{4.2}$$

One argues the entropy inequality in exactly the same manner as in the previous section, with the same assumptions and conclusions of section 3, with the additional exception that one uses the evolution equation for λ, (2.4), to reduce the dissipation inequality in the manner explained in section 2. The energy flux vector is identified by the requirement of positivity of the left-hand side of (2.15), which leads to

$$\boldsymbol{q} = -k\vec{\nabla}T + Q_{hc}\boldsymbol{s}\,. \tag{4.3}$$

The vector $\boldsymbol{s}$ can be chosen according to Fick's law such that

$$\boldsymbol{s} = d\,\vec{\nabla}\lambda\,. \tag{4.4}$$

The Helmholtz free energy is designated as $\psi = \psi_1+\psi_2+\psi_3+\psi_{4a}+\psi_{4b}+\psi_{4c}+\psi_5$, where ψ_{1-4c} are defined in the previous section and ψ_5 is the chemical enthalpy $\psi_5 = -Q_{hc}\lambda$. The configurational stress is again of the form $\boldsymbol{\xi} = \rho\gamma_\phi\vec{\nabla}\phi$. The representation of the stress is

$$\begin{aligned}\boldsymbol{\sigma} = {}& \mu_s\frac{\rho}{\rho_0}\boldsymbol{B} - \mu_c\frac{\rho}{\rho_0}III_{\boldsymbol{B}}^{-(\nu_s/(1-2\nu_s)}\boldsymbol{I} + \mu_l\frac{\rho}{\rho_o}III_{\boldsymbol{B}}^{1/3}\boldsymbol{I}\\ & - \alpha_c(\phi)K\frac{\rho}{\rho_0}(T-T_0)\boldsymbol{I} - \rho R_g(\phi)T\boldsymbol{I}\\ & - \rho\gamma_\phi\vec{\nabla}\phi\otimes\vec{\nabla}\phi + \nu_f(\vec{\nabla}\cdot\boldsymbol{v})\boldsymbol{I} + 2\mu_f\boldsymbol{D}\,.\end{aligned} \tag{4.5}$$

The various scalar material properties identified previously, such as $c_v, \gamma_\phi, \ldots$, now can also have explicit dependence on λ as well as ϕ and T.

A revised list of the governing equations for the combined model with reaction and phase change is

$$\dot{\rho} + \rho\vec{\nabla}\cdot\boldsymbol{v} = 0\,, \tag{4.6}$$

$$\rho\,\dot{\boldsymbol{v}} = \vec{\nabla}\cdot\boldsymbol{\sigma} + \rho\boldsymbol{f}\,, \tag{4.7}$$

$$\begin{aligned}\rho c_v\dot{T} = {}& \vec{\nabla}\cdot(k\vec{\nabla}T) + \boldsymbol{\sigma}^{diss}:\vec{\nabla}\boldsymbol{v} + B\dot{\phi}^2 + \rho T\frac{\partial^2\psi}{\partial T\partial\boldsymbol{F}}\boldsymbol{F}^T:\vec{\nabla}\boldsymbol{v}\\ & + \rho T\frac{\partial^2\psi}{\partial T\partial\phi}\dot{\phi} + \rho T\frac{\partial^2\psi}{\partial T\partial\vec{\nabla}\phi}\cdot\dot{\vec{\nabla}\phi} + \rho\,Q_{hc}\,\Omega + \rho r\,,\end{aligned} \tag{4.8}$$

$$B\dot{\phi} = \vec{\nabla}\cdot(\rho\gamma_\phi\vec{\nabla}\phi) - \rho\frac{\partial\psi}{\partial\phi}\,, \tag{4.9}$$

$$\rho\dot{\lambda} = \vec{\nabla}\cdot(d\vec{\nabla}\lambda) + \rho\,\Omega\,, \tag{4.10}$$

$$\dot{\boldsymbol{F}} = \boldsymbol{L}\boldsymbol{F}\,. \tag{4.11}$$

With the specific constitutive forms chosen for ψ, the energy equation becomes

$$\begin{aligned}\rho c_v\dot{T} = {}& \vec{\nabla}\cdot(k\vec{\nabla}T)+\nu_f(\vec{\nabla}\cdot\boldsymbol{v})^2+2\mu_f\boldsymbol{D}:\boldsymbol{D}+B\dot{\phi}^2-\alpha_c(\phi)K\frac{\rho}{\rho_0}T(\vec{\nabla}\cdot\boldsymbol{v})-\rho R_g(\phi)T(\vec{\nabla}\cdot\boldsymbol{v})\\ & + \left\{-\frac{\alpha_c'(\phi)}{2}K\frac{\rho}{\rho_0}T\ln(III_{\boldsymbol{B}}) - \rho\frac{R_g'(\phi)}{2}T\ln(III_{\boldsymbol{B}}) - \rho c_v'(\phi)T\ln\left(\frac{T}{T_0}\right)\right.\\ & \left. + \rho\left[\beta_m'(\phi)\frac{T}{T_m}Q_m + \beta_v'(\phi)\frac{T}{T_v}Q_v\right]\right\}\dot{\phi} + \rho Q_{hc}\Omega + \rho r,\end{aligned} \tag{4.12}$$

and the evolution law for ϕ becomes

$$\begin{aligned}B\dot{\phi} = {}& \vec{\nabla}\cdot\left(\rho\gamma_\phi\vec{\nabla}\phi\right) + \rho c_v'(\phi)\left[T\ln\left(\frac{T}{T_0}\right) - (T-T_0)\right]\\ & - \frac{\mu_s'(\phi)}{2}\frac{\rho}{\rho_0}(I_{\boldsymbol{B}}-3) - \frac{\mu_c'(\phi)}{2}\frac{\rho}{\rho_0}\frac{(1-2\nu_s)}{\nu_s}\left(III_{\boldsymbol{B}}^{-\nu_s/(1-2\nu_s)}-1\right) + \frac{3\mu_l'(\phi)}{2}\frac{\rho}{\rho_o}III_{\boldsymbol{B}}^{1/3}\\ & + \frac{\alpha_c'(\phi)}{2}K\frac{\rho}{\rho_0}(T-T_0)\ln(III_{\boldsymbol{B}}) + \frac{1}{2}\rho R_g'(\phi)T\ln(III_{\boldsymbol{B}})\\ & - \rho\frac{1}{2}\Psi^{well}\frac{\partial\mathcal{F}}{\partial\phi} - \rho\left[\beta_m'(\phi)\left(\frac{T}{T_m}-1\right)Q_m + \beta_v'(\phi)\left(\frac{T}{T_v}-1\right)Q_v\right].\end{aligned} \tag{4.13}$$

4.1. Material transition functions. An important ingredient of our model is the use of ϕ-dependent material properties or material transition functions. Earlier in section 2.4.2, we encountered $\beta_m(\phi), \beta_v(\phi)$ in the specification of the phase transition energy density, $\mu_c(\phi), \mu_l(\phi), \mu_s(\phi), \alpha_c(\phi), R_g(\phi)$ in the specification of the strain energy density, $c_v(\phi)$ in the specification of the thermal energy density, as well as functions associated with dissipative processes like $\nu_f(\phi)$. The model assumes that these functions have limiting pure phase values when $\phi = 0, 1, 2$. The structure of these functions has an influence on the exact details of the spatial structure of the transition layers and their dynamics when particular problems are solved. However, one makes an implicit assumption that when the transitions occur in thin layers relative to other geometric lengths, the structure within the layer does not strongly influence the information transmitted across the layer. This modeling precept is consistent with the use of viscous dissipation to describe continuum shock structure when the shock is molecularly thin.

For illustration sake, Figures 2 and 3 show typical transition functions that we have used to carry out representative simulations discussed in the companion paper [13]. These functions are constructed from simple polynomials in ϕ and their smooth extensions. The figures clearly show the basic properties that are required. For example, in Figure 2(c), the representation of the thermal expansion parameter $\alpha_c(\phi)$, which has the same (constant) value in the solid and liquid phase, is zero in the gas phase. Another example is that $\beta'_m(\phi)$ is zero for all values of ϕ except for those between 0 and 1, and terms that multiply $\beta'_m(\phi)$ are only involved in the solid to liquid transitions of melting or freezing and are totally absent in the liquid-gas transition of evaporation and condensation.

5. Some limiting cases.

5.1. Pure phases. The results for pure phases can be identified by the constitutive forms for the stress tensor. First we will consider the solid, $\phi = 0$, in the additional limit of small strain. The small strain limit is represented in terms of the displacement gradient $\boldsymbol{H} = \boldsymbol{F} - \boldsymbol{I}$, where $|\boldsymbol{H}| << 1$. Define the small strain tensor $\boldsymbol{E} = (\boldsymbol{H} + \boldsymbol{H}^T)/2$, and the left Cauchy–Green tensor can be written as $\boldsymbol{B} = \boldsymbol{F}\boldsymbol{F}^T = \boldsymbol{I} + 2\boldsymbol{E} + \boldsymbol{H}\boldsymbol{H}^T$. Our limiting form for the stress relation reduces to

$$\boldsymbol{\sigma} = -\alpha_{solid}\,\frac{\rho}{\rho_0} K(T - T_0)\boldsymbol{I} + \frac{2\,\mu_{solid}\;\nu_{solid}}{1 - 2\,\nu_{solid}}\, I_{\boldsymbol{E}}\,\boldsymbol{I} + 2\,\mu_{solid}\,\boldsymbol{E}\,, \tag{5.1}$$

When one considers the limit of a liquid, $\phi = 1$, the expression for the stress becomes

$$\boldsymbol{\sigma} = -\alpha_{liquid}\,K\frac{\rho}{\rho_0}(T - T_0)\,\boldsymbol{I} - \mu_{liquid}\,\frac{\rho}{\rho_0}\left(\left(\frac{\rho}{\rho_0}\right)^{2\,\nu_{liquid}/(1-2\,\nu_{liquid})} - \left(\frac{\rho}{\rho_0}\right)^{-2/3}\right)\boldsymbol{I} + \nu_{liquid}(\vec{\nabla}\cdot\boldsymbol{v})\boldsymbol{I} + 2\mu_{liquid}\boldsymbol{D}; \tag{5.2}$$

similarly for the limit of the gas, $\phi = 2$, the expression for the stress becomes

$$\boldsymbol{\sigma} = -\rho R_{gas} T\boldsymbol{I} + \nu_{gas}(\vec{\nabla}\cdot\boldsymbol{v})\boldsymbol{I} + 2\mu_{gas}\boldsymbol{D}\,. \tag{5.3}$$

5.2. Motionless phase transition. In this case we simply assume that the system is nearly motionless with $\boldsymbol{v} \approx 0$ and consider the pure phase change from solid

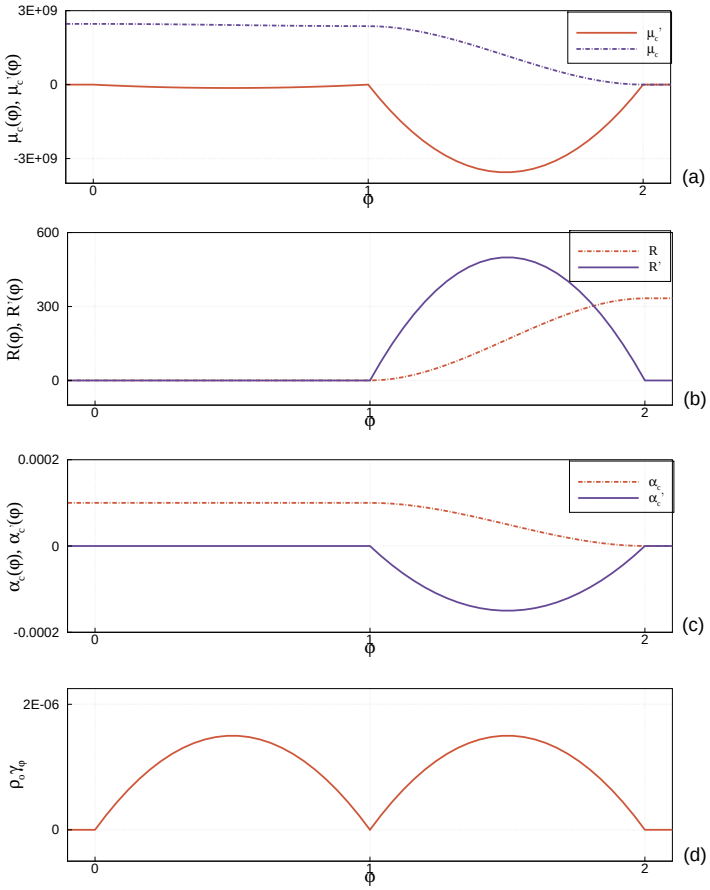


FIG. 2. *Plots of transition functions for HMX simulation and their derivatives with respect to the phase variable. Shear modulus, ideal gas constant, thermal expansion coefficient, and phase diffusion coefficient are shown from top to bottom.*

to liquid with no chemical reaction. In addition, we neglect the thermal expansion configurational forces, consistent with a nearly zero velocity field, and the thermal dissipation associated with the phase transition. Further we assume that ϕ is in the range $0 \leq \phi \leq 1$ and $\mathcal{F}(\phi)$ is effectively a double-well potential. We take the specific heat to be constant and are left with a thermal-diffusional model for the temperature and phase field given by the equations

$$\rho c_v \dot{T} = \vec{\nabla} \cdot (k \vec{\nabla} T) + \rho \beta'_m(\phi) \frac{T}{T_m} Q_m \dot{\phi} \tag{5.4}$$

and

$$B\dot{\phi} = \vec{\nabla} \cdot (\rho \gamma_\phi \vec{\nabla} \phi) - \rho \frac{1}{2} \Psi^{well} \frac{\partial \mathcal{F}}{\partial \phi} - \rho \beta'_m(\phi) \left(\frac{T}{T_m} - 1 \right) Q_m. \tag{5.5}$$

These equations are a generalized form of a thermally dependent Ginzburg–Landau theory of phase transitions often cited in discussions of solidification of binary alloys (see, for example, Wheeler, Boettinger, and McFadden [14].) Simple systems

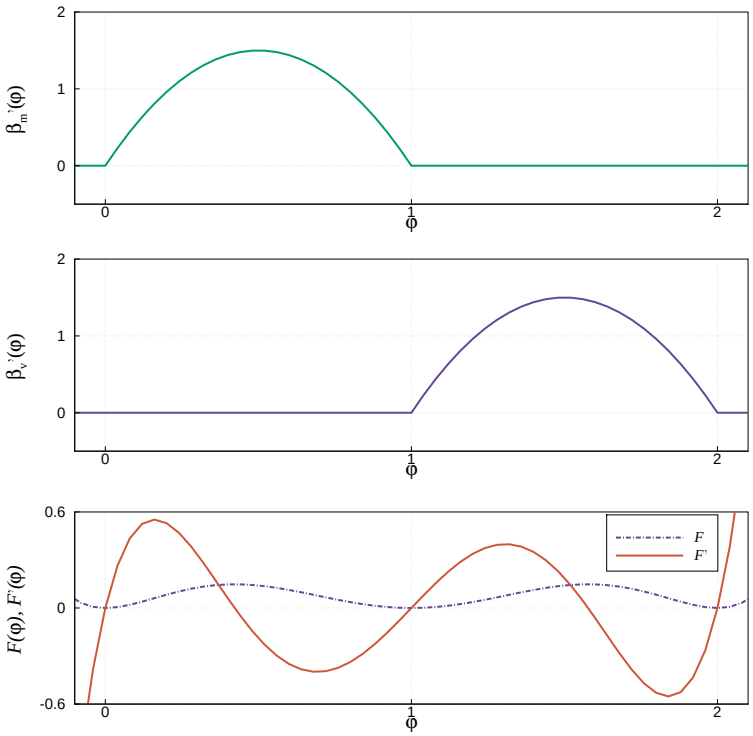


FIG. 3. *φ-dependent transfer functions (derivatives) for heat of phase transformations, β'_m and β'_v. The third figure depicts the triple-well Ginzburg–Landau potential function and its derivative.*

of this form, with a double-well potential and a single latent heat term, that have been analyzed in the literature have been shown to correspond to various forms of the classical (sharp interface) description of phase transitions. Further analysis leads to modified Stefan problems that incorporate surface tension and kinetic undercooling [8].

5.3. Relation to the simpler theory of quasi-static phase transformation. Here we briefly discuss the manner in which our model relates to the theory of quasi-static phase transformations that are a part of classical equilibrium thermodynamics. We assume that the changes in the state in the material happen so slowly that all inertial effects can be neglected and that the material undergoes only isotropic volume changes that are measured by changes in the density. The stress is spherical so that $\boldsymbol{\sigma} = -p\boldsymbol{I}$. The deformation is homogeneous such that $\boldsymbol{x} = s\boldsymbol{X}$, with $\boldsymbol{F} = s\boldsymbol{I}$, $\det(\boldsymbol{F}) = s = (\rho_0/\rho)$, $\boldsymbol{B} = (\rho_0/\rho)^2\boldsymbol{I}$, and strain invariants $III_{\boldsymbol{B}} = (\rho_0/\rho)^2$ and $I_{\boldsymbol{B}} - 3 = 3[(\rho_0/\rho)^2 - 1]$. One neglects all spatial gradients.

Next we consider the volume changes that occur as the temperature rises when the material is subjected to constant volumetric heating (given by constant r), under isobaric (constant pressure) conditions. For simplicity, we will also assume that the specific heat is constant in all phases. Then the change in the thermodynamic states would be controlled by a simplified version of the energy equation (for the

temperature) and the phase evolution equation. These are written as

$$\rho c_v \frac{\partial T}{\partial t} = \rho \left(\beta'_m(\phi) \frac{T}{T_m} Q_m + \beta'_v(\phi) \frac{T}{T_v} Q_v \right) \frac{\partial \phi}{\partial t} + \rho r \,, \tag{5.6}$$

$$B \frac{\partial \phi}{\partial t} = -\rho \frac{1}{2} \Psi^{well} \frac{\partial \mathcal{F}}{\partial \phi} - \rho \left[\beta'_m(\phi) \frac{T - T_m}{T_m} Q_m + \beta'_v(\phi) \frac{T - T_v}{T_v} Q_v \right], \tag{5.7}$$

and for the purpose of illustration, (4.5) is simplified by linearizing ρ about ρ_0 in the solid and liquid phases to obtain the thermal equation of state, a relation between p, ρ, T, and ϕ,

$$p = \frac{6\,\mu_c(\phi)\nu_s}{1 - 2\nu_s} \left(\frac{\rho}{\rho_0} - 1 \right) + \alpha_c K \frac{\rho}{\rho_0} (T - T_0) - \rho R_g T. \tag{5.8}$$

The above equations are solved subject to the initial condition that the material is initially solid and, at the reference temperature, $\phi(0) = 0$ and $T(0) = T_0$. For constant pressure, a specified temperature, and ϕ, (5.8) determines the specific volume, $V = 1/\rho$. The solution of the initial value problem for T and ϕ determines a trajectory in T, V, ϕ-space at fixed p. A typical solution shows that as the temperature rises in the solid, the volume increases along the isobar. A phase transition (change in ϕ) does not take place till the temperature nears the melting temperature, T_m. Above that temperature local analysis shows that a change in stability of the state $\phi = 0$ occurs and then the transition from $\phi = 0$ to $\phi = 1$ occurs. Since the volumetric change is small (4% or less), the deviation in a T, v isobar is not large in some sense. As the temperature continues to rise, the second phase transition occurs near the vaporization temperature, T_v. Since the thermal equation of state is effectively modeled by the ideal gas law, a rather large change in the specific volume occurs. Finally, after the phase transition to vapor is completed and $\phi = 2$ is reached, the temperature continues to climb on the gas phase isobar with increasing volume. Figure 4 show plots of a T, V-trajectory for a isobaric phase transition for the HMX-like material described in [13]. Figure 5 shows the corresponding ϕ, V-trajectory at different pressures. Again, the purpose here is simply to illustrate that conventional notions of quasi-static phase transformations described in classical thermodynamics are embedded in this model.

6. Special forms of the model for three simple motions. In this concluding section we write out special and exact forms of the differential equations for the model when the material undergoes three simple motions: (i) evolution at constant volume, (ii) one-dimensional, time-dependent, longitudinal motion, and (iii) one-dimensional, time-dependent shear motion. All three are very important in the analysis of ignition of EMs. The three cases are the exclusive subject of the companion paper [13], in which numerical simulation and the properties of the model are discussed further.

6.1. Constant volume evolution and thermal explosion. A simple but extremely important subcase that is studied extensively in combustion theory describes the constant volume thermal explosion, where the velocity $\boldsymbol{v}$ and all spatial gradients are exactly zero. The density is constant hence the volume of a material particle is constant. For illustration, we neglect thermal expansion and assume constant specific heat and gas constants. We are left with three ODEs in time for the temperature, phase change, and reaction progress:

$$\rho c_v \frac{\partial T}{\partial t} = \rho \left(\beta'_m(\phi) \frac{T}{T_m} Q_m + \beta'_v(\phi) \frac{T}{T_v} Q_v \right) \frac{\partial \phi}{\partial t} + \rho Q_{hc} \Omega + \rho r \,, \tag{6.1}$$

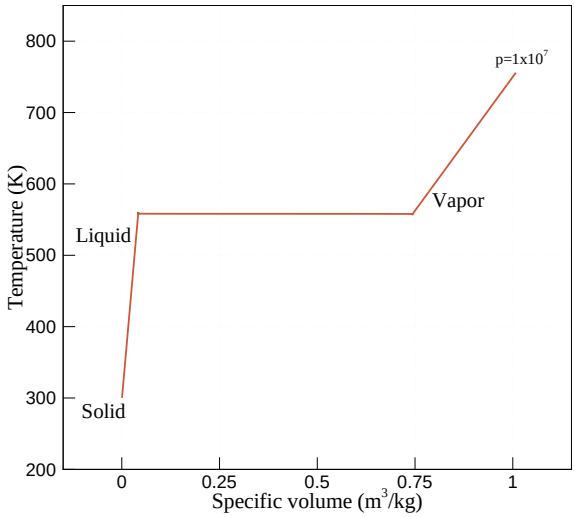


FIG. 4. *T, V-trajectory on an isobar ($p = 10^7$ Pa) under the quasi-static assumptions. The large volume jump from liquid to gas happens at nearly constant temperature T_v.*

$$B\frac{\partial \phi}{\partial t} = -\rho \frac{1}{2}\Psi^{well}\frac{\partial \mathcal{F}}{\partial \phi} - \rho\left[\beta_m'(\phi)\frac{T-T_m}{T_m}Q_m + \beta_v'(\phi)\frac{T-T_v}{T_v}Q_v\right], \tag{6.2}$$

$$\frac{\partial \lambda}{\partial t} = \Omega. \tag{6.3}$$

If one discards phase change, we recover the equations from standard combustion theory for constant volume thermal explosion, $c_v(\partial T/\partial t) = Q_{hc}\,\Omega\,, (\partial\lambda/\partial t) = \Omega$. Of course, the more interesting behavior occurs when phase change is included. The typical dynamics of these ODEs are discussed at length in [13].

6.2. Longitudinal motion. Next we turn to specializations of the equations to simplified motions that lead to PDEs in one space dimension and one time dimension; this is particularly suited to the study of ignition phenomena in EMs (which is one of our main concerns). First we consider longitudinal compression associated with a flyer-plate impact test. In this idealization, an infinite slab experiences a displacement loading normal to its surface. Specifically we consider the following one-dimensional motion:

$$x_1 = X_1 + f_1(X_1, t), \qquad x_2 = X_2, \qquad x_3 = X_3\,, \tag{6.4}$$

where f_1 is the 1-displacement.

For this motion, there is one nonzero velocity component, $v_1 = \partial f_1/\partial t|_{\boldsymbol{X}}(X_1, t)$, $\boldsymbol{F}$ is diagonal with $F_{11} = \partial x_1/\partial X_1 = 1 + f_1'$, and $F_{22} = F_{33} = 1$. The density is related to the single strain gradient by $1 + f_1' = \rho_0/\rho$. Also, $\boldsymbol{B}$ is diagonal with $B_{11} = (1 + f_1')^2$, $B_{22} = 1$, $B_{33} = 1$. The first and third invariants of $\boldsymbol{B}$ are $I_{\boldsymbol{B}} =$

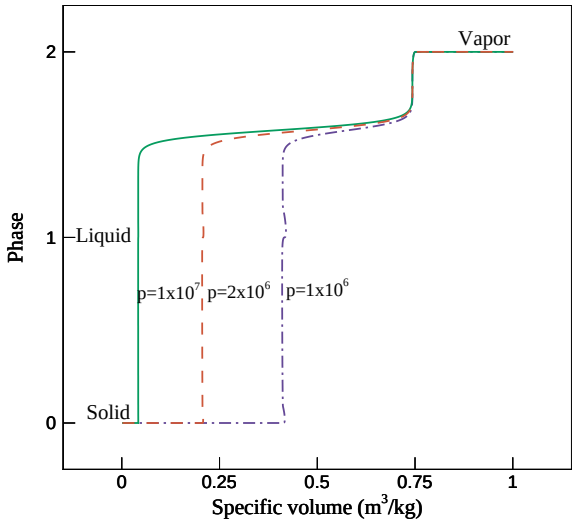


FIG. 5. *Phase-V trajectory of constant pressure under the thermo-quasistatic assumption.*

$2 + (1 + f_1')^2$ and $III_{\boldsymbol{B}} = (1 + f_1')^2$, with $III_{\boldsymbol{B}} = (\rho_0/\rho)^2$ and $I_{\boldsymbol{B}} - 3 = (\rho_0/\rho)^2 - 1$. Hence we use the density as the independent strain measure and replace f_1'. The one nonzero component of the velocity gradient and rate of strain tensor are, respectively, $L_{11} = D_{11} = \partial v_1/\partial x_1$. Also, $(\vec{\nabla}\phi \otimes \vec{\nabla}\phi)_{11} = (\partial\phi/\partial x_1)^2$. It then follows that all the shear stresses are zero, $\sigma_{12} = \sigma_{23} = \sigma_{13} = 0$, and the normal stresses σ_{11} are given by

$$\sigma_{11} = -\mu_c \left(\frac{\rho}{\rho_0}\right)\left[\left(\frac{\rho}{\rho_0}\right)^{\frac{2\nu_s}{1-2\nu_s}} - \left(\frac{\rho}{\rho_0}\right)^{-2}\right] - \alpha_c K \frac{\rho}{\rho_0}(T - T_0) - \rho R_g T - \rho\gamma_\phi \left(\frac{\partial\phi}{\partial x_1}\right)^2 + (\nu_f + 2\mu_f)\frac{\partial v_1}{\partial x_1}\,. \tag{6.5}$$

The other normal stress are the same as the σ_{11} stress, minus the phase stress, i.e., $\sigma_{22} = \sigma_{33} = \sigma_{11} + \rho\gamma_\phi(\partial\phi/\partial x_1)^2$.

The specific governing equations for longitudinal compression are the mass and momemtum equations

$$\frac{\partial\rho}{\partial t} + v_1\frac{\partial\rho}{\partial x_1} + \rho\frac{\partial v_1}{\partial x_1} = 0\,, \tag{6.6}$$

$$\rho\left(\frac{\partial v_1}{\partial t} + v_1\frac{\partial v_1}{\partial x_1}\right) = \frac{\partial}{\partial x_1}\left\{-\mu_c\left(\frac{\rho}{\rho_0}\right)\left[\left(\frac{\rho}{\rho_0}\right)^{\frac{2\nu_s}{1-2\nu_s}} - \left(\frac{\rho}{\rho_0}\right)^{-2}\right] - \alpha_c K\frac{\rho}{\rho_0}(T - T_0) - \rho R_g T - \rho\gamma_\phi\left(\frac{\partial\phi}{\partial x_1}\right)^2 + (\nu_f + 2\mu_f)\frac{\partial v_1}{\partial x_1}\right\} \tag{6.7}$$

and the energy balance, phase evolution, and reaction progress evolution equations that take the specific forms

$$
\begin{aligned}
\rho c_v \left(\frac{\partial T}{\partial t} + v_1 \frac{\partial T}{\partial x_1}\right) &= \frac{\partial}{\partial x_1}\left(k \frac{\partial T}{\partial x_1}\right) + (\nu_f + 2\mu_f)\left(\frac{\partial v_1}{\partial x_1}\right)^2 \\
&+ B\dot{\phi}^2 - \left[\alpha_c K \frac{\rho}{\rho_0} T + \rho R_g(\phi) T\right] \frac{\partial v_1}{\partial x_1} \\
&+ \left\{ -\frac{1}{2}\alpha_c'(\phi) K \frac{\rho}{\rho_0} T \ln(III_{\boldsymbol{B}}) - \frac{1}{2}\rho R_g'(\phi) T \ln(III_{\boldsymbol{B}}) - \rho c_v'(\phi) T \ln(T/T_0) \right. \\
&\left. + \rho \left[\beta_m'(\phi) \frac{T}{T_m} Q_m + \beta_v'(\phi) \frac{T}{T_v} Q_v\right] \right\} \dot{\phi} + \rho\, Q_{hc} \Omega + \rho r,
\end{aligned} \tag{6.8}
$$

$$
\begin{aligned}
&B\left(\frac{\partial \phi}{\partial t} + v_1 \frac{\partial \phi}{\partial x_1}\right) = \frac{\partial}{\partial x_1}\left(\rho \gamma_\phi \frac{\partial \phi}{\partial x_1}\right) + \rho c_v'(\phi)\left[T \ln\left(\frac{T}{T_0}\right) - (T - T_0)\right] \\
&- \frac{\mu_s'(\phi)}{2}\frac{\rho}{\rho_0}(I_{\boldsymbol{B}} - 3) - \frac{\mu_c'(\phi)}{2}\frac{\rho}{\rho_0}\frac{(1-2\nu_s)}{\nu_s}\left(III_{\boldsymbol{B}}^{-\nu_s/(1-2\nu_s)} - 1\right) - \frac{3\,\mu_l'(\phi)}{2}\frac{\rho}{\rho_o} III_{\boldsymbol{B}}^{1/3} \\
&+ \frac{\alpha_c'(\phi)}{2} K \frac{\rho}{\rho_0}(T - T_0)\ln(III_{\boldsymbol{B}}) + \frac{1}{2}\rho R_g'(\phi) T \ln(III_{\boldsymbol{B}}) \\
&- \frac{1}{2}\rho \Psi^{well}\frac{\partial \mathcal{F}}{\partial \phi} - \rho\left[\beta_m'(\phi)\frac{T - T_m}{T_m} Q_m + \beta_v'(\phi)\frac{T - T_v}{T_v} Q_v\right],
\end{aligned} \tag{6.9}
$$

and

$$
\rho\left(\frac{\partial \lambda}{\partial t} + v_1 \frac{\partial \lambda}{\partial x_1}\right) = \frac{\partial}{\partial x_1}\left(d \frac{\partial \lambda}{\partial x_1}\right) + \rho\, \Omega\,. \tag{6.10}
$$

6.3. Shear motion. Now we turn to specialization of the equations to shear motion, which again leads to PDEs in one space dimension, transverse to the motion, and one time dimension. A nominal geometry is a slab of fixed thickness loaded on one surface with constant velocity while the other is fixed. The bottom surface is taken to be fixed (zero displacement) for the entire duration of the test. Specifically, we consider the following one-dimensional motion:

$$
x_1 = X_1 + f_1(X_2, t), \qquad x_2 = X_2 + f_2(X_2, t), \qquad x_3 = X_3\,, \tag{6.11}
$$

where f_1 and f_2 are the in-plane displacements, which can also be regarded as functions of the spatial coordinate and time x_2, t. Corresponding to this motion, one has the velocities with dependencies $v_1(x_2, t), v_2(x_2, t)$, and $v_3 = 0$ and $\partial/\partial x_1 = \partial/\partial x_3 = 0$. The expression of the material time derivative is given by $\dot{()} = \partial/\partial t + v_2\, \partial/\partial x_2$.

The shear deformation is described by

$$
\begin{aligned}
(\boldsymbol{F})_{ij} &= \frac{\partial x_i}{\partial X_j} = \begin{bmatrix} 1 & f_1' & 0 \\ 0 & 1 + f_2' & 0 \\ 0 & 0 & 1 \end{bmatrix}_{ij}, \\
(\boldsymbol{B})_{ij} &= (\boldsymbol{F}\boldsymbol{F}^T)_{ij} = \begin{bmatrix} 1 + {f_1'}^2 & f_1'(1 + f_2') & 0 \\ f_1'(1 + f_2') & (1 + f_2')^2 & 0 \\ 0 & 0 & 1 \end{bmatrix}_{ij},
\end{aligned} \tag{6.12}
$$

$$(\boldsymbol{L})_{ij} = (\vec{\nabla}\boldsymbol{v})_{ij} = \frac{\partial v_i}{\partial x_j} = \begin{bmatrix} 0 & \frac{\partial v_1}{\partial x_2} & 0 \\ 0 & \frac{\partial v_2}{\partial x_2} & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad (\boldsymbol{D})_{ij} = \begin{bmatrix} 0 & \frac{1}{2}\frac{\partial v_1}{\partial x_2} & 0 \\ \frac{1}{2}\frac{\partial v_1}{\partial x_2} & \frac{\partial v_2}{\partial x_2} & 0 \\ 0 & 0 & 0 \end{bmatrix}_{ij}. \tag{6.13}$$

The invariants of $\boldsymbol{B}$ are computed as $I_{\boldsymbol{B}} = 1+{f_1'}^2+(1+f_2')^2+1$ and $III_{\boldsymbol{B}} = (1+f_2')^2 = (\rho_0/\rho)^2$ with $1+f_2' = \rho_0/\rho$. Also $I_{\boldsymbol{B}} - 3 = (\rho_0/\rho)^2 - 1 + {f_1'}^2$. In addition, from the kinematic identity, $\dot{\boldsymbol{F}} = \boldsymbol{L}\boldsymbol{F}$, we obtain two nontrivial relations $\dot{f}_1' = (1+f_2')\partial v_1/\partial x_2$ and $\dot{f}_2' = (1+f_2')\partial v_2/\partial x_2$, where the material derivative is $\dot{\overline{()}} = \partial/\partial t + v_2\,\partial/\partial x_2$. The second of the two results just restates mass conservation and is equivalent to replacing $1+f_2'$ with ρ_0/ρ. But the first is an independent expression for the shear strain, which can be recast in terms of the density and transverse velocity gradient as

$$\dot{\overline{(f_1')}} = \left(\frac{\rho_0}{\rho}\right)\frac{\partial v_1}{\partial x_2}. \tag{6.14}$$

Finally, the contribution to the configurational stress has only one nonzero component, $(\vec{\nabla}\phi\otimes\vec{\nabla}\phi)_{22} = (\partial\phi/\partial x_2)^2$.

Using the density ρ and the shear strain f_1' as the two independent kinematic variables, we can now write down expressions for the components of the stress tensor. The cross-plane shear stresses are zero, i.e., $\sigma_{13} = \sigma_{23} = 0$. The in-plane shear stress σ_{12} is given by the expression

$$\sigma_{12} = \mu_s\, f_1' + \mu_f \frac{\partial v_1}{\partial x_2}. \tag{6.15}$$

The in-plane normal stress σ_{22} is given by

$$\sigma_{22} = -\mu_c\left(\frac{\rho}{\rho_0}\right)\left[\left(\frac{\rho}{\rho_0}\right)^{\frac{2\nu_s}{1-2\nu_s}} - \left(\frac{\rho}{\rho_0}\right)^{-2}\right] - \alpha_c K\frac{\rho}{\rho_0}(T-T_0) - \rho R_g T - \rho\gamma_\phi\left(\frac{\partial\phi}{\partial x_2}\right)^2 + (\nu_f + 2\mu_f)\frac{\partial v_2}{\partial x_2}. \tag{6.16}$$

The specific governing equations for the shear motion for the full model are

$$\frac{\partial\rho}{\partial t} + v_2\frac{\partial\rho}{\partial x_2} + \rho\frac{\partial v_2}{\partial x_2} = 0, \tag{6.17}$$

$$\rho\left(\frac{\partial v_1}{\partial t} + v_2\frac{\partial v_1}{\partial x_2}\right) = \frac{\partial}{\partial x_2}\left[\mu_s f_1' + \mu_f\frac{\partial v_1}{\partial x_2}\right], \tag{6.18}$$

$$\rho\left(\frac{\partial v_2}{\partial t} + v_2\frac{\partial v_2}{\partial x_2}\right) = \frac{\partial}{\partial x_2}\left\{-\mu_c\left(\frac{\rho}{\rho_0}\right)\left[\left(\frac{\rho}{\rho_0}\right)^{\frac{2\nu_s}{1-2\nu_s}} - \left(\frac{\rho}{\rho_0}\right)^{-2}\right] - \alpha_c(\phi)K\frac{\rho}{\rho_0}(T-T_0) - \rho R_g(\phi)T - \rho\gamma_\phi\left(\frac{\partial\phi}{\partial x_2}\right)^2 + (\nu_f+2\mu_f)\frac{\partial v_2}{\partial x_2}\right\}, \tag{6.19}$$

$$
\begin{aligned}
(6.20)\quad \rho c_v\left(\frac{\partial T}{\partial t}+v_2\frac{\partial T}{\partial x_2}\right) &= \frac{\partial}{\partial x_2}\left(k\frac{\partial T}{\partial x_2}\right)+\left[\mu_f\left(\frac{\partial v_1}{\partial x_2}\right)^2+(\nu_f+2\mu_f)\left(\frac{\partial v_2}{\partial x_2}\right)^2\right]\\
&\quad + B\dot{\phi}^2-\left[\alpha_c K\frac{\rho}{\rho_0}T+\rho R_g T\right]\frac{\partial v_2}{\partial x_2}\\
&\quad +\left\{-\frac{1}{2}\alpha_c'(\phi)K\frac{\rho}{\rho_0}T\ln(III_{\boldsymbol{B}})-\frac{1}{2}\rho R_g'(\phi)T\ln(III_{\boldsymbol{B}})-\rho c_v'(\phi)T\ln(T/T_0)\right.\\
&\quad \left.+\rho\left[\beta_m'(\phi)\frac{T}{T_m}Q_m+\beta_v'(\phi)\frac{T}{T_v}Q_v\right]\right\}\dot{\phi}+\rho\, Q_{hc}\Omega+\rho r\,,
\end{aligned}
$$

$$
\begin{aligned}
(6.21)\quad B\left(\frac{\partial \phi}{\partial t}+v_2\frac{\partial \phi}{\partial x_2}\right) &= \frac{\partial}{\partial x_2}\left(\rho\gamma_\phi\frac{\partial \phi}{\partial x_2}\right)+\rho c_v'(\phi)\left[T\ln(T/T_0)-(T-T_0)\right]\\
&\quad -\frac{\mu_s'(\phi)}{2}\frac{\rho}{\rho_0}(I_{\boldsymbol{B}}-3)-\frac{\mu_c'(\phi)}{2}\frac{\rho}{\rho_0}\frac{(1-2\nu_s)}{\nu_s}\left(III_{\boldsymbol{B}}^{-\nu_s/(1-2\nu_s)}-1\right)-\frac{3\,\mu_l'(\phi)}{2}\frac{\rho}{\rho_o}III_{\boldsymbol{B}}^{1/3}\\
&\quad +\frac{\alpha_c'(\phi)}{2}K\frac{\rho}{\rho_0}(T-T_0)\ln(III_{\boldsymbol{B}})+\frac{1}{2}\rho R_g'(\phi)T\ln(III_{\boldsymbol{B}})\\
&\quad -\frac{1}{2}\rho\Psi^{well}\frac{\partial \mathcal{F}}{\partial \phi}-\rho\left[\beta_m'(\phi)\frac{T-T_m}{T_m}Q_m+\beta_v'(\phi)\frac{T-T_v}{T_v}Q_v\right]
\end{aligned}
$$

and for chemical reaction are

$$
(6.22)\qquad \rho\left(\frac{\partial \lambda}{\partial t}+v_2\frac{\partial \lambda}{\partial x_2}\right)=\frac{\partial}{\partial x_2}\left(d\,\frac{\partial \lambda}{\partial x_2}\right)+\rho\Omega\,.
$$

Finally, the kinematic relation (6.14) for the shear strain (which must be included) is expressed as

$$
(6.23)\qquad \frac{\rho}{\rho_0}\left(\frac{\partial f_1'}{\partial t}+v_2\frac{\partial f_1'}{\partial x_2}\right)=\frac{\partial v_1}{\partial x_2}\,.
$$

7. Conclusions. We have posed a three-dimensional model for a representative energetic material with two independent state variables that represent the change in phase and the extent of exothermic reaction. The model has a context and formulation in which it is thermodynamically consistent. This is in contrast to other models which may not be self-consistent because the constitutive theory is invoked a posteriori. Gurtin's notion of a fundamental balance of configurational forces leads to evolution laws for the phase variable. Limiting forms of this model are consistent with classical theories, but the model also yields limiting forms that can describe the transition between two phases, if desired. The combined model is very rich in the sense that the coupling between phase evolution and the energy equations is complex, due in part to the necessary partition of the Helmholtz free energy.

In [13] we use experimental data based on the behavior and properties of HMX to study representative dynamics of the three simple motions discussed in section 6. The examples we develop show a variety of behavior observed over many time and length scales. Strain localization and phase transition phenomena are observed, as well as many other complex phenomena.

Acknowledgment. Discussions with E. Fried are warmly acknowledged.

REFERENCES

[1] J. C. Oxley, *The chemistry of explosives,* in Explosive Effects and Applications, J. A. Zuckas and W. P. Walters, eds., Springer, New York, 1997, pp. 137–172.

[2] W. Fickett and W. C. Davis, *Detonation*, University of California Press, Berkeley, CA, 1979.

[3] F. A. Williams, *Combustion Theory*, 2nd ed., Benjamin Cummings, Menlo Park, CA, 1985.

[4] T. L. Boggs, *The thermal behavior of (RDX) and (HMX),* in Fundamentals of Solid-Propellant Combustion, K. K. Kuo and M. Summerfeld, eds., Progress in Astronautics and Aeronautics 90, AIAA, 1984, Reston, VA, pp. 121–175.

[5] T. B. Brill, *Multiphase chemistry considerations at the surface of burning nitramine monopropellants,* J. Propulsion Power, 11 (1995), pp. 740–750.

[6] D. Chakraborty, R. P. Muller, S. Dasgupta, and W. A. Goddard, III, *The mechanism for unimolecular decomposition of RDX (*1, 3, 5-*trinitro*-1, 3, 5,-*triazine); an ab initio study,* J. Phys. Chem. A, 104 (2000), pp. 2261–2272.

[7] D. Bedrov, G. D. Smith, and T. D. Sewell, *Temperature-dependent shear viscosity coefficient of octahydro*-1, 3, 5, 7-*tetranitro*-1, 3, 5, 7-*tetrazocine (HMX): A molecular dynamics simulation study,* J. Chem. Phys., 112 (2000), pp. 7203–7208.

[8] G. Caginalp, *Stefan and Hele-Shaw type models as asymptotic limits of the phase-field equations,* Phys. Rev. A, 39 (1989), pp. 5887–5896.

[9] M. E. Gurtin, *Configurational Forces as Basic Concepts of Continuum Physics*, Appl. Math. Sci. 137, Springer, New York, 2000.

[10] J. D. Buckmaster and G. S. S. Ludford, *Theory of Laminar Flames,* Cambridge University Press, Cambridge-New York, 1982.

[11] I. Muller, *Thermodynamics*, Pitman, Boston, 1985.

[12] R. M. Bowen, *Part* I: *Theory of mixtures,* in Continuum Physics, Vol. III Mixtures and Electromagnetic Field Theories, A. C. Eringen, ed., Academic Press, New York, 1976, pp. 1–127.

[13] J. J.-I. Yoh, D. S. Stewart, and G. A. Ruderman, *A thermomechanic model for energetic materials with phase transformations: Analysis of simple motions,* SIAM J. Appl. Math., 63 (2002), pp. 538–563.

[14] A. A. Wheeler, W. J. Boettinger, and G. B. McFadden, *Phase-field model for isothermal phase transitions in binary alloys,* Phys. Rev. A., 45 (1992), pp. 7424–7439.

[15] A. D. Drozdov, *Finite Elasticity and Viscoelasticity,* World Scientific, Singapore, 1996.

SIAM J. APPL. MATH.
Vol. 63, No. 2, pp. 538–563

A THERMOMECHANICAL MODEL FOR ENERGETIC MATERIALS WITH PHASE TRANSFORMATIONS: ANALYSIS OF SIMPLE MOTIONS*

JACK JAI-ICK YOH†, D. SCOTT STEWART†, AND GREGORY A. RUDERMAN‡

Abstract. This paper examines the behavior of a thermomechanical model for energetic materials posed in the companion paper and specifically analyzes three simple motions: (i) constant volume evolution, (ii) one-dimensional, time-dependent longitudinal compression (expansion), and (iii) time-dependent shear. The model describes phase transitions from solid to liquid to gas and exothermic chemical reaction. Thermal and mechanical properties are matched to the explosive HMX in order to illustrate representative dynamics and transitions. Constant volume thermal explosion, shock melting, and shear localization are demonstrated.

Key words. combustion, phase transformations, energetic materials

AMS subject classifications. 74A50, 74F10, 74F25, 74A15, 80A22, 80A25

PII. S003613990139026X

1. Introduction.

This paper (paper II) is the second of two papers that describe a continuum model for the behavior of a condensed phase energetic material that undergoes phase transformation. Such materials are often used in explosive and pyrotechnic systems and are commonly known as solid explosives. Explosive materials are usually stable solids at room temperature and pressure, and when subjected to sufficiently strong mechanical or thermal stimulus, they undergo transitions to liquid and gas before releasing the bulk of their stored energy by chemical reaction mainly in the gas phase. Paper I [2] presented the continuum formulation that describes phase transitions from solid to liquid to gas. The model also includes energy-release due to chemical reaction. The state of the phase and the progress of the chemical reaction are represented by two thermodynamically independent variables, ϕ and λ. The phase variable ϕ takes on the value 0 for a pure solid, 1 for a pure liquid, and 2 for a pure gas. The progress of the (exothermic) chemical reaction is represented by λ, which ranges from 0 (unreacted) to 1 (completely reacted).

Most of paper I explains the model's formulation and assumptions and the restricted form of the constitutive theory based on standard arguments from the second law of thermodynamics. Following Gurtin's suggestion [1], configurational forces are assumed to be in global and local balance and further arguments lead to the derivation of an evolution law for ϕ, which is of the advection, diffusion, reaction type. Following combustion theory for a reactive mixture, an evolution law for the reaction progress variable λ is posited as a fundamental law.

Hence, our model is fully three-dimensional and is thermodynamically and tensorially consistent. Specialization of the model and limiting forms are examined in

*Received by the editors June 4, 2001; accepted for publication (in revised form) April 29, 2002; published electronically November 19, 2002. This work was supported by the U.S. Air Force. This work was carried out with resources from the U.S. Air Force Research Laboratory, Armament Directorate, Eglin AFB, Florida, F08630-95-004, F08630-00-1-0002 and the U.S. Air Force Office of Scientific Research, Physical Mathematics Directorate, F49620-96-1-0260.

http://www.siam.org/journals/siap/63-2/39026.html

†Department of Theoretical and Applied Mechanics, University of Illinois, Urbana-Champaign, Urbana IL 61801 (yoh1@llnl.gov, dss@uiuc.edu).

‡Air Force Research Laboratory, Edwards Air Force Base, Edwards AFB, CA 93542 (Gregory.Ruderman@edwards.af.mil).

paper I, and the equations for the special cases of constant volume evolution, one-dimensional, time-dependent longitudinal compression motion and time-dependent shear motion are obtained. Solutions to initial boundary-value problems for the equations for these simple motions illustrate the behavior of the model and reveal its properties. The results testify to the model's potential suitability for modeling complex phenomena that involve both phase transformation and chemical reaction in one combined framework. Material constants and properties of the energetic material (solid explosive) HMX are used to determine representative values for the model. These include properties such as (but not limited to) the elastic properties, viscosities, specific heats, gas constant, heats of melting (fusion), vaporization (condensation), and combustion (detonation).

In section 2 the equations for the three-dimensional model are given. In section 3 we discuss how we assigned the material properties of HMX to the model. In section 4 the special forms of the equations for the three simple motions are indicated (and the reader is referred to paper I). We also solve the case for constant volume evolution and discuss the properties of the underlying ODEs and their dynamics. To avoid unnecessary repetition of previously stated equations, we will refer to the equations as follows. For (6.1) of paper I, we will write (I 6.1). In section 5 the numerical methodology is given for longitudinal and shear motions. Section 6 presents representative numerical solutions for mechanically induced phase transformation and includes examples of interesting properties of the model such as shear localization and shock melting.

2. Mathematical formulations.

2.1. Kinematics and some definitions. The coordinates of position in the lab-frame are given by $\boldsymbol{x}$ and the initial position of the material particles is given by $\boldsymbol{X}$. The mapping of the deformations is given by $\boldsymbol{x} = \boldsymbol{x}(\boldsymbol{X}, t)$. The deformation gradient $\boldsymbol{F}$ is defined by the derivative $\boldsymbol{F} = \partial\boldsymbol{x}/\partial\boldsymbol{X}$. The left Cauchy–Green tensor $\boldsymbol{B} = \boldsymbol{F}\boldsymbol{F}^T$ is used to describe finite deformations. The velocity is defined by the time derivative of the particle trajectories $\boldsymbol{v} = (\partial\boldsymbol{x}/\partial t)_{\boldsymbol{X}}$. The velocity gradient is the gradient of the velocity field defined by the tensor $\boldsymbol{L} = \vec{\nabla}\boldsymbol{v}$. Let the dot notation, $\dot{\overline{()}}$, refer to the material derivative. The rate of stretching tensor is given by $\boldsymbol{D} = [\vec{\nabla}\boldsymbol{v} + (\vec{\nabla}\boldsymbol{v})^\top]/2$. The time derivative of the deformation gradient is $\dot{\boldsymbol{F}} = \boldsymbol{L}\boldsymbol{F}$. Consideration of conservation of mass relates the instantaneous density ρ to a reference (ambient) density of the solid, ρ_0, by $\det(\boldsymbol{F}) = \rho_0/\rho$ as well as $\det(\boldsymbol{B}) = (\rho_0/\rho)^2$.

2.2. General formulation. The derivation of the model and its constitutive specification was the principal subject of paper I [2]. The arguments in paper I developed a ϕ-dependent constitutive expression for the stress $\boldsymbol{\sigma}$ as

$$\begin{aligned}\boldsymbol{\sigma} = {} & \mu_s \frac{\rho}{\rho_0}\boldsymbol{B} - \mu_c \frac{\rho}{\rho_0} III_{\mathbf{B}}^{-[\nu_c/(1-2\nu_c)]}\boldsymbol{I} + \mu_l \frac{\rho}{\rho_0} III_{\mathbf{B}}^{1/3}\boldsymbol{I} \\ & - \alpha_c K \frac{\rho}{\rho_0}(T - T_0)\boldsymbol{I} - \rho R_g T \boldsymbol{I} \\ & - \rho\gamma_\phi \vec{\nabla}\phi \otimes \vec{\nabla}\phi + \nu_f(\vec{\nabla}\cdot\boldsymbol{v})\boldsymbol{I} + 2\mu_f \boldsymbol{D}\,. \end{aligned} \tag{2.1}$$

The material properties μ_c, μ_s, μ_l, α_c, R_g, γ_ϕ, ν_f, and μ_f are assumed to be functions of ϕ such that they are nonzero in the appropriate phase and are zero otherwise. The shear modulus μ_s is associated with a Blatz–Ko compressible solid and μ_l is associated with a liquid. The function μ_c represents the shear modulus of the condensed phase

such that $\mu_c = \mu_s + \mu_l$, with the properties that $\mu_c(0) = \mu_s(0) = \mu_{solid}$, $\mu_l(0) = 0$, $\mu_s(1) = 0$, $\mu_c(1) = \mu_l(1) = \mu_{liquid}$, and $\mu_c(2) = \mu_s(2) = \mu_l(2) = 0$. The function α_c is associated with a thermal expansion stress, R_g is associated with the ideal gas constant in a gas, ν_f and μ_f are associated with strain rate generated viscous stress, and γ_ϕ is associated with phase change induced stresses that act in regions with nonzero phase gradients. Derivatives of the ϕ-dependent functions appear as $\alpha'(\phi, T) = \partial\alpha/\partial\phi|_T$. The "$s$" subscript refers to the solid phase, the "l" subscript refers to the liquid phase, and the "c" subscript refers to the condensed phase. Similarly the "f" subscript refers to the fluid properties for both liquid and gas phases. If spelled out, the subscript "solid," "liquid," or "gas" refer to a constant material property. The various scalar material properties, such as c_v, γ_ϕ, could have explicit dependence on λ as well as ϕ and T. The governing equations for the model with reaction and phase change (without body forces) are

$$\dot{\rho} + \rho\vec{\nabla}\cdot\boldsymbol{v} = 0\,, \tag{2.2}$$

$$\rho\,\dot{\boldsymbol{v}} = \vec{\nabla}\cdot\boldsymbol{\sigma}\,, \tag{2.3}$$

$$\begin{aligned}
\rho c_v\dot{T} = {}& \vec{\nabla}\cdot(k\vec{\nabla}T) + \nu_f(\vec{\nabla}\cdot\boldsymbol{v})^2 + 2\mu_f\boldsymbol{D}:\boldsymbol{D} + B\dot{\phi}^2 \\
& - \alpha_c K\frac{\rho}{\rho_0}T(\vec{\nabla}\cdot\boldsymbol{v}) - \rho R_g T(\vec{\nabla}\cdot\boldsymbol{v}) \\
& + \left\{-\frac{\alpha_c'(\phi)}{2}K\frac{\rho}{\rho_0}T\ln(III_{\mathbf{B}}) - \rho\frac{R_g'(\phi)}{2}T\ln(III_{\mathbf{B}}) - \rho c_v'(\phi)T\ln(T/T_0)\right. \\
& \left. + \rho\left[\beta_m'(\phi)\frac{T}{T_m}Q_m + \beta_v'(\phi)\frac{T}{T_v}Q_v\right]\right\}\dot{\phi} + \rho Q_{hc}\,\Omega + \rho r\,,
\end{aligned} \tag{2.4}$$

$$\begin{aligned}
B\dot{\phi} = {}& \vec{\nabla}\cdot\left(\rho\gamma_\phi\vec{\nabla}\phi\right) + \rho c_v'(\phi)\left[T\ln(T/T_0) - (T - T_0)\right] \\
& - \frac{\mu_s'(\phi)}{2}\frac{\rho}{\rho_0}(I_{\mathbf{B}} - 3) - \frac{\mu_c'(\phi)}{2}\frac{\rho}{\rho_0}\frac{(1-2\nu_c)}{\nu_c}\left(III_{\mathbf{B}}^{-\nu_c/(1-2\nu_c)} - 1\right) \\
& + \frac{3\mu_l'(\phi)}{2}\frac{\rho}{\rho_0}III_{\mathbf{B}}^{1/3} \\
& + \frac{\alpha_c'(\phi)}{2}K\frac{\rho}{\rho_0}(T - T_0)\ln(III_{\mathbf{B}}) + \frac{1}{2}\rho R_g'(\phi)T\ln(III_{\mathbf{B}}) \\
& - \rho\frac{1}{2}\Psi^{well}\frac{\partial F}{\partial\phi} - \rho\left[\beta_m'(\phi)\left(\frac{T}{T_m} - 1\right)Q_m + \beta_v'(\phi)\left(\frac{T}{T_v} - 1\right)Q_v\right],
\end{aligned} \tag{2.5}$$

$$\rho\dot{\lambda} = \vec{\nabla}\cdot(d\vec{\nabla}\lambda) + \rho\,\Omega\,, \tag{2.6}$$

$$\dot{\boldsymbol{F}} = \boldsymbol{L}\boldsymbol{F}\,. \tag{2.7}$$

Table 1 gives the values for material properties that appear in (2.1)–(2.7). The values are based upon HMX along with references to the data source or a notation that we have used a model value.

2.3. Material transition functions. An important ingredient of our model is the use of ϕ-dependent material properties, or material transition functions. Their most prominent use is in the definition of the source terms in the ϕ-evolution equation and the energy (temperature) equation. Also, functions such as $\mu_c(\phi)$, $\mu_s(\phi)$, $\mu_l(\phi)$, $\alpha_c(\phi)$, $R_g(\phi)$, $c_v(\phi)$ all change with the phase variable ϕ. The model assumes that these are defined in such a way so that they take on proper values for pure phases when the material has the limiting pure-phase values for ϕ. The ϕ-dependent material

TABLE 1
Material properties typical of HMX.

Material property	Value	Refs.
density of β-HMX (ρ_0)	1.71 g/cm^3	[7]
density of liquid-HMX	1.65 g/cm^3	[13]
specific heat at constant volume (c_v)	1.5×10^3 J/kg K	[7]
isothermal bulk modulus ($K = \rho \left.\frac{\partial p}{\partial \rho}\right\|_T$)	13.5 GPa	[7]
shear modulus (μ_{solid})	2.46 GPa	modeled
shear modulus (μ_{liquid})	2.37 GPa	modeled
Poisson's ratio (ν_c)	0.414	calculated
viscosity coefficient (μ_f)	0.45 N s/m^2	[5]
bulk viscosity coefficient (ν_f)	-2/3 μ_f	Stokes hypothesis
thermal expansion coefficient (α_{solid})	0.000134 1/K [7]	
thermal conductivity (k)	.36 W/m K	[7]
phase diffusion coefficient ($\rho\gamma_\phi$)	1.0×10^{-6} m kg/s^2	modeled
universal gas constant (R_u)	8313 J/kmole K	
molar weight of β-HMX	296.2 kg/kmole	[5]
gas constant per unit mass (R_{gas})	300 J/kg K	modeled
melting temperature (T_m)	558 K	[5]
vaporization temperature (T_v)	588 K	modeled
heat of melting (Q_m)	-200×10^3 J/kg	[5]
heat of vaporization (Q_v)	-100×10^3 J/kg	modeled
heat of combustion (Q_{hc})	5.0×10^6 J/kg	[8]
rate of heat source (ρr)	5000 J/m^{3}s	modeled
frequency of Arrhenius kinetic (A)	9.3×10^{16} 1/s	[8]
activation temperature (E_a/R_u)	24660 K	[8]
depth of phase well (Ψ^{well})	550 J/kg	modeled
multiplication factor of $\dot{\phi}$ (B)	1.5 kg/m s	modeled

transition functions used for this paper are listed in the appendix and are made up of simple, smooth, or piece-wise smooth polynomials in ϕ.

Figure 1 shows the material transition functions $\beta'_m(\phi)$, $\beta'_v(\phi)$ that are used to construct ψ_2, which represents free-energy changes during phase transition and is given by

$$\psi_2 = \frac{1}{2}\Psi^{well}F(\phi) + \beta_m(\phi)Q_m\Big(\frac{T}{T_m} - 1\Big) + \beta_v(\phi)Q_v\Big(\frac{T}{T_v} - 1\Big), \tag{2.8}$$

where $\Psi^{well} > 0$ is a constant that describes the potential and $Q_m < 0$ and $Q_v < 0$ are constants representing the heats of melting and vaporization. The constants $T_m > 0$ and $T_v > 0$ are melting and vaporization temperatures. Here we assume a specific form for $F(\phi)$ (listed in the appendix) that is a smooth positive definite function with isolated zeros at $\phi = 0, 1$, and 2 representing three local minima. In addition, $F(\phi)$ is assumed to be locally quadratic near the zeros at $\phi = 0, 1$ and 2, i.e., near $\phi = 0$, $F \sim \phi^2$, near $\phi = 1$, $F \sim (\phi - 1)^2$, and near $\phi = 2$, $F \sim (\phi - 2)^2$. The function $\beta_m(\phi)$ is assumed to be smooth and monotonically increasing and has values from 0 to 1 on the range $0 \le \phi \le 1$ with zero derivative elsewhere. Similarly, the function $\beta_v(\phi)$ is assumed to be monotonically increasing with values from 0 to 1 on the range $1 \le \phi \le 2$. The derivative of transition energy density $\partial\psi_2/\partial\phi$ generates source terms

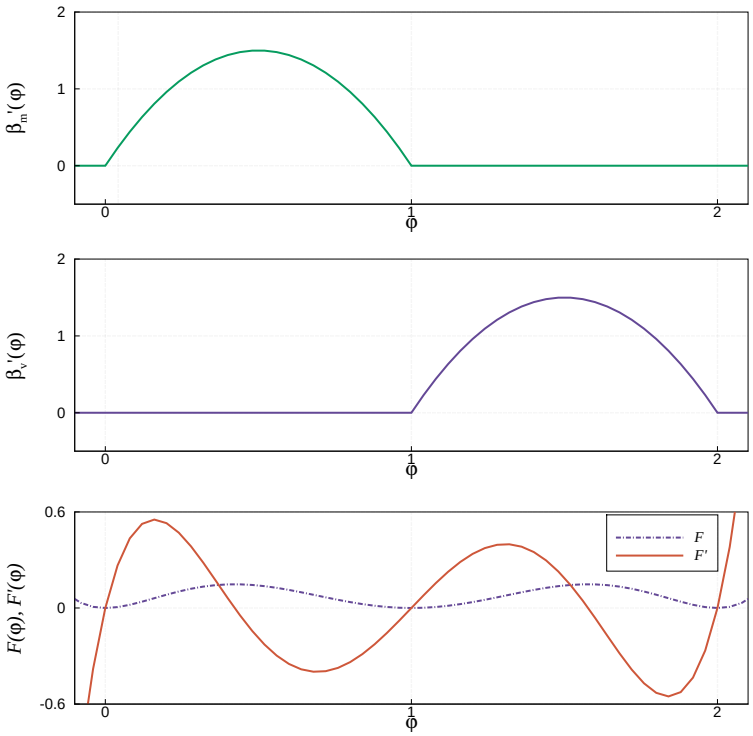


FIG. 1. *Material transition functions β'_m and β'_v and Ginzburg–Landau potential F used to prescribe the phase transition energy density ψ_2.*

in the phase equation represented as

$$\frac{\partial \psi_2}{\partial \phi} = \frac{1}{2}\Psi^{well}\frac{\partial F}{\partial \phi}(\phi) + \beta'_m(\phi)Q_m\Big(\frac{T}{T_m} - 1\Big) + \beta'_v(\phi)Q_v\Big(\frac{T}{T_v} - 1\Big). \tag{2.9}$$

Figure 2 illustrates the assumed dependence of $\psi_2(\phi, T)$ on ϕ and T. Starting from (a) through (d), the temperature T is raised from below T_m to above T_v, representing a standard melting-evaporation process. The transition energy density in (a) has its minimum at $\phi = 0$. As T is increased through T_m and then T_v, we see a shift in the global minima from pure solid to solid-liquid and to liquid-vapor. As T eventually exceeds T_v as shown in (d), the energy minimizing well shifts to a vapor state at $\phi = 2$. Figure 3 shows examples of the other material transition functions and their derivatives.

3. Matching material properties to HMX. Here we discuss our fit of the model's material properties to mimic an energetic material like HMX. Figure 4 shows a pressure-temperature plane that indicates regions where, from classical and experimental considerations, HMX can be considered to be a static solid, liquid, or a gas in thermodynamic equilibrium. Some of the boundaries (specifically the solid/liquid boundary) are known from experiment. Note that solid phases of HMX are not differentiated here, and it is assumed that the β-phase of solid HMX is representative. The solid/liquid boundary is of particular interest and is computed via a Kraut–Kennedy law. It is well known that HMX liquid is quite unstable [3] and once the liquid phase

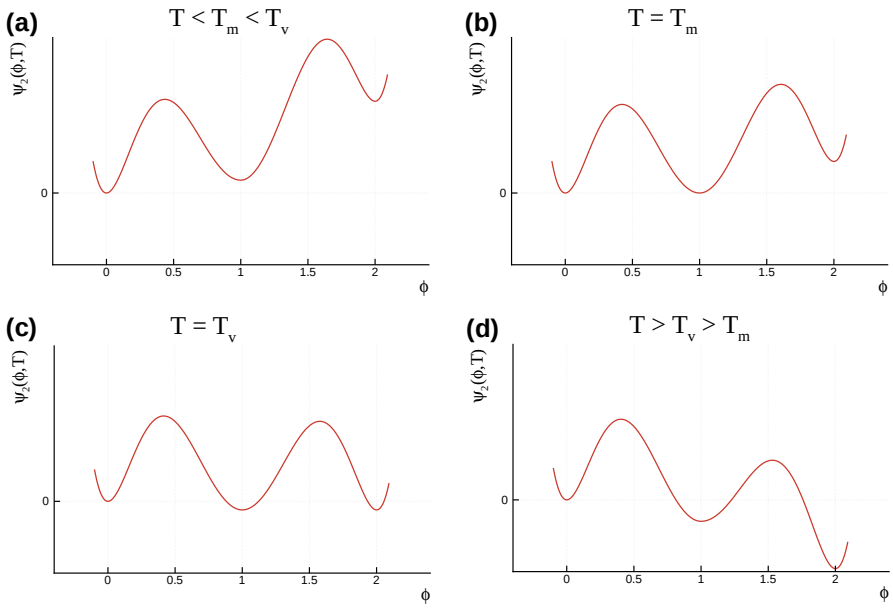


FIG. 2. *Plot of ψ_2 as a function of ϕ with T variation.*

appears it quickly evolves into gas, partly from exothermic energy released by chemical reaction in the condensed phase. For the purpose of our early modeling efforts, we have decided to match the HMX melt temperature to the experimental melt temperature $T = 558$ K and the evaporation temperature at $T = 588$ K. Figure 4 shows a shaded box that represents the range of temperatures and pressures (level of stress) predicted by computation with model.

In our model, p-V isotherms (where $V = 1/\rho$) can be obtained directly from (2.1) by setting all derivatives equal to zero and by assuming a homogeneous deformation such that $\boldsymbol{B} = (\rho_0/\rho)^{2/3}\boldsymbol{I}$ and $\boldsymbol{\sigma} = -p\boldsymbol{I}$ (where p is the pressure) to obtain

$$p = \mu_c \left(\frac{\rho}{\rho_0}\right)\left[\left(\frac{\rho}{\rho_0}\right)^{\frac{2\nu_c}{1-2\nu_c}} - \left(\frac{\rho}{\rho_0}\right)^{-\frac{2}{3}}\right] + \alpha_c K \frac{\rho}{\rho_0}(T - T_0) - \rho R_g T\,. \tag{3.1}$$

HMX liquid is approximately 4% less dense than HMX solid [13]. Figure 5 shows a plot of an isotherm computed from (3.1) with values shown in Table 1. Experimental data points on the solid isotherms obtained by Yoo and Cynn [10] are shown for comparison. Since HMX liquid is so chemically unstable, experimental data for the liquid isotherm is not available. One implication of the lower density for HMX liquid is that the isothermal sound speed (the negative slope of the p-V isotherm) is greater in the solid than in the liquid. Figure 6 shows a plot of an isotherm computed from (3.1) for the ideal gas term that is proportional to R_g. Figure 7 shows a representative isotherm on log-scales at 300 K, for the full range of values for the model when the material is solid, liquid, or gas, as computed from (3.1).

4. Form of the model for three simple motions. Here we consider the differential equations for the model when the material undergoes three simple motions: (i) evolution at constant volume, (ii) time-dependent longitudinal motion, and (iii)

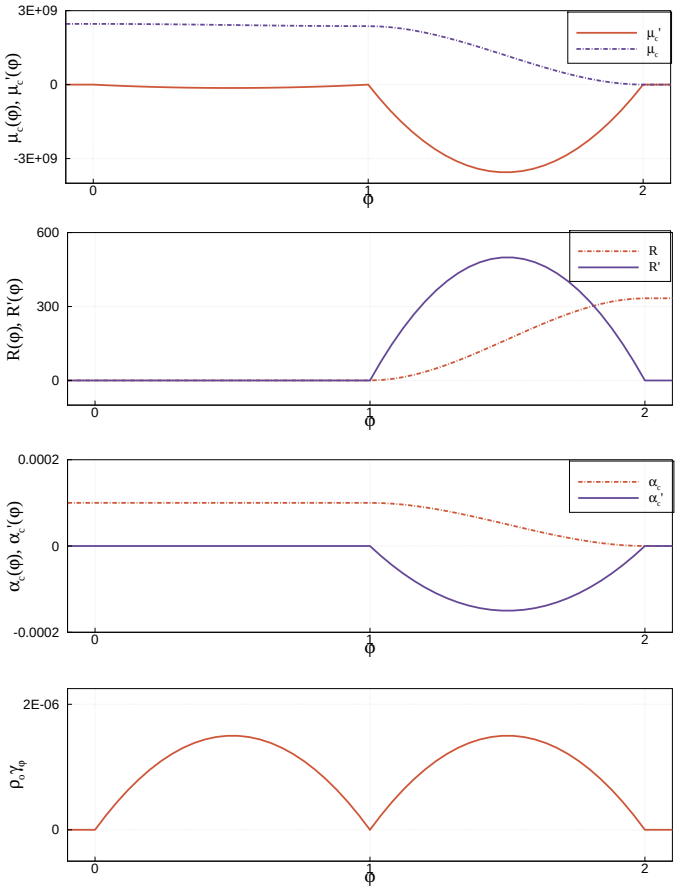


FIG. 3. *Plots of transition functions and their derivatives for the HMX simulant derivatives with respect to the phase variable. Shear modulus, ideal gas constant, thermal expansion coefficient, and phase diffusion coefficient are shown from top to bottom.*

one-dimensional, time-dependent shear motion. All three cases are amenable to extensive computational and theoretical analysis and their discussion reveals the underlying mathematical properties of the model. All three are very important in the traditional analysis of ignition of energetic materials. The reduction for the three special motions follow directly from the general form of equations (2.1)–(2.7) and were derived in the last section of paper I.

4.1. Evolution at constant volume. An important simple case often considered in combustion theory describes constant volume thermal explosion, where the velocity $\boldsymbol{v}$ as well as all spatial gradients are exactly zero and the density is constant. For illustration (in this section only), we neglect thermal expansion and assume a constant specific heat and gas constant. We are left with three ODEs in time for the temperature T, phase variable ϕ, and reaction progress variable λ (see (I 6.1), (I 6.2), and (I 6.3)). If phase change is discarded, we recover the equations from standard combustion theory for constant volume thermal explosion, $c_v(\partial T/\partial t) = Q_{hc}\,\Omega\,, (\partial\lambda/\partial t) = \Omega$. Of course, the more interesting behavior occurs when phase change is included.

Figure 8 shows an example of the time evolution of constant volume heating

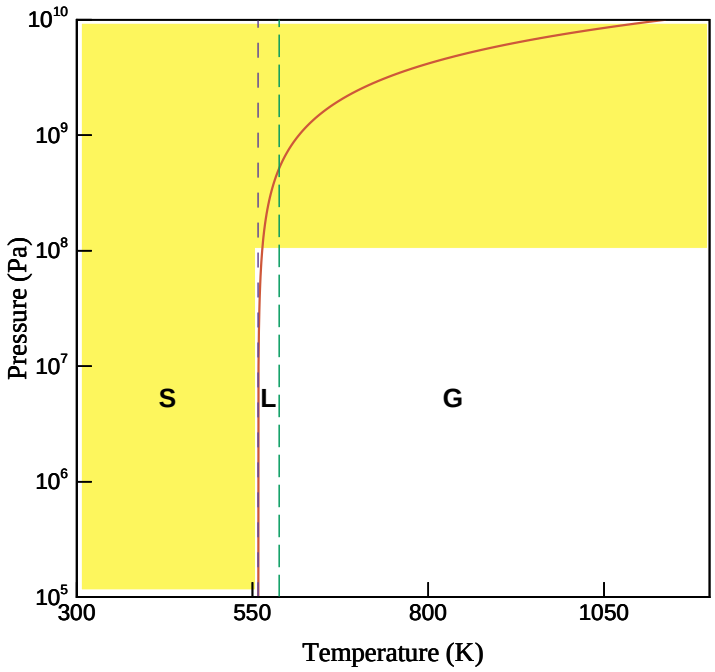


FIG. 4. *The solid curve is the melt temperature-pressure relation for β-HMX given by the semi-empirical Kraut–Kennedy law* [4], [7]. *Dashed line and long-dashed line are constant melt temperature and vapor temperature used in the current numerical simulation, respectively.*

without chemical reaction starting from a solid ($\phi = 0$) at an initial temperature of $T = 300$ K. The heating rate r and the kinetic parameter B control the transformation rates. It is possible to see (at least qualitatively) all of the phase change behaviors expected during constant volume heating. As heat is first applied, the temperature increases linearly. As the temperature increases further, the material begins to melt and the endothermic process absorbs heat from the system. (In Figure 8 the slight temperature decrease is barely visible in the nearly constant temperature interval.) At the completion of the phase transformation to liquid, the temperature rises in the liquid at a constant rate until the vaporization temperature is reached and second phase transition from liquid to gas phase occurs. After that, constant rate heating in the gas occurs.

The sharp transitions that are apparent in Figure 8 (for example, near the times $t = 0.05$ and $t = 0.09$ sec) are the result of a bifurcation and a change of stability in the ODEs near the transition temperatures T_m and T_v. To see this clearly, consider the stability of the solid phase during constant rate heating. The temperature and the phase variable can be represented to leading order as $T = T_0 + r\,t/c_v$ and $\phi = 0$ (with $\lambda = 0$ for all time), so that a stability analysis assumes that T and ϕ take the form

$$T = T_0 + \frac{r}{c_v}t + T'(t) + \cdots, \qquad \phi = \phi'(t) + \cdots, \tag{4.1}$$

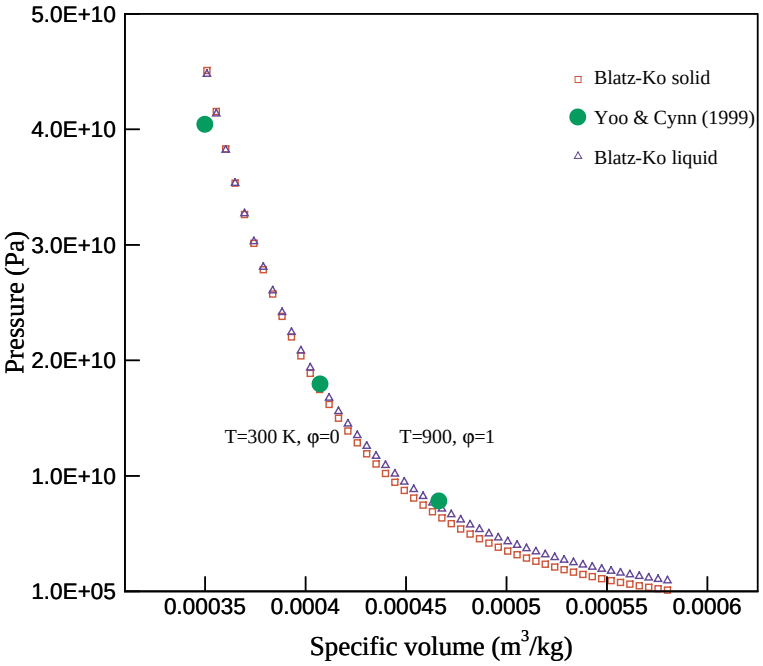


FIG. 5. *P-V isotherms at $T = 300\ K$ and $T = 900\ K$. $\left|\frac{dp}{dV}\right|_{\phi=0} > \left|\frac{dp}{dV}\right|_{\phi=1}$ implies that the speed of sound is greater in solid than in the liquid HMX.*

where T' and ϕ' are assumed to be small. The linearization of (I 6.1), (I 6.2) with $\beta'_m \approx 6\phi'$ and $\partial F/\partial\phi \approx 8\phi'$ is straightforward and leads to equations for T', ϕ':

$$\frac{dT'}{dt} = 0\,, \tag{4.2}$$

$$\frac{d\phi'}{dt} = -\frac{\rho}{B}\left\{4\Psi^{well} - 6Q_m\left(1 - \frac{T^{(0)}}{T_m}\right)\right\}\phi'\,, \tag{4.3}$$

where $T^{(0)}$ is the leading-order temperature found from simple constant rate heating

$$T^{(0)} = T_0 + \frac{r}{c_v}t\,. \tag{4.4}$$

The stability properties of the solution for ϕ' are governed by the sign of $d\phi'/dt$ found on the right-hand side of (4.3). For early times, the argument is always negative, since $T^{(0)} < T_m$ and $Q_m < 0$. Consequently, the solution is exponentially stable. (It is a simple matter to write down the exact solution of (4.3).) As the temperature rises, the stability changes as $d\phi'/dt$ changes sign, and this time it is found by setting the right-hand side of (4.3) exactly equal to zero. In this case the leading-order temperature is

$$T^{(0)} = T_m - \frac{2}{3}\frac{\Psi^{well}}{Q_m}T_m\,. \tag{4.5}$$

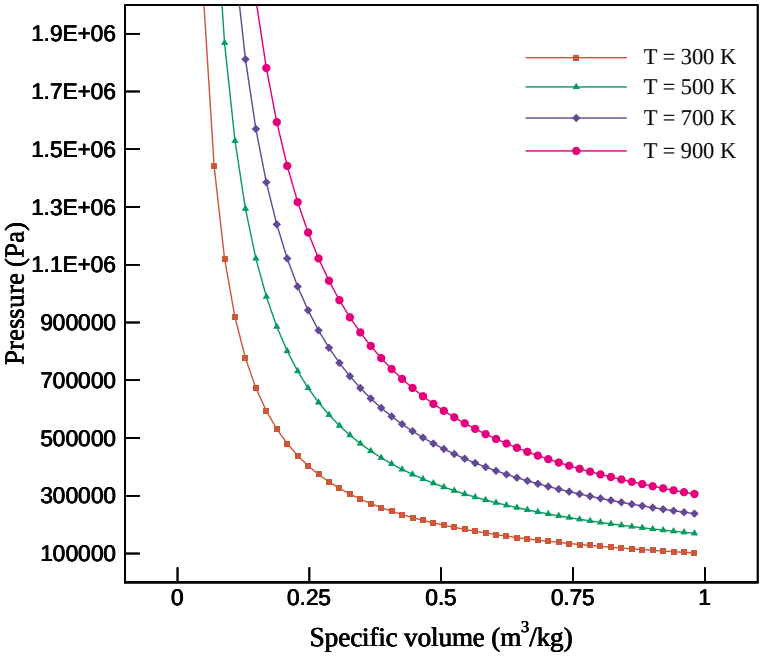


FIG. 6. *P-V isotherms at four different temperatures for HMX vapor ($\phi = 2$).*

For the case where $|\Psi^{well}/Q_m| << 1$, the phase transition temperature associated with this simple change of stability is close to T_m. So we find that below the melt temperature the perturbations are stable, but near the melt temperature the stability changes type and becomes unstable. Any perturbation grows and subsequently an abrupt transition occurs from $\phi = 0$. Another point is that our assumed properties for F strictly require a nonzero perturbation of ϕ to be combined with heating in order to observe a phase transition. In cases other than constant volume evolution, other source terms exist in the ϕ-evolution equation (specifically those related to derivatives of the Helmholtz free energy associated with deformation) and they can be the source of thermomechanical disturbances that can grow when the phase becomes dynamically unstable.

Figure 9 shows a representative solution to (I 6.1)–(I 6.3) with an Arrhenius form assumed for $\Omega = A(\phi)(1-\lambda)\exp[-E_a/(R_u T)]$. The function $A(\phi)$ is chosen to be zero in the solid phase and takes the value listed in Table 1 in the gas phase. Initially the material is solid and cold and heated at a uniform rate. So the phase transformations from solid to liquid to gas occur in the same way as shown in Figure 8. However, for this example, once the gas is in abundance, the chemical reaction starts and the gas undergoes a classically well understood constant volume thermal explosion. If we had chosen to adopt a more complex kinetic form for Ω, reaction could take place first in the liquid phase. In the near future, we plan to use more realistic kinetic scheme for HMX. Clearly there is the flexibility within this formulation to model many aspects of condensed phase energy release.

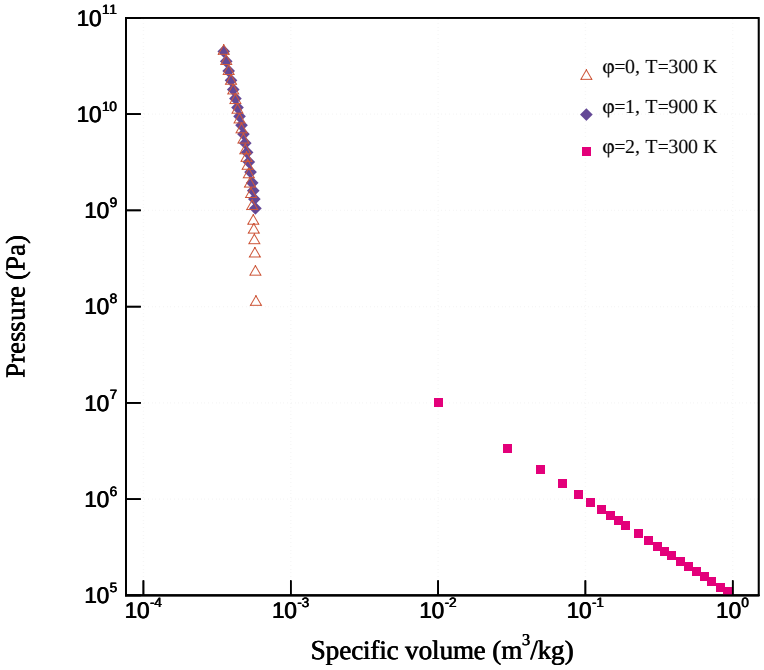


FIG. 7. *P-V isotherms for solid, liquid, and vapor HMX at $T = 300$ K, drawn to a single range of P-V axes.*

4.2. Longitudinal motion. Next we consider simple longitudinal motion. Typically, explosives are tested by subjecting them to impact with a flyer-plate. In an idealization of this experiment, an infinite slab experiences a displacement loading normal to its surface. As a computational matter, the same flow can be modeled as a reverse impact experiment, where the sample is set into uniform motion and suddenly comes to rest at the origin. We must consider following one-dimensional motion:

$$x_1 = X_1 + f_1(X_1, t)\,, \quad x_2 = X_2\,, \quad x_3 = X_3\,, \tag{4.6}$$

where f_1 is the displacement in the 1-direction. There is one nonzero velocity component, $v_1 = \partial f_1/\partial t|_{\boldsymbol{X}}(X_1, t)$, and $\boldsymbol{F}$ and $\boldsymbol{B}$ are both diagonal, and $B_{11} = (1 + f_1')^2$, $B_{22} = 1, B_{33} = 1$. The first and third invariants of $\boldsymbol{B}$ are $I_{\boldsymbol{B}} = 2 + (1 + f_1')^2$ and $III_{\boldsymbol{B}} = (1 + f_1')^2 = (\rho_0/\rho)^2$. Then $I_{\mathbf{B}} - 3 = (\rho_0/\rho)^2 - 1$. Hence we use the density as the independent strain measure and replace f_1'. The one nonzero component of the rate of strain tensor is $D_{11} = \partial v_1/\partial x_1$. Also $(\vec{\nabla}\phi \otimes \vec{\nabla}\phi)_{11} = (\partial\phi/\partial x_1)^2$. It follows that all the shear stresses are zero, $\sigma_{12} = \sigma_{23} = \sigma_{13} = 0$, and the normal stress σ_{11}, is given by (I 6.5).

The governing equations for longitudinal compression are the mass and momentum equations in (I 6.6), (I 6.7), the energy equation (written in temperature form) in (I 6.8), and the phase and reaction progress evolution equations, given by (I 6.9) and (I 6.10). This special formulation is a set of five PDEs for the dependent variables ρ, v_1, T, ϕ, and λ in terms of the independent variables x_1 and t.

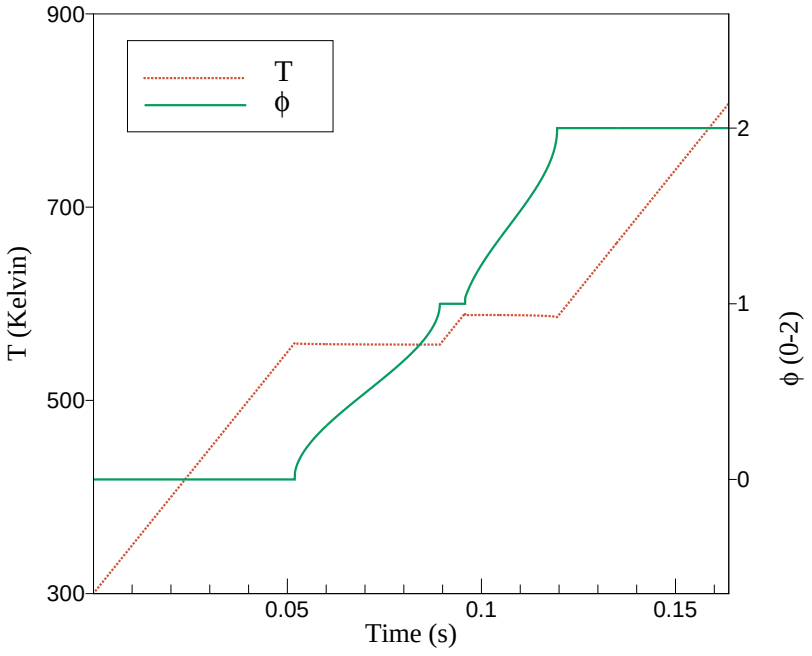


FIG. 8. *Constant volume phase transformation without reaction.*

4.3. Shear motion. The specialization of the equations to shear motion leads to PDEs in one space variable and time. The nominal geometry is a slab of fixed thickness loaded at one surface with constant velocity. The other surface is taken to be fixed (zero displacement) for the entire duration of the test. We consider the following shear motion:

$$x_1 = X_1 + f_1(X_2, t), \qquad x_2 = X_2 + f_2(X_2, t), \qquad x_3 = X_3\,, \tag{4.7}$$

where f_1 and f_2 are the in-plane displacements, which can also be regarded as functions of the spatial coordinate and time x_2, t. Corresponding to this motion one has the velocities with $v_1(x_2,t), v_2(x_2,t)$, $v_3 = 0$, and $\partial/\partial x_1 = \partial/\partial x_3 = 0$. The expression of the material time derivative is given by $\dot{\overline{()}} = \partial/\partial t + v_2\,\partial/\partial x_2$. The shear deformation is described by (I 6.12) and (I 6.13). The invariants of $\boldsymbol{B}$ are computed as $I_{\mathbf{B}} = 1 + {f_1'}^2 + (1+f_2')^2 + 1$ and $III_{\mathbf{B}} = (1+f_2')^2 = (\rho_0/\rho)^2$ with $1 + f_2' = \rho_0/\rho$. Also $I_{\mathbf{B}} - 3 = (\rho_0/\rho)^2 - 1 + {f_1'}^2$. In addition, from the kinematic identity $\dot{\boldsymbol{F}} = \boldsymbol{L}\boldsymbol{F}$, we obtain two nontrivial relations $\dot{f}_1' = (1+f_2')\partial v_1/\partial x_2$ and $\dot{f}_2' = (1+f_2')\partial v_2/\partial x_2$, where the material derivative is $\dot{\overline{()}} = \partial/\partial t + v_2\,\partial/\partial x_2$. The second of the two results is equivalent to replacing $1 + f_2'$ with ρ_0/ρ. The first is an independent expression for the shear strain which can be recast in terms of the density and transverse velocity gradient as in (I 6.14). Finally, $\vec{\nabla}\phi \otimes \vec{\nabla}\phi$ has only one nonzero component, $(\vec{\nabla}\phi \otimes \vec{\nabla}\phi)_{22} = (\partial\phi/\partial x_2)^2$.

We use the density ρ and the shear strain f_1' as the two independent kinematic

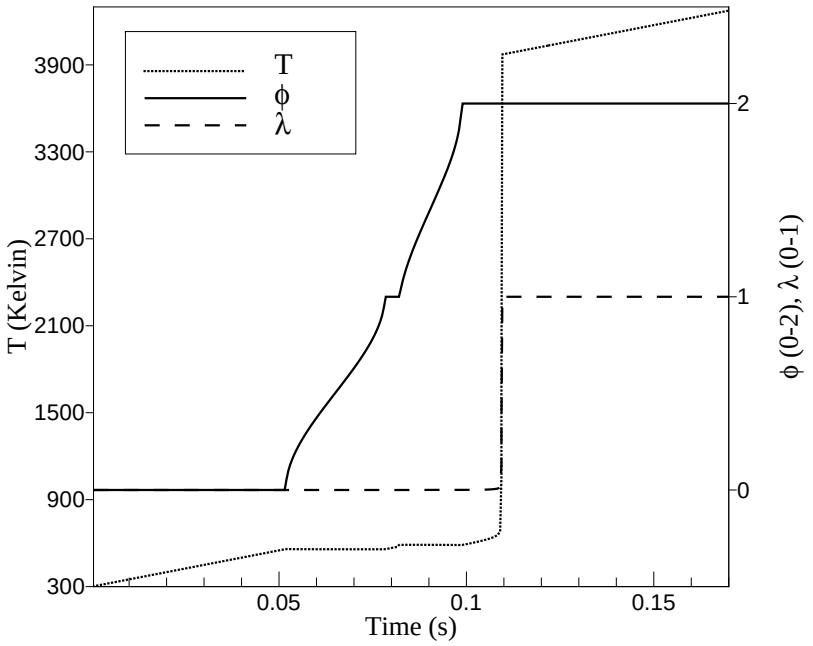


FIG. 9. *Constant volume thermal explosion.*

variables. We can now write down expressions for the components of the stress tensor. The cross-plane shear stresses are zero, i.e., $\sigma_{13} = \sigma_{23} = 0$. The in-plane shear stress σ_{12} is given by the expression (I 6.15). The in-plane normal stress σ_{22} is given by (I 6.16).

The specific governing equations for the shear motion are listed in (I 6.17)–(I 6.22). Finally the kinematic relation (I 6.14) for the shear strain (which must be included) is expressed as in (I 6.23). This special formulation is a set of seven PDEs for the dependent variables $\rho, v_1, v_2, T, \phi, \lambda$, and f_1' in terms of the independent variables x_2 and t.

5. Numerical methodology. We have implemented an efficient high-order temporal scheme for stiff equations based on the method of lines (MOL) to solve for longitudinal and shear motions. The MOL can be implemented for various choices of spatial discretization. For discretization of convective terms we use a fourth-order convex essentially nonoscillatory (ENO) method [6] combined with a third-order, low-storage, semi-implicit Runge–Kutta method [9] for the MOL-ODEs. We do not describe the ENO discretization here. Interested readers are referred to [6].

5.1. Description of low-storage semi-implicit Runge–Kutta solver. A more comprehensive discussion of the temporal scheme can be found in [9], and only a brief description of the method is given below. To solve a system of autonomous ODEs of the form $\boldsymbol{u}' = \boldsymbol{f}(\boldsymbol{u}) + \boldsymbol{g}(\boldsymbol{u})$, we use an explicit scheme for the nonstiff term $\boldsymbol{f}$ and an implicit scheme for the stiff term $\boldsymbol{g}$. We solve the system in an explicit/implicit

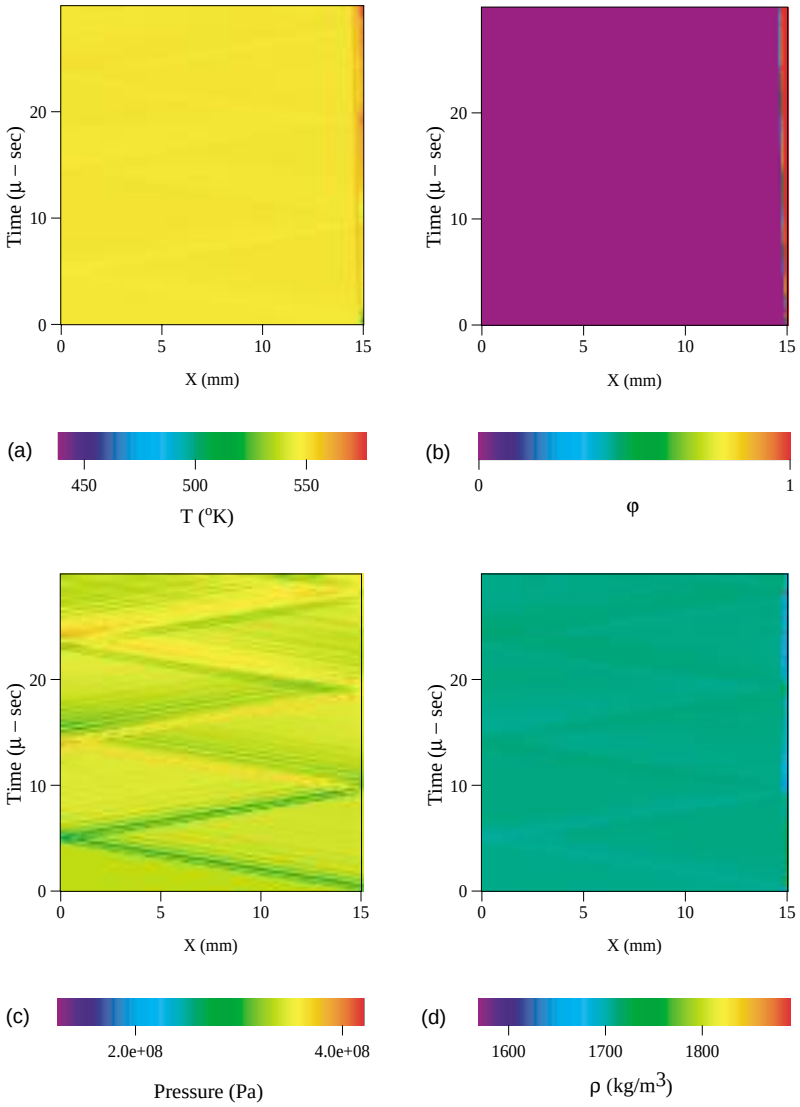


FIG. 10. *Temperature, phase, pressure, and density fields for a representative shear experiment* ($v_{shear} = 600\ m/s$, $T_0 = 550\ K$).

hybrid fashion to achieve high-order accuracy and stiffly stable calculation. A typical third-order method of this kind is given below:

$$\begin{aligned} \boldsymbol{k}_j &= a_j \boldsymbol{k}_{j-1} + h[\boldsymbol{f}(\boldsymbol{u}_{j-1}) + \boldsymbol{g}(\boldsymbol{u}_{j-1} + \bar{c}_j \boldsymbol{k}_{j-1} + c_j \boldsymbol{k}_j)], \\ \boldsymbol{u}_j &= \boldsymbol{u}_{j-1} + b_j \boldsymbol{k}_j \\ &\quad (j = 1, \ldots, 3), \end{aligned} \tag{5.1}$$

where h is the time increment, and the coefficients of the scheme are as follows:

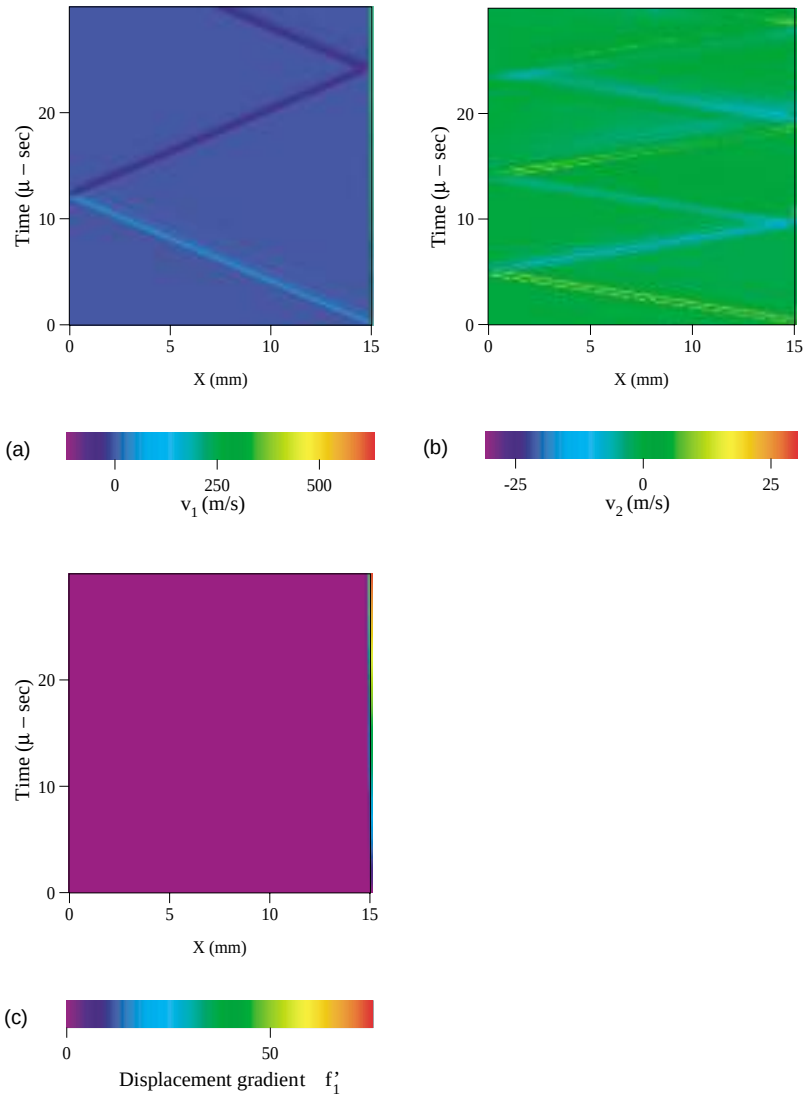


FIG. 11. *Shear velocity (v_1), compression velocity (v_2), and displacement gradient (du_1/dX_2) fields for a representative shear experiment ($v_{shear} = 600$ m/s, $T_0 = 550$ K).*

$$b_1 = \frac{1}{3}, \quad b_2 = \frac{15}{16}, \quad b_3 = \frac{8}{15}, \quad a_2 = -\frac{5}{9}, \quad a_3 = -\frac{153}{128},$$
$$c_1 = \frac{1}{5}, \quad c_2 = \frac{49}{75}, \quad c_3 = \frac{143}{600}, \quad \bar{c}_2 = -\frac{59}{135}, \quad \bar{c}_3 = -\frac{5283}{25600} \tag{5.2}$$

with $a_1 = 0, \bar{c}_1 = 0$.

In many instances where implicit calculation is not required, one can simply assign zero to the stiff vector $\boldsymbol{g}$ and assign the entire source as a nonstiff vector $\boldsymbol{f}$ and the standard explicit Runge–Kutta scheme is recovered.

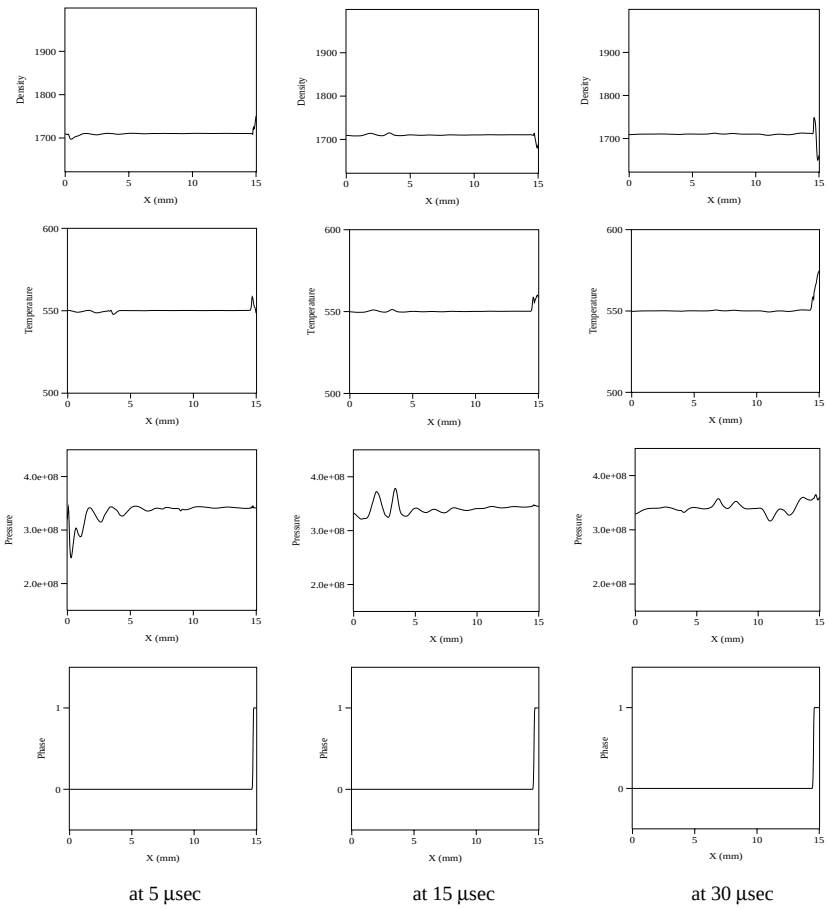


FIG. 12. *Snapshots of density, temperature, pressure, and phase field (from top to bottom) taken at time* $t = 5,\ 15,\ 30\,\mu sec$ *from Figures* 10 *and* 11 *of the plane shearing experiment.*

5.2. Implementation. Before starting the computation one writes the governing PDEs in a conservative form such that limiting forms of the equations admit discontinuous solutions which are also admitted by the numerical approximation. Further, the stiff and nonstiff terms must be intelligently separated. In particular, convective terms, which are a priori discretized in space via a fourth-order convex ENO scheme, are always treated as nonstiff terms. The viscous stress terms of momentum equations are treated as nonstiff and are discretized by a fourth-order central differencing. Only the reaction source term, Ω, is treated as stiff and is subjected to the implicit numerical procedure. Otherwise the explicit method solves all the remaining terms of the equations.

We consider the shear motion to illustrate the numerical implementation. After converting the equations into a conservative form and separating the stiff and nonstiff terms, we can write the conservative variables and the fluxes as follows:

$$\text{(5.3)}\quad \boldsymbol{u} = \begin{bmatrix} \rho \\ \rho v_1 \\ \rho v_2 \\ \rho c_v T \\ \rho\phi \\ \rho f_1' \\ \rho\lambda \end{bmatrix}, \quad \boldsymbol{f} = \begin{bmatrix} -\frac{\partial}{\partial x_2}(\rho v_2) \\ -\frac{\partial}{\partial x_2}(\rho v_1 v_2 + \sigma_{12}) \\ -\frac{\partial}{\partial x_2}(\rho v_2 v_2 + \sigma_{22}) \\ -\frac{\partial}{\partial x_2}(\rho c_v T v_2) + \omega_1 \\ -\frac{\partial}{\partial x_2}(\rho\phi v_2) + \omega_2 \\ -\frac{\partial}{\partial x_2}(\rho f_1' v_2) + \rho_0 \frac{\partial v_1}{\partial x_2} \\ -\frac{\partial}{\partial x_2}(\rho\lambda v_2) \end{bmatrix}, \quad \text{and } \boldsymbol{g} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \rho\Omega \end{bmatrix},$$

where ω_1 and ω_2 are the right-hand-side source terms of T and ϕ. The convective terms are discretized by the fourth-order convex ENO scheme [6] and the resulting semidiscretized equations $\boldsymbol{u}_t = \boldsymbol{f} + \boldsymbol{g}$ are a system of autonomous ODEs in $\boldsymbol{u}$ and are integrated in time via the third-order Runge–Kutta method as discussed earlier.

6. Simulations of longitudinal and shear motions. We have validated the code written for the full model through a series of graduated tests. Since equations that correspond to classical elastodynamics and classical gas dynamics can be obtained simply by suppressing the appropriate terms, limiting versions of the code can be used to solve problems with exact solutions, like standard Riemann problems or small amplitude linear wave propagation. These tests are fully documented in Yoh's Ph.D. thesis [11]. Similar test cases can be found in Ruderman's thesis [12]. For example, Riemann problems have been computed for a special case of an ideal gas. In the special limiting case of small-displacement elasticity for shear motions, with the assumption of constant material properties, one can show that there are dilatation waves that travel at $\sqrt{(\lambda_s + 2\mu_s)/\rho_0}$ and shear waves that travel at $\sqrt{\mu_s/\rho_0}$, where $\lambda_s = 2\mu_s\nu_c/(1 - 2\nu_c)$.

6.1. One-dimensional shear motions. Here we discuss representative solutions to an initial boundary-value problem that represents numerical experiments for shear motion. The problem set up is as follows: A slab of material 15 mm thick in the x_2-direction is initially at an elevated temperature and suddenly subjected to a constant velocity shearing motion at the edge $x_2 = 15$ mm while the edge at $x_2 = 0$ is held fixed. The material is thermally insulated. For the purpose of these experiments, the gas phase is suppressed and does not appear, hence the transitions documented here occur only between solid and liquid. We show representative results for two different initial conditions. First, we consider the initial temperature at 550 K with the constant shear velocity of 600 m/s, dubbed shear case A. In the second case, the initial temperature is slightly above the melting transition temperature at 560 K with a lower shearing velocity of 200 m/s, dubbed shear case B. Shear case B exhibits more complex dynamics associated with multiple regions of phase change. Both cases show generic elastic wave interactions and reflections within solid-fluid regions. The computational domain has 500 points spread uniformly over 15 mm.

Figures 10 and 11 show x_2-t contour plots of the thermodynamic variables T, ϕ, p, ρ, the velocities v_1 and v_2, and the displacement gradient $f_1'(X_2)$. Initially the hot sample, just below the melting temperature at 550 K, is exposed to the wall shear at 600 m/s. The rapid shearing at $x_2 = 15$ mm produces sufficient heating to cause rapid melting in a thin layer near the moving boundary. This is easily observed in Figure 10(a) and 10(b) for the temperature T and phase variable ϕ, respectively. The shearing motion is then confined mostly to a thin shear layer as seen in Figure 11(a) for velocity v_1 (in the direction of the imposed motion at $x_2 = 15$ mm). Note that the shear wave in the solid associated with v_1 is clearly observed as a wave that initially

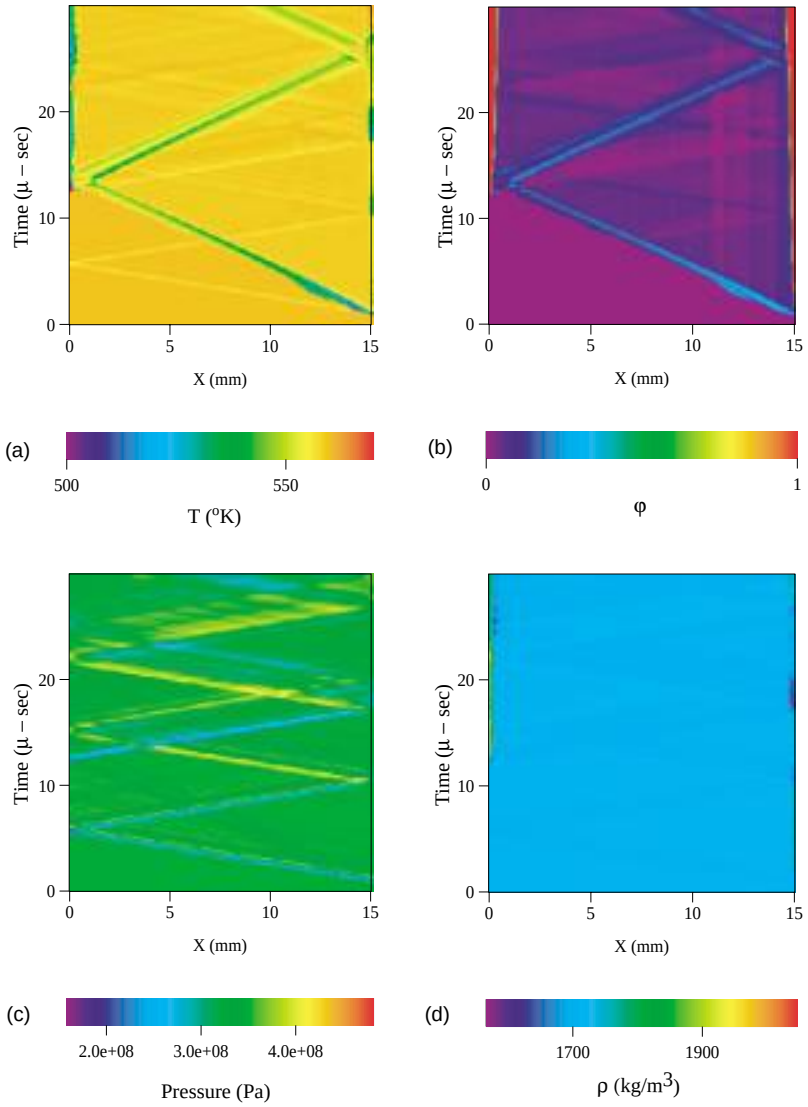


FIG. 13. *Temperature, phase, pressure, and density fields for a representative shear experiment* ($v_{shear} = 200$ *m/s*, $T_0 = 560$ *K*).

enters the domain at $x_2 = 15$ mm and travels toward $x_2 = 0$ mm and subsequently reflects off the stationary wall.

Figure 11(b) for v_2 displays waves that travel at the dilatational wave speed, which is approximately twice the shear wave speed. The dilatational waves are generated by the initial growth of the melted layer and are associated with pressure waves of magnitude of approximately 10^8 Pa = 1 kbar. Note that the initial stress in the system is elevated due to the effect of thermal expansion at the initially raised temperature. Close inspection of the temperature and pressure fields shown in Figure 10(a), (c) shows evidence of high frequency acoustic waves that can be traced to reflections and transmissions of waves through the solid/liquid interface near $x_2 = 15$ mm.

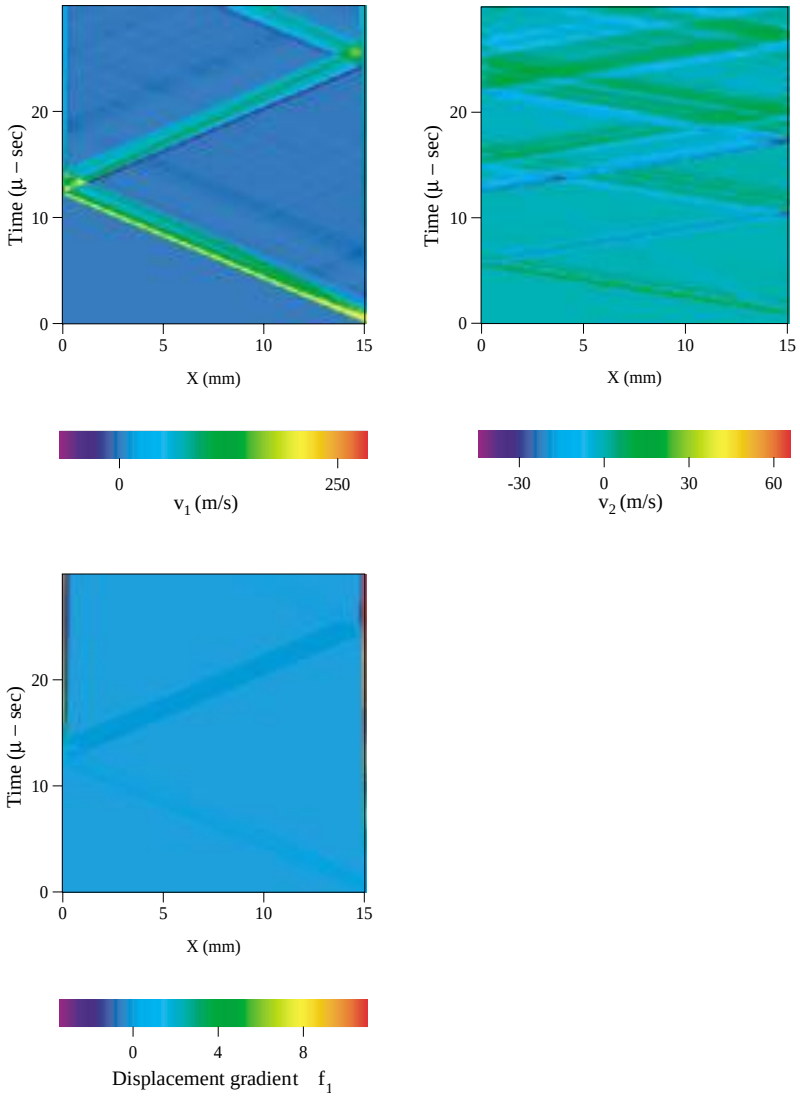


FIG. 14. *Shear velocity (v_1), compression velocity (v_2), and displacement gradient (du_1/dX_2) fields for a representative shear experiment ($v_{shear} = 200$ m/s, $T_0 = 560$ K).*

Figure 12 shows computed profiles for shear case A for ρ, T, p, and ϕ at times 5, 15, and 30 μsec, which represents time cuts across Figure 10(a)–(d). The profiles show elevated temperatures and phase change (melting) confined to the layer near the $x_2 = 15$ mm boundary. The fluctuations in the pressure, density, and temperature profiles are the result of the acoustic disturbances propagating through the solid and across the solid/liquid layer.

Note that the layer of liquid that develops at $x_2 = 15$ mm is a localized shear layer and can be thought of as a shear band. The material in the melt layer has very large v_1-velocities and subsequently undergoes large deformation. The material in the solid phase essentially remains fixed in place as the fluid layer slides across it.

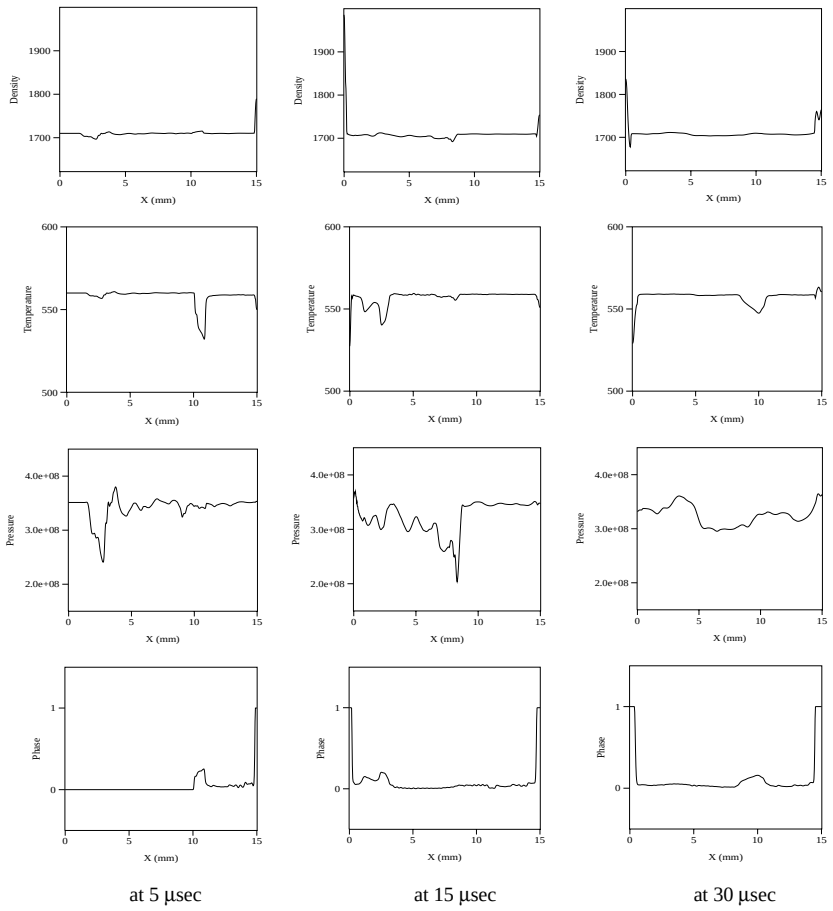


FIG. 15. *Instantaneous profiles of density, temperature, pressure, and phase field (from top to bottom) taken at time* $t = 5,\ 15,\ 30\,\mu$*sec corresponding to Figures* 13 *and* 14 *of the plane shearing experiment.*

Shear case B has the solid with its initial temperature slightly above the melt temperature, suddenly subjected to a (lower) constant shear motion of 200 m/s. As in case A, a melt layer forms near $x_2 = 15$ mm and the dilation wave travels across the slab. After reflection at the fixed wall, a second melt layer develops near $x_2 = 0$ mm. Figures 13, 14, and 15 show the additional complexity in the x_2-t record. The second melt layer causes additional scatter of waves generated near the $x_2 = 15$ mm boundary, and, in turn, the growth of the layer generates additional disturbances which transmit through the regions. One recalls that there are additional terms in the ϕ-evolution equation that are associated with the deformational part of the stress. We clearly see that the stress waves (by themselves) can induce the phase transformation. One sees transient phase generation carried on the subcharacteristics in the phase variable plot Figure 13(b). The next set of experiments for longitudinal motions illustrates shock melting.

6.2. One-dimensional longitudinal motion: Reverse impact. The results discussed next are for two different longitudinal motions where a HMX specimen of thickness 15 mm is initially solid at the melt temperature ($T = 558$ K) and subjected

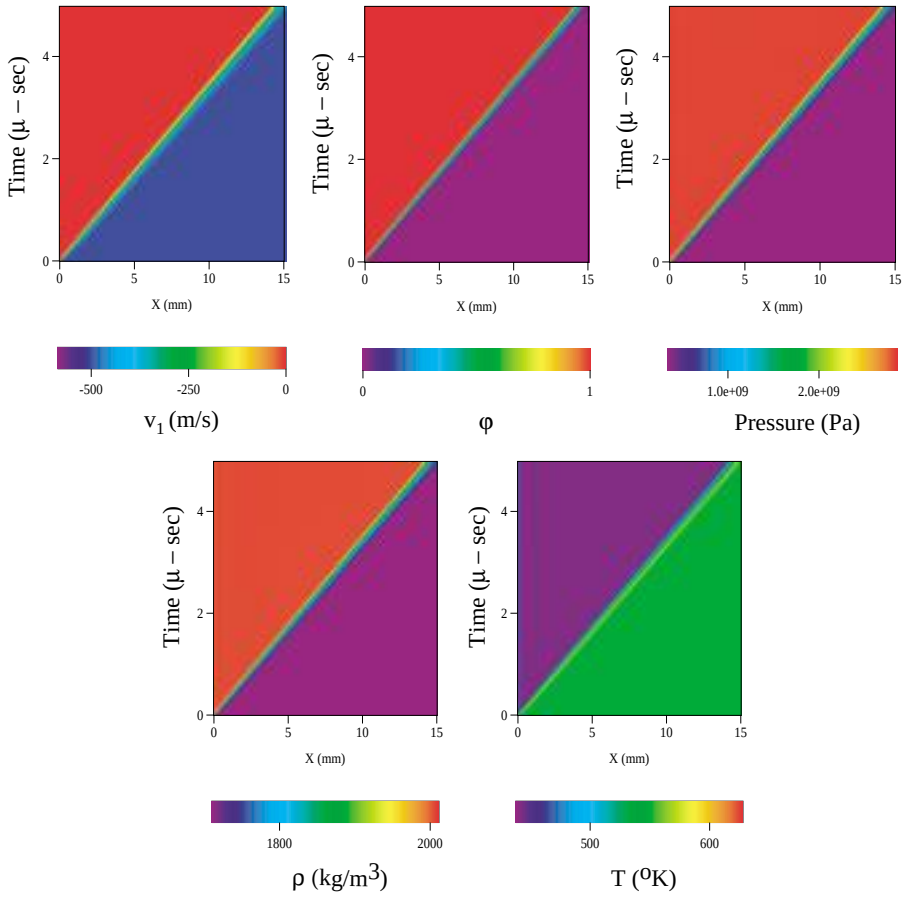


FIG. 16. *Velocity, phase, pressure, density and temperature fields for a representative reverse-impact (longitudinal) experiment ($v_{impact} = -500$ m/s, $T_0 = 550$ K).*

to a reverse impact at speed −500 m/s for longitudinal case A and −200 m/s for longitudinal case B. The computational domain spans the 0.015 m with 500 mesh points.

For longitudinal case A, Figures 16 and 17 illustrate the phenomenon of shock melting as predicted by the model. Figure 16(a)–(e) clearly shows the emergence of a shock wave from the stationary wall into the oncoming stream. Ahead of the shock the material is solid with $\phi = 0$; behind the shock the material is liquid with $\phi = 1$. The model predicts a shock with definite spatial structure as illustrated by the structure profiles taken at $t = 3\,\mu$sec and shown in Figure 17. In the shocked state, where the material has liquefied, there is a significant pressure increase to about 2.8 GPa (28 kbar)—a 20% density increase—and a drop in the temperature due to the endothermic nature of the phase transformation. The pressure and density rise monotonically across the shock structure. The temperature increases slightly, then drops with the onset of the phase transformation from solid to liquid. Throughout the structure, the phase changes monotonically from solid to liquid.

Longitudinal case B corresponds to a reverse-impact experiment where the impact speed is reduced to −200 m/s but the initial temperature is raised slightly to 560 K,

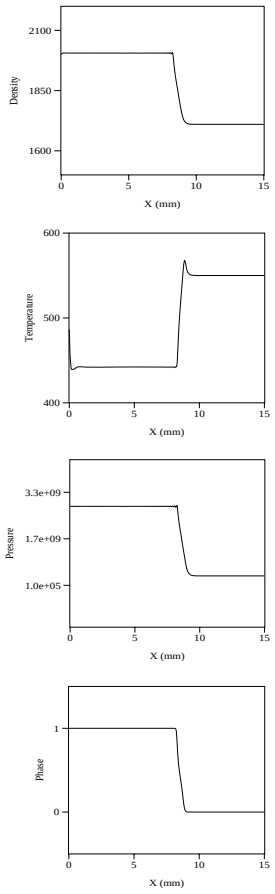


FIG. 17. *Snapshots of density, temperature, pressure, and phase field (from top to bottom) taken at time* $t = 3\,\mu$*sec from Figure* 16 *of the longitudinal exercise.*

just two degrees above the melt temperature. Figure 18 shows the x_1-t contour plots. Figure 19 shows corresponding line cuts taken at time $t = 3\,\mu$sec. Similar to longitudinal case A, shock induced phase transformation occurs; however, a stable intermediate phase is produced behind the shock front with $\phi = 0.33$. Interestingly, the model can be shown to allow these intermediate states in ϕ due to the contributions of the other stress-dependent source terms proportional to $\mu_s'(\phi), \mu_c'(\phi), \alpha_c'(\phi)$, etc. as found in (2.7). A complete analysis of all possible ϕ-states and their stability is beyond the scope of this paper. However, we can illustrate the stability of the intermediate state for longitudinal case B by a numerical evaluation as follows. We take the evolution equation for ϕ, (I 6.9), to be rewritten as $\frac{\partial \phi}{\partial t} = -v_1 \frac{\partial \phi}{\partial x_1} + w_2$, where w_2 is the source term for the material derivative of ϕ. We then take the shock structure as obtained numerically at $t = 3\,\mu$sec for both longitudinal cases A and B and plot $\partial\phi/\partial t$ versus ϕ in Figure 20. Stable equilibria points (in ϕ) are found by the zeros of $\partial\phi/\partial t$. For longitudinal case A, only $\phi = 0$ and $\phi = 1$ are stable with $\partial\phi/\partial t = 0$. But for longitudinal case B, the intermediate state $\phi = 0.33$ is found to be stable. Our numerical experiments suggest that increasing the intensity of the reverse impact causes the intermediate states to disappear with $\phi = 0, 1$ as the only

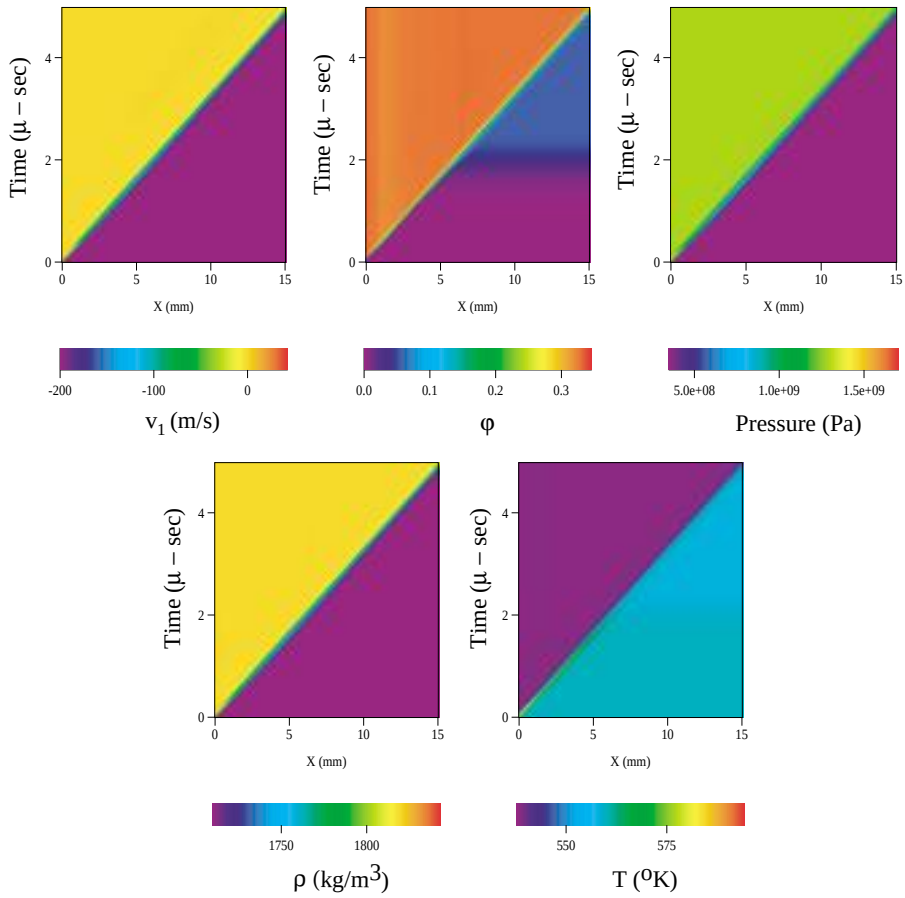


FIG. 18. *Velocity, phase, pressure, density and temperature fields for a representative reverse-impact (longitudinal) experiment ($v_{impact} = -200$ m/s, $T_0 = 560$ K).*

stable equilibria.

7. Conclusions. We have illustrated that our model, fitted to a real material, leads to predictions of simple motions (constant volume evolution, shear motion, and longitudinal motion) that are plausible. The model has the property that the constitutive theory automatically changes with the phase and is consistent with classical properties of that phase. We have shown that it is possible to fit the model to the known behavior of a real material.

Although idealized, the representative numerical experiments exhibit extremely rich behaviors. Strain localization phenomena occurred via melting in thin layers in many of the trials we have conducted. The phase change phenomena is directly coupled to the material loading through the change in material type and changes in properties that are carried with the phase. We are ready to apply this new continuum model to more complex physical problems of interest to us. Of course, extremely interesting and varied mathematical problems, such as steady traveling waves and their multidimensional stability, will arise that can profitably be analyzed by asymptotic means. The models embedded within this larger model may have greater application

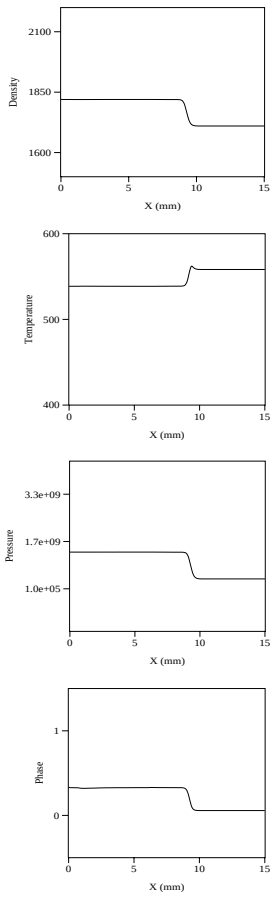


FIG. 19. *Snapshots of density, temperature, pressure, and phase field (from top to bottom) taken at time* $t = 3\,\mu$*sec from Figure* 18 *of the longitudinal experiment.*

to the general theory of phase transformation. Of specific near term interest to us is a detailed study of the mechanically induced ignition of an energetic solid. We also plan to pursue a simplified version of this model to more fully examine the processes of classical melting/freezing and vaporization/condensation in the context of the model. We also anticipate the near term application of the model to problems of vaporizing fuels and condensed phase propellant combustion.

Appendix. List of ϕ-dependent functions for $0 \le \phi \le 2$.

$$
\begin{aligned}
F(\phi) &= [\phi(\phi-1)(\phi-2)]^2 \\
F'(\phi) &= 2\phi(4 - 18\phi + 26\phi^2 - 15\phi^3 + 3\phi^4),
\end{aligned}
$$

$$
\begin{aligned}
\beta'_m(\phi) &= \begin{cases} 6\phi(1-\phi) & \text{for } 0 \le \phi \le 1, \\ 0 & \text{otherwise,} \end{cases} \\
\beta'_v(\phi) &= \begin{cases} 6(\phi-1)(2-\phi) & \text{for } 1 \le \phi \le 2, \\ 0 & \text{otherwise,} \end{cases}
\end{aligned}
$$

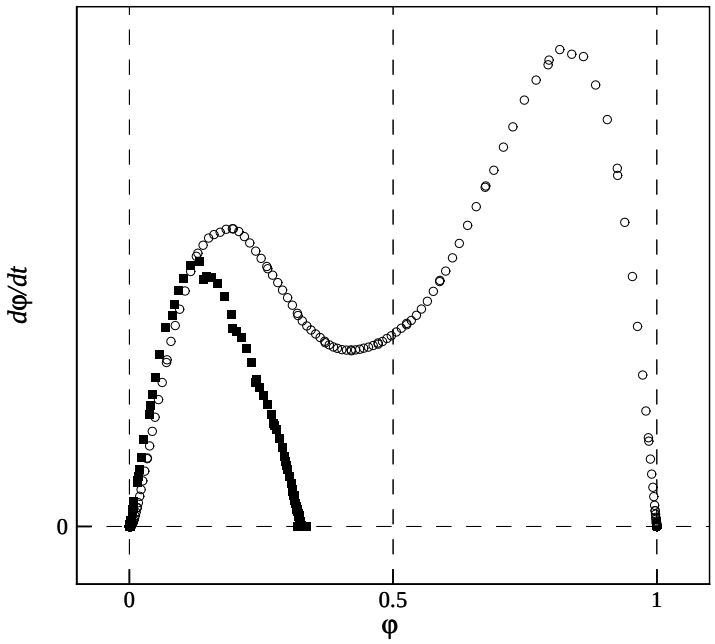


FIG. 20. *$d\phi/dt$ versus ϕ for the two specialized reverse-impact experiments discussed. By varying the initial temperature, T_0, meta-stable state ($\phi \approx 0.33$) is shown as a local equilibrium point on the experiment represented by the square symbols. In contrast, the experiment shown by hollow circles suggests that $\phi = 1$ is the only stable equilibria once the initial state is perturbed about the unstable point at $\phi = 0$, corresponding to the solid state under the impact loading.*

$$\mu_c(\phi) = \begin{cases} 2(\mu_{solid} - \mu_{liquid})\phi^3 - 3(\mu_{solid} - \mu_{liquid})\phi^2 + \mu_{solid} & \text{for } 0 \le \phi \le 1, \\ 2(\mu_{liquid})(\phi-1)^3 - 3(\mu_{liquid})(\phi-1)^2 + \mu_{liquid} & \text{for } 1 \le \phi \le 2, \end{cases}$$

$$\mu_c'(\phi) = \begin{cases} 6(\mu_{solid} - \mu_{liquid})\phi^2 - 6(\mu_{solid} - \mu_{liquid})\phi & \text{for } 0 \le \phi \le 1, \\ 6(\mu_{liquid})(\phi-1)^2 - 6(\mu_{liquid})(\phi-1) & \text{for } 1 \le \phi \le 2, \end{cases}$$

$$\mu_s(\phi) = \begin{cases} 2(\mu_{solid})\phi^3 - 3(\mu_{solid})\phi^2 + \mu_{solid} & \text{for } 0 \le \phi \le 1, \\ 0 & \text{for } \phi > 1, \end{cases}$$

$$\mu_s'(\phi) = \begin{cases} 6(\mu_{solid})\phi^2 - 6(\mu_{solid})\phi & \text{for } 0 \le \phi \le 1, \\ 0 & \text{otherwise,} \end{cases}$$

$$\mu_l(\phi) = \begin{cases} 2(-\mu_{liquid})\phi^3 - 3(-\mu_{liquid})\phi^2 & \text{for } 0 \le \phi \le 1, \\ 2(\mu_{liquid})(\phi-1)^3 - 3(\mu_{liquid})(\phi-1)^2 + \mu_{liquid} & \text{for } 1 \le \phi \le 2, \end{cases}$$

$$\mu_l'(\phi) = \begin{cases} 6(-\mu_{liquid})\phi^2 - 6(-\mu_{liquid})\phi & \text{for } 0 \le \phi \le 1, \\ 6(\mu_{liquid})(\phi-1)^2 - 6(\mu_{liquid})(\phi-1) & \text{for } 1 \le \phi \le 2, \end{cases}$$

$$R(\phi) = \begin{cases} 2(-R_{gas})(\phi-1)^3 - 3(-R_{gas})(\phi-1)^2 & \text{for } 1 \le \phi \le 2, \\ 0 & \text{for } \phi < 1, \end{cases}$$

$$R'(\phi) = \begin{cases} 6(-R_{gas})(\phi-1)^2 - 6(-R_{gas})(\phi-1) & \text{for } 1 \le \phi \le 2, \\ 0 & \text{otherwise,} \end{cases}$$

$$\alpha_c(\phi) = \begin{cases} 2(\alpha_{solid})(\phi-1)^3 - 3(\alpha_{solid})(\phi-1)^2 + \alpha_{solid} & \text{for } 1 \le \phi \le 2, \\ \alpha_{solid} & \text{for } \phi < 1, \end{cases}$$

$$\alpha_c'(\phi) = \begin{cases} 6(\alpha_{solid})(\phi-1)^2 - 6(\alpha_{solid})(\phi-1) & \text{for } 1 \le \phi \le 2, \\ 0 & \text{otherwise,} \end{cases}$$

$$\rho\gamma_\phi(\phi) = \begin{cases} 6(-\rho\gamma_\phi)\phi^2 - 6(-\rho\gamma_\phi)\phi & \text{for } 0 \le \phi \le 1, \\ 6(-\rho\gamma_\phi)(\phi-1)^2 - 6(-\rho\gamma_\phi)(\phi-1) & \text{for } 1 \le \phi \le 2, \end{cases}$$

$$c_v(\phi) = c_v$$

REFERENCES

[1] M. E. Gurtin, *Configurational Forces as Basic Concepts of Continuum Physics*, Appl. Math. Sci. 137, Springer-Verlag, New York, 2000.

[2] G. A. Ruderman, D. S. Stewart, and J. J.-I. Yoh, *A thermomechanical model for energetic materials with phase transformations*, SIAM J. Appl. Math., 63 (2002), pp. 510–537.

[3] T. B. Brill, *Multiphase chemistry considerations at the surface of burning nitramine monopropellants*, J. Propulsion Power, 11 (1995), pp. 740–750.

[4] J.-P. Poirier, *Introduction to the Physics of the Earth's Interior*, Cambridge University Press, Cambridge, UK, 1991.

[5] B. M. Dobratz and P. C. Crawford, *LLNL Explosive Handbook*, Lawrence Livermore National Laboratory, Livermore, CA, 1985.

[6] X.-D. Liu and S. Osher, *Convex ENO high order multi-dimensional schemes without field by field decomposition or staggered grids*, J. Comput. Phys., 142 (1998), pp. 304–330.

[7] R. Menikoff and T. D. Sewell, *Constituent properties of HMX needed for meso-scale simulations*, Appl. Phys. Rev., submitted.

[8] C. M. Tarver, S. K. Chidester, and A. L. Nichols, III, *Critical conditions for impact- and shock-induced hot spots in solid explosives*, J. Phys. Chem., 100 (1996), pp. 5794–5799.

[9] J. J. Yoh and X. Zhong, *Low-storage semi-implicit Runge–Kutta schemes for chemically reacting flow computations*, J. Comput. Phys., submitted.

[10] C. Yoo and H. Cynn, *Equation of state, phase transition, decomposition of β-hmx (octahydro-1,3,5,7-tetranitro-1,3,5,7-tetrazoncine) at high pressures*, J. Chem. Phys., 111 (1999), pp. 10229–10235.

[11] J. J. Yoh, *Thermomechanical and Numerical Modeling of Energetic Materials and Multimaterial Impact*, Ph.D. thesis, University of Illinois, Urbana-Champaign, IL, 2001.

[12] G. A. Ruderman, *A Continuum Thermomechanical Model for Energetic Materials*, Ph.D. thesis, University of Illinois, Urbana-Champaign, IL, 1998.

[13] D. Bedrov, G. D. Smith, and T. D. Sewell, *Temperature dependent shear viscosity coefficients of HMX, a molecular dynamics simulation study*, J. Chem. Phys., 112 (2000), pp. 7203–7208.

SIAM J. APPL. MATH.
Vol. 63, No. 2, pp. 564–592

EULER'S ELASTICA AND CURVATURE-BASED INPAINTING*

TONY F. CHAN†, SUNG HA KANG†, AND JIANHONG SHEN‡

Abstract. Image inpainting is a special image restoration problem for which image prior models play a crucial role. Euler's elastica was first introduced to computer vision by Mumford [*Algebraic Geometry and its Applications*, Springer-Verlag, New York, 1994, pp. 491–506] as a curve prior model. By functionalizing the elastica energy, Masnou and Morel [*Proceedings of the 5th IEEE International Conference Image Processing*, 3 (1998), pp. 259–263] proposed an elastica-based variational inpainting model. The current paper is intended to contribute to the development of its mathematical foundation and the study of its properties and connections to the earlier works of Bertalmio, Sapiro, Caselles, and Ballester [*SIGGRAPH* 2000, ACM Press, New York, 2000] and Chan and Shen [*J. Visual Comm. Image Rep.*, 12 (2001), pp. 436–449]. A computational scheme based on numerical PDEs is presented, which allows the automatic handling of topologically complex inpainting domains.

Key words. inpainting, elastica, prior models, Bayesian, variational method, bounded variation, curvature, transport, diffusion, numerical PDE

AMS subject classifications. Primary, 94A08; Secondary, 68U10, 65K10

PII. S0036139901390088

1. Introduction.

Among museum conservators and restoration artists, inpainting refers to the practice of retouching or recovering damaged ancient paintings [14, 40]. The goal is to remove the cracks or recover the missing patches in an undetectable manner.

The term *digital inpainting* was initially introduced into digital image processing by Bertalmio, Sapiro, Caselles, and Ballester in [3], where the authors first made an innovative construction of a third-order PDE inpainting model. Equally important in [3] is that the authors demonstrated the broad applications of digital inpainting in film restoration, text removal, scratch removal, and special effects in movies. The same group of authors has also recently developed a variational inpainting model based on a joint cost functional on the gradient vector field and gray values [1]. An earlier variational inpainting model was studied by Masnou and Morel [27] in the context of disocclusion in computer vision. Recently, Chan and Shen have proposed the *total variation* (TV) inpainting model [8] and a new PDE inpainting model based on *curvature driven diffusions* (CDD) [9]. In [8], Chan and Shen also discovered novel applications in digital zooming and primal-sketch-based [26] image coding schemes.

Inpainting is essentially an interpolation problem, with special focus on situations in which image information is partially missing or inaccessible on certain two-dimensional (2-D) regions. What makes image inpainting highly nontrivial is the complexity of image functions, caused by the richness of geometric structures and a large dynamic range of scales.

A simple but often sufficient mathematical model for generic nontexture images

*Received by the editors May 25, 2001; accepted for publication (in revised form) May 27, 2002; published electronically November 19, 2002. This research was supported by the NSF under grants DMS-9973341 and DMS-0202565 and by the ONR under grant N00014-02-1-0015.
http://www.siam.org/journals/siap/63-2/39008.html
†Department of Mathematics, UCLA, Los Angeles, CA 90095 (chan@math.ucla.edu, skang@math.ucla.edu).
‡School of Mathematics, University of Minnesota, Minneapolis, MN 55455 (jhshen@math.umn.edu).

is the BV (*bounded variation*) space, where the most crucial low-level visual cue—the edge—is permissible (see Rudin and Osher [36], Rudin, Osher, and Fatemi [37], and Chambolle and Lions [7] for examples). Yet, in both the approximation community and that of numerical analysis, as far as we know, there has not been much work on the inpainting of BV functions. A recent related paper by Cohen et al. [12] discusses the nonlinear approximation of BV functions, motivated by the wavelets thresholding technique and the Rudin–Osher–Fatemi denoising model [37]. The nonlinear interpolation studied therein is for the near-optimal approximation and representation of a given *complete* noisy image u_0. But inpainting has a different objective, which is to recover the entire clean image u from a given *incomplete* noisy image u_0 observed only outside a missing domain K.

The inpainting of BV images is generally an ill-posed problem, which can be seen more clearly through a one-dimensional (1-D) example. Imagine that we know the values (or even the derivatives) of a function f at $a-h$ and $a+h$. If f is smooth, then as $h \to 0$ we can apply smooth interpolants such as Lagrange's and Hermite's to infer the values of f on $(a-h, a+h)$ with a certain degree of precision guaranteed. But for a BV function f, all such smooth interpolants fail to work properly no matter how small h is, since a "widthless" jump can always occur in $(a-h, a+h)$. Under the TV Radon measure, a single point in 1-D can have nonzero mass, which makes the corresponding interpolation problem ill-posed generally. (In the high-dimensional case, the BV interpolant performs better than in 1-D, due to its connection to the *minimal surface* problem. See, for example, [8, 18].)

The good news is that images as BV functions are not too intractable. Each image is a 2-D projection of a window of the 3-D world, in which individual objects often have their geometric or surface reflectivity regularities. Such regularities partially diminish the ill-posedness of the inpainting problem.

Given an image, if we partially cover it with a piece of paper of moderate size and ask a person to guess what has been occluded in the original image, most people will come up with a "rational" best guess. For example, if a green apple in a photo is partially occluded by a piece of paper, then one often first estimates the occluded boundary and then inpaints with the green color over the occluded area belonging to the apple. All these decisions are realized by the best guesses or, more scientifically, by Bayesian inference [16, 21, 30]. The two ingredients of Bayesian inference are the *prior* model and *data* model. The data model is simple for most inpainting problems: the available data is simply a part of a complete image that we try to restore. Thus the prior model plays a crucial role in our inference process. For the apple experiment mentioned above, the *a priori* knowledge of the shape and color of an apple is helpful for a person trying to make a meaningful inpainting.

In order to develop a general inpainting model, one should never rely on the *prior* model of a *specific* class of image objects (such as apples). The model must employ generic regularities to better condition the ill-posedness. The Rudin–Osher–Fatemi BV [37] and Mumford–Shah's object-and-edge models [31] are the two most well-known prior models for generic nontexture images. However, as Chan and Shen [8] and Esedoglu and Shen [15] discussed, they become less suitable in some large-scale inpainting problems where they lead to a violation of the *connectivity principle* in vision psychology [20]. As a result, it is clear that a good inpainting model must consider *high-order* geometric information such as the curvature feature of the level sets [9].

In the current paper, we study a variational inpainting model that has arisen

from a second-order plane curve model—Euler's elastica. The gap between a prior model for curves and that for images is formally bridged by the level set concept: generally, a curve prior model can always be "lifted" to an image prior model once being imposed on all the level sets of an image (similar to the *coarea* formula in the theory of BV functions [18]). Indeed, this is how Masnou and Morel first proposed this model for image inpainting in [27]. Our current paper is intended to (a) study the mathematical foundation and properties of the inpainting models based on elasticas and curvatures, (b) explore the connection of this work to the earlier empirical works on PDE-based inpainting models, and (c) construct numerical PDE schemes for the associated nonlinear PDEs. Compared with Masnou and Morel's linear programming algorithm in [27], the numerical PDE approach is more flexible in that it frees one from laboring over edge detection or pixel coupling along the boundaries, and also that it lifts the topological restriction on the inpainting domain K [27].

Euler's elasticas were first introduced and seriously studied in computer vision by Mumford [29] as a prior curve model. They were employed in the visual disocclusion program in [32] to smoothly connect occluded edges and T-junctions. Earlier in approximation theory, elasticas had been introduced as *nonlinear splines* by Birkhoff and De Boor [4].

We are now ready to introduce the layout of the paper. We begin with a brief introduction to the elasticas (section 2). Then, as in classical interpolation theory, in section 3, we first study the generic local models for nontexture images. By the method of moving frames, we are able to inpaint or interpolate the missing T-junctions or corners inside the inpainting domain based on the elastica interpolant. We then explain the approach that leads to Masnou and Morel's algorithm on individually "engineering" isophotes (i.e., level sets) [27]. A level set-type idea [34] formally "lifts" this isophote-based model to an energy functional that acts directly on gray scale images, a process that we shall call "functionalization."

The rest of the paper is devoted to the mathematical analysis and computational implementation of elastica-based inpainting models.

In section 4, by introducing the concept of the *weak curvature* of a general BV function, we legitimize the functionalization of the elastica energy in the BV space. The direct method for elastica inpainting is generally difficult due to the lack of classical fine properties (such as convexity and lower semicontinuities [2]). But for the extreme case of TV inpainting, we are able to rigorously establish the existence theorems based on the theory of BV functions. We also discuss why the nonuniqueness of the solutions to a general inpainting model should be somehow appreciated instead of being cursed. The last part of the section discusses various relaxation schemes for the constraints of the inpainting energy.

In section 5, we derive the formal Euler–Lagrange equation for a general curvature-based inpainting model. The most interesting result is the structure of the associated flux field $\vec{V}$, which provides a formal way to unify the earlier empirical work of Bertalmio et al. [3] on transport-based inpainting, and that of Chan and Shen on CDD-based inpainting [9]. We therefore conjecture that transport and diffusion are the two universal infinitesimal mechanisms for any PDE-based inpainting schemes.

In section 6, we explain our numerical schemes for the associated nonlinear Euler–Lagrange equation and demonstrate several computational examples. The computational schemes have been inspired by various techniques in the classical literature of computational fluid dynamics.

In the last section, we conclude with a brief discussion on the connection between

the elastica inpainting model and a recent model by Ballester et al. [1] on the joint interpolation of both the normal field $\vec{n} = \nabla u/|\nabla u|$ and gray values u.

2. Euler's elastica and its Bayesian rationale. A curve Γ is said to be *Euler's elastica* if it is the equilibrium curve of the elasticity energy

$$E_2[\gamma] = \int_\gamma (a + b\kappa^2) ds, \tag{2.1}$$

where ds denotes the arc length element, $\kappa(s)$ the scalar curvature, and a, b two positive constant weights. Extra constraints may include the positions and normal directions of the two ends. Euler obtained the energy in 1744 in studying the steady shape of a thin and torsion-free rod under external forces (see [23]).

Since both the arc length and curvature are intrinsic geometric features of a curve, the elastica energy naturally extends to the curves on a general Riemannian manifold M. For example, if M is embedded in a Euclidean space R^N, a curve γ on M can be expressed by the embedded coordinates

$$s \to \vec{x}(s) = (x_1(s), \dots, x_N(s)).$$

Then $\vec{t} = d\vec{x}/ds$ is the tangent and $\prod_{\vec{x}} d\vec{t}/ds = \kappa\vec{n}$ defines the curvature, with $\prod_{\vec{x}}$ representing the orthogonal projection from $T_{\vec{x}}R^N$ to $T_{\vec{x}}M$. For a general Riemannian manifold M, the intrinsic derivative $d\vec{t}/ds$ is defined by the Levi–Civita connection or the covariant derivative (see Chern, Chen, and Lam [11], for example). Such extension onto general manifolds is motivated by inpainting problems on arbitrary surfaces in R^3, for example, the inpainting of an incomplete image on the surface of a Coke can in computer graphics.

By the calculus of variations, it can be shown that an elastica must satisfy

$$2\kappa''(s) + \kappa^3(s) = \frac{a}{b}\kappa(s).$$

For example, Mumford gives a detailed derivation in [29]. More generally, if the elastica lives on a Riemannian surface, there will be an extra term due to the curving of the surface:

$$2\kappa''(s) + \kappa^3(s) + 2G(s)\kappa(s) = \frac{a}{b}\kappa(s),$$

with $G(s)$ denoting the Gaussian curvature of the surface. More studies in elasticas on general Riemannian manifolds can be found in Langer and Singer [22].

The elastica was first seriously studied from the computer vision point of view in Mumford's paper [29], the introduction section of which provides a delightful view of its mathematical history. According to [29], the key link between the elastica and computer vision is founded on the interpolation capability of elasticas, as initially proposed by Birkhoff and De Boor [4]. Such "nonlinear splines" [4], like classical polynomial splines, are natural tools for completing the missing or occluded edges [32].

A remarkable feature of elasticas revealed by Mumford [29] is their Bayesian rationale, which enhances the interpolating role of elasticas in image and vision analysis. It also sheds light on the choice of "2" for the curvature power in the energy formula (2.1). Here we present a slightly polished version of Mumford's original argument.

Consider the random walk of a drunk initially staying at the origin of a 2-D ground. Assume that each individual step is straight. For some given fixed integer N, we try to understand the distribution of all possible N-step polygonal walks. The moving characteristics are the following:

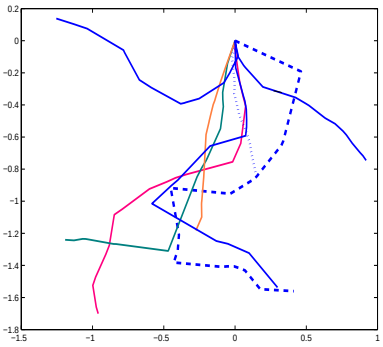


FIG. 2.1. *Some sample paths for the drunk's walking.*

(a) Let h_k denote the step size of the kth step. Then $\{h_k \ : \ k = 1, 2, \dots, N\}$ are independently and identically distributed (i.i.d) of exponential type $\lambda \exp(-\lambda h)$ for some positive mean $1/\lambda$.

(b) Let θ_k denote the orientation of the kth step, measured by the angle between the walking direction and the x-axis, and define $\theta_0 = 0$. Let $\Delta\theta_k = \theta_k - \theta_{k-1}$ $(k = 1, 2, \dots, N)$ denote the turn made at the kth step. The basic assumption is that, at each step k, the larger the step size h_k is, the more uncertain the turn $\Delta\theta_k$ will become. Precisely, $\Delta\theta_k$ is a Gaussian of $N(0, h_k\sigma^2)$. Yet

$$\left\{ n_k = \Delta\theta_k / \sqrt{h_k} \ : \ k = 1, 2, \dots, N \right\}$$

is an independent set and also independent of all the h_k's.

Thus, an N-step polygonal walk γ is completely determined by the data

$$\{h_1, \dots, h_N\} \cup \{\Delta\theta_1, \dots, \Delta\theta_N\},$$

and the likelihood is quantified by

$$\lambda^N \exp(-\lambda(h_1 + \cdots + h_N))\, dh_1 \cdots dh_N \\ \times (\sqrt{2\pi}\sigma)^{-N} \exp\left(-\frac{1}{2\sigma^2} \left[\frac{(\Delta\theta_1)^2}{h_1} + \cdots + \frac{(\Delta\theta_N)^2}{h_N} \right] \right) dn_1 \cdots dn_N.$$

Up to a multiplicative constant, this probability density function is exactly

$$\exp\left(-\lambda L(\gamma) - \frac{1}{2\sigma^2} \|\kappa^2\|_\gamma \right),$$

where L denotes the length and $\|\kappa^2\|_\gamma$ the discrete analogy of $\int_\gamma \kappa^2 ds$. Therefore, the minimization of the elastica energy (2.1), with $a = \lambda$ and $b = 1/(2\sigma^2)$, is equivalent to the maximum likelihood (ML) estimation of such random curves.

Remark 1. Notice that the walking drunk model presented here does not come from the discrete sampling of the Brownian motion, since for the latter (in 2-D), h^2 (not h) is exponential and the turn $\Delta\theta$ is uniform along the unit circle. The dependence of the turns on the step sizes makes the paths smoother than the sampling of Brownian motions, which makes the model a valuable curve model in computer vision, where regularity is important. Figure 2.1 shows a computer simulation of the paths.

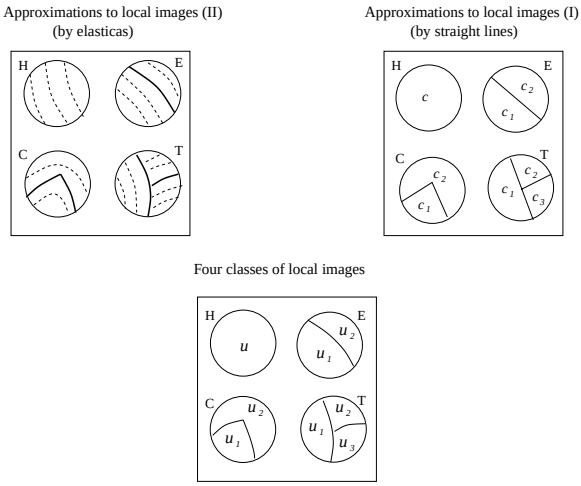


FIG. 3.1. *Local (nontexture) image models and their approximation.*

3. Local (nontexture) image models and elastica isophotes.

3.1. Generic local (nontexture) image models. To model the inpainting of nontexture images, it is important to start with local and small inpainting domains. Focusing on locality is a common practice in numerical mathematics, where from differentiation, integration, and interpolation, to optimization, most well-known numerical schemes (such as the Runge–Kutta schemes, Simpson's integration rule, and the Newton–Raphson searching strategy for zeros or valleys) are inspired by the local models of the target functions, such as the linear model, parabolic model, and so on [19, 39]. Digital inpainting, after all, is the numerical interpolation of 2-D images. Therefore, it is crucial to first understand what an image looks like locally.

Imagine that we have a *small* aperture with a circular shape of radius r, and that we focus on only the part that is captured within when it is moved over a 2-D gray image. Suppose that the image only contains man-made nontexture objects, and that their characteristic scale $l \gg r$. What sorts of local image patches will be observed most frequently?

They can be grouped into four classes, labeled by "H," "E," "C," and "T" (see Figure 3.1), as follows:

(a) Class H. A local image patch belongs to this class if and only if it falls within the *homogeneous* regions. Such a patch has very little intensity variation.

(b) Class E. This occurs when the aperture captures a fraction of the smooth *edge* between two objects or homogeneous regions.

(c) Class C. Like class E, but the aperture captures a recognizable *corner*. Corners are also a universal feature of man-made objects, from tables, windows, and books, to posters.

(d) Class T. This is the case in which the aperture captures a *T-junction*. T-junction is an important cue for occlusion and the perception of orders in the lost dimension of range [32]. A local T-junction patch is characterized by three homogeneous gray values u_1, u_2, and u_3 and two smooth edges—one meets the boundary at its two ends while the other at only one end due to occlusion.

Notice that homogeneous regions are 2-D objects, edges are 1-D objects, and both corners and T-junctions are isolated 0-D objects. Therefore, heuristically, in terms of the probability (or frequency) of being observed through a small aperture, we have the following relation:

$$\text{Prob}(H) \gg \text{Prob}(E) \gg \text{Prob}(C) \text{ or } \text{Prob}(T).$$

Of course, the probabilities do not necessarily represent their perceptual significance in terms of vision inference. In fact, it seems that often the scattered singular features can generate strong response from the vision system (e.g., the wavelets idea [13, 5, 35]).

Notice that class C is indeed very much "man-made" like its ancestors in our 3-D world, in the sense that a local engineering (or small perturbation) of the corner can easily change a class C patch to class E. In this sense class C is unstable and nongeneric.

3.2. Local edge interpolation by elasticas: Moving frames. Suppose an image u_0 has a local patch K missing, and we try to inpaint $u_0\big|_K$ based on the available information surrounding K.

By checking the available data close to K, we can easily determine which class $u_0\big|_K$ belongs to: H, E, or T. (Without a priori knowledge, one can never distinguish the nongeneric class C from the generic class E.) The existence of a corner or the exact configuration of the T-shape inside the inpainting domain will require extra information or additional models and algorithms (especially for the coupling of the three end points in the T-junction case). (For the corner case, for example, see Shah's recent work on elasticas with hinges [38].) In this paper, we shall assume that eventually we do know the exact class that $u_0\big|_K$ belongs to and the precise configurations.

The first level of approximation will be based on the *straight line* curve model (refer to the upper-right panel in Figure 3.1). For a class H type, one can average the available boundary pixel values and inpaint $u_0\big|_K$ by this mean value c. For a class E type, we connect the two edge ends by a straight segment and inpaint the two objects u_1 and u_2 by their boundary mean values c_1 and c_2. For a class C type, we make two straight shoots from the boundary end points into the inpainting domain. The orientations follow their cue left outside K. The two straight lines generate a corner and also segment the patch K into two objects u_1 and u_2. Then u_1 and u_2 are inpainted by their boundary mean values. For the last class T, we connect the two coupled boundary end points by a straight line and shoot from the third one directly toward the interior of K, as for class C. Then the three segmented objects u_1, u_2, and u_3 are inpainted by their boundary mean values c_1, c_2, and c_3.

A second level of approximation is based on the *elastica* curve model (the upper-left panel in Figure 3.1). That is, we shall interpolate the boundary end points by elasticas instead of by straight line segments. We can improve the inpainting accuracy by further approximating each individual (nonedge) isophote by an elastica for the segmented regions, instead of simply making constant approximations. Therefore, we need first to inpaint the missing edges including corners and T-junctions to reduce classes E, C, and T to class H.

Along the boundary, each end point can be represented by $(p, \vec{n})$, with p denoting its position and $\vec{n}$ the normal to the edge, which can be computed from the image available outside the inpainting domain K.

(a) For class E, to inpaint the missing smooth edge, we employ the elastica Γ that satisfies the boundary conditions $(p_1, \vec{n}_1)$ and $(p_2, \vec{n}_2)$.

(b) For class C, to inpaint the corner, we take a *moving frame* approach. The corner is represented by an affine frame $(p; \vec{n}_1', \vec{n}_2')$, with p denoting its unknown position, and $\vec{n}_1'$ and $\vec{n}_2'$ the two unknown unit normals to the smooth edges coming from the boundary end points $(p_1, \vec{n}_1)$ and $(p_2, \vec{n}_2)$. For each i, the energy (2.1) of the elastica that meets the requirement at $(p_i, \vec{n}_i)$ and $(p, \vec{n}_i')$ is denoted by

$$E_2((p_i, \vec{n}_i), (p, \vec{n}_i')).$$

Then the corner is inpainted by a joint optimization:

$$\min_{(p;\vec{n}_1',\vec{n}_2')} E_2((p_1, \vec{n}_1), (p, \vec{n}_1')) + E_2((p_2, \vec{n}_2), (p, \vec{n}_2')). \tag{3.1}$$

For example, if in the elastica energy (2.1) we choose a large ratio b/a (or $a = 0$ for the extremal case), then the solution to (3.1) shall be very close to the straight line shooting method mentioned in the first level of approximation.

(c) For class T, we first inpaint the two coupled end points by an elastica Γ (provided that we have established such coupling from other models or algorithms, as pointed out in the second paragraph). To inpaint the occluded edge from the clue of the third boundary end point $(p_3, \vec{n}_3)$, we again take the *moving frame* approach. Let $(p, \vec{n}_3')$ denote the junction position and the normal direction of the occluded edge. Then the junction inpainting is completed by solving

$$\min_{(p,\vec{n}_3')} E_2((p_3, \vec{n}_3), (p, \vec{n}_3')). \tag{3.2}$$

Here an admissible p must stay on the disoccluded edge Γ.

3.3. Local inpainting by individually engineering the isophotes. After the feature edges have all been interpolated, all four classes of local image inpainting are essentially reduced to the inpainting of class H, the homogeneous patches. In the same fashion as above, such patches can be inpainted by having the broken isophotes interpolated by elasticas *one by one* from the boundary information. This is exactly the idea underlying Masnou and Morel's dynamical programming algorithm [27].

Generically, one can assume that the missing smooth patch $u_0\big|_K$ is *regular* in the sense that it lies close to a regular point where ∇u_0 is nonzero (or by first applying a small step of Gaussian diffusion). Thus the isophotes of u_0 on K are well defined and distinguishable, and each Γ_λ is uniquely labeled by its gray level $u_0 \equiv \lambda$.

The trace of each Γ_λ on the boundary tells the coupling rule of boundary pixels. Suppose that $p_1, p_2 \in \partial K$ share the same gray level λ, and that the normals computed from the available image data outside K are $\vec{n}_1$ and $\vec{n}_2$. Then we inpaint the λ-isophote Γ_λ by an elastica Γ_λ':

$$\Gamma_\lambda' = \operatorname*{argmin}_{\gamma_\lambda \vdash ((p_1,\vec{n}_1),(p_2,\vec{n}_2))} \int_{\gamma_\lambda} (a + b\kappa^2) ds = \operatorname*{argmin}_{\gamma_\lambda \vdash ((p_1,\vec{n}_1),(p_2,\vec{n}_2))} E_2[\gamma_\lambda], \tag{3.3}$$

where $\vdash$ means subject to the boundary conditions: γ_λ goes through p_1 and p_2, and $\dot{\gamma}_\lambda \perp \vec{n}_i$ at the two ends. Notice that generally Γ_λ' does exist but may not be unique (see [22, 29]).

As λ varies according to the available boundary data u_0, (3.3) gives a family of (and theoretically infinitely many) elasticas. On the other hand, if we denote this

bundle of elasticas by

$$\mathcal{F}' = \{\Gamma'_\lambda \ : \ 0 \le \lambda \le 1\},$$

then it is easy to see that $\mathcal{F}'$ is also the minimizer of the following energy for all boundary admissible curve bundles $\mathcal{F} = \{\gamma_\lambda \ : \ 0 \le \gamma \le 1\}$:

$$E[\mathcal{F}] = \int_0^1 E_2[\gamma_\lambda]\, d\lambda \tag{3.4}$$

or, more generally,

$$E_w[\mathcal{F}] = \int_0^1 w(\lambda) E_2[\gamma_\lambda]\, d\lambda, \tag{3.5}$$

with some positive weight function $w(\lambda)$ (whose influence in application will be explained later). However, due to the lack of communication among the elasticas, there exist two potential problems:

(1) two different elastica interpolants Γ'_λ and Γ'_μ with $\lambda \neq \mu$ can meet inside the inpainting domain K, while the original isophotes never met;
(2) even putting problem (1) on hold, generally, it is not guaranteed that the elastica bundle

$$\mathcal{F}' = \{\Gamma'_\lambda \ : \ 0 \le \lambda \le 1\}$$

does "weave" the entire inpainting domain K and leave no "holes." Thus, the inpainting can still be incomplete.

These issues have been taken care in Masnou and Morel's algorithm [27]. Here we propose to resolve these issues by working with the level-set function u_K. (A similar philosophy has now made the *level-set method* of Osher and Sethian [34] a great success in the numerical computation of various interface motion problems.) An admissible curve bundle $\mathcal{F} = \{\gamma_\lambda\}_\lambda$, which not only satisfies the boundary conditions but also avoids the two problems, is uniquely and fully characterized by an inpainting function u_K that is "tangent" to u_0 along ∂K. Conversely, working with u_K instead of the individual isophotes automatically resolves these problems. Thus, first we need to translate the elasticity energies (3.4) or (3.5) into ones that are directly applicable to the inpainting function u_K.

4. The elastica inpainting model.

4.1. The functionalized elastica energy. Let $u = u_K$ be an admissible inpainting, assumed to be smooth enough so that all the conventional differentiations make sense in the following computation. Along any isophote $\gamma_\lambda : u \equiv \lambda$, the curvature of the oriented curve is given by

$$\kappa = \nabla \cdot \vec{n} = \nabla \cdot \left(\frac{\nabla u}{|\nabla u|} \right).$$

On the other hand, let dt denote the length element along the normal direction $\vec{n}$ (or along the steepest ascent integral curve). Then

$$\frac{d\lambda}{dt} = |\nabla u| \qquad \text{or} \qquad d\lambda = |\nabla u|\, dt.$$

Therefore, the integrated elastica energy (3.5) now passes on to u by

$$J[u] = E_w[\mathcal{F}] \tag{4.1}$$

$$= \int_0^1 w(\lambda) \int_{\gamma_\lambda : u=\lambda} (a + b\kappa^2) ds \, d\lambda \tag{4.2}$$

$$= \int_1^0 \int_{\gamma_\lambda : u=\lambda} w(u) \left(a + b \left(\nabla \cdot \frac{\nabla u}{|\nabla u|} \right)^2 \right) |\nabla u| dt \, ds \tag{4.3}$$

$$= \int_K w(u) \left(a + b \left(\nabla \cdot \frac{\nabla u}{|\nabla u|} \right)^2 \right) |\nabla u| \, dx, \tag{4.4}$$

since dt and ds represent a couple of orthogonal length elements. Now the energy is completely expressed in terms of the inpainting u itself. Notice that this formal derivation is much like the coarea formula for BV functions [18].

The weight function $w(\lambda)$ can be set to 1. In applications, it can be defined based on the histogram $h(\lambda)$ of the given image. ($h(\lambda)$ denotes the frequency of pixels with gray value λ.) The histogram of an image typically contains several "humps," each of which corresponds to an object. Since human observers are very sensitive to the regularity of object boundaries, we may assign a high weight to the pixels whose gray values are typically near the "valleys" of the histogram. Therefore we may choose the weight function in the form of

$$w(\lambda) = W(1 - h(\lambda)),$$

with $W(h)$ denoting a suitable positive and increasing function.

4.2. Admissible inpainting and the weak form of curvature. From now on, let us consider the functionalized Euler's elastica energy

$$J_2[u] = \int_K \left(a + b \left(\nabla \cdot \frac{\nabla u}{|\nabla u|} \right)^2 \right) |\nabla u| \, dx, \tag{4.5}$$

with the conditions that

$$u\big|_{\Omega \setminus K} = u_0\big|_{\Omega \setminus K}, \qquad \int_{\partial K} |Du| = 0, \qquad \text{and } |\kappa(p)| < \infty \quad \text{a.e. along } \partial K, \tag{4.6}$$

where a.e. is in the sense of 1-D Hausdorff measure.

We have assumed that the original complete image u_0 (typically on a square domain Ω) belongs to BV(Ω) and has the property that

$$\int_{\partial K} |Du_0| = 0 \tag{4.7}$$

in the sense of the Radon measure $\int |Du_0|$. Under such an assumption, the second boundary condition on u follows naturally. This condition can be made more explicit by the *trace* of BV functions [18]. Let u^- and u^+ denote the interior and exterior traces of u along ∂K with respect to K. Then we have

$$\int_{\partial K} |Du| = \int_{\partial K} |u^+ - u^-| \, d\mathcal{H}_1.$$

Thus the second condition is equivalent to the continuity condition

$$u^- = u^+ = u_0^+ \qquad \text{a.e. along } \partial K \text{ by } d\mathcal{H}_1.$$

We shall call assumption (4.7) on the original complete image the *feasibility condition* for low-level inpainting models, which do not employ global feature recognition or learning. It requires that there be no *essential* overlap between the boundary of the inpainting domain K and the edges of 2-D objects in the image. Imagine, in contrast, the opposite situation, in which an object is completely missing along its boundary. Then no existing image information can possibly bring it back in the absence of high-level intelligence (such as face recognition and symmetry detection). A low-level inpainting model, after all, is expected to interpolate incomplete objects based only on the hints they leave on the exterior of the inpainting domain.

Finally, the last condition in (4.6) demands finite curvatures along the inpainting boundary. Therefore a sudden turn of isophotes is not permitted along ∂K, and the condition is therefore a first-order continuity constraint.

Although the concept of "curvature" for a BV function has been used for both the elastica energy and the boundary conditions, its meaning has very much stayed at a formal level, since an average BV function lacks the necessary regularity for discussing curvatures in the conventional sense.

Therefore we introduce the concept of *weak curvature*, which may not be the only possible generalization but seems to be general enough to serve image analysis.

Suppose $u \in \mathrm{BV}(K)$. Then $|Du|(\cdot) = \int_\cdot |Du|$ is a finite Radon measure on K, and for any open subset $Q \subset K$,

$$\int_Q |Du| = \sup_{\mathbf{g} \in C_0^1(Q, B_1)} \int_Q u \, \nabla \cdot \mathbf{g} \, dx,$$

where B_1 denotes the unit ball centered at the origin in R^2. Let $\mathrm{supp}(|Du|)$ denote the support of the TV measure. Then for any $p \in \mathrm{supp}(|Du|)$ and any of its open neighborhoods N_p,

$$|Du|(N_p) = \int_{N_p} |Du| > 0.$$

Let ρ be a fixed radially symmetric nonnegative mollifier with compact support and unit total integral, and set (for 2-D)

$$\rho_\sigma = \frac{1}{\sigma^2} \rho\left(\frac{x}{\sigma}\right) \qquad \text{and} \qquad u_\sigma = \rho_\sigma * u.$$

Then we define the *weak absolute curvature* $\tilde{\kappa}(p)$ of u at p by

$$\tilde{\kappa}(p) = \limsup_{\sigma \to 0} \left| \nabla \cdot \left(\frac{\nabla u_\sigma}{|\nabla u_\sigma|} \right)(p) \right|, \tag{4.8}$$

where for those σ's that give $|\nabla u_\sigma(p)| = 0$ we define $\nabla \cdot (\nabla u_\sigma / |\nabla u_\sigma|)$ to be ∞. Finally, for any pixel p outside $\mathrm{supp}(|Du|)$, we assign 0 to $\tilde{\kappa}(p)$, since u is a.e. constant near a neighborhood of p. Thus the weak absolute curvature is well defined everywhere for an arbitrary BV function.

There are two important situations in image analysis in which the weak curvature is indeed the ordinary curve curvature for $p \in \mathrm{supp}(|Du|)$. The first situation is presented as follows.

PROPOSITION 4.1. *Suppose that $u \in C^2(K)$ and that $p \in K$ is a regular pixel: $\nabla u(p) \neq \mathbf{0}$. Then $\tilde{\kappa}(p) = |\kappa(p)|$.*

Proof. Assume that the mollifier is supported on the unit ball B_1. From the definition of convolution

$$u_\sigma(q) = \rho_\sigma * u(q) = \int_{B_1} \rho(y) u(q + \sigma y)\, dy,$$

it is easy to see that there is a small neighborhood N_p and some positive number a so that $u_\sigma(q)$ is C^2 over $(\sigma, q) \in (-a, a) \times N_p$. Since ∇u is continuous and nonvanishing at p, we can further refine N_p and a so that all $\nabla u(q + \sigma y)$ (with $(\sigma, q, y) \in (-a, a) \times N_p \times B_1$) are concentrated enough around $\nabla u(p)$. Then, thanks to the averaging property of the mollifier, all $\nabla u_\sigma(q) = (\nabla u)_\sigma(q)$ are nonvanishing, which makes

$$\vec{n}_\sigma(q) = \frac{\nabla u_\sigma(q)}{|\nabla u_\sigma(q)|}$$

C^1 on $(-a, a) \times N_p$. Then $\kappa_\sigma(q) = \nabla \cdot \vec{n}_\sigma(q)$ is well defined on N_p and continuous in σ, especially at p:

$$|\kappa(p)| = \lim_{\sigma \to 0} |\kappa_\sigma(p)| = \tilde{\kappa}(p). \qquad \square$$

The second case occurs when p lies on an intensity edge between two objects.

PROPOSITION 4.2. *Suppose an oriented curve segment γ is a C^2 submanifold in K. Assume that near a given pixel $p \in \gamma$, on one side of γ, $u = c^+$, and on the other, $u = c^-$, two constant gray values. Then $\tilde{\kappa}(p) = |\kappa(p)|$.*

Proof. Since curvature is a second-order local feature, we can replace γ near p by the curvature circle with radius $r = 1/|\kappa(p)|$ and centered at $q = p - r\vec{n}$, where $\vec{n}$ is one of the unit normal vectors at p. Then both the data (c^+, c^-) and the geometry γ are locally rotationally invariant (with respect to the center) in a neighborhood of p. Since the mollifier ρ is radially symmetric and compactly supported, as long as σ is small enough, $u_\sigma = \rho_\sigma * u$ must also be locally rotationally invariant with respect to the center, which means that locally near p, γ is also an isophote of u_σ. Thus under the same orientation of γ, $\kappa_{\sigma \to 0}(p) = \kappa(p)$ and $\tilde{\kappa}(p) = \lim_\sigma |\kappa_\sigma(p)| = |\kappa(p)|$. This completes the proof. $\square$

One useful property of $\tilde{\kappa}$ which can be proven easily is its invariance under the linear scaling of gray values. Let $\tilde{\kappa}_f(p)$ denote the weak curvature of a function f at p.

PROPOSITION 4.3. *Let $u \in BV(K)$ and $v = a + bu$ for some constants a and $b \neq 0$. Then for any $p \in K$, $\tilde{\kappa}_u(p) = \tilde{\kappa}_v(p)$.*

With the help of the concept of weak curvature, the functionalized elastica energy (4.5) can be rigorously defined. A function $u \in \mathrm{BV}(K)$ is said to be *admissible* if

$$\tilde{\kappa} \in L^2(K, |Du|).$$

For all such functions, the generalized elastica energy

$$J_2[u] = \int_K (a + b\tilde{\kappa}^2) |Du| \tag{4.9}$$

is well defined and finite. Together with the boundary conditions (4.6), it defines the so-called *elastica inpainting model*.

The elastica inpainting model is difficult to analyze in terms of existence and uniqueness, due to the nonconvexity of the energy and the involvement of curvature. However, there is indeed a special, simple, yet very useful case for which one can carry out the analysis successfully. This is the TV inpainting model first proposed and implemented in Chan and Shen [8]. In the next subsection, we study the existence and uniqueness of this special case.

4.3. The TV inpainting model of Chan and Shen. The TV inpainting model of Chan and Shen [8] is an extreme case of elastica inpainting in which we weigh highly against the total variation, i.e., $a/b = \infty$. Thus one is led to the minimization of

$$\mathrm{TV}(u) = \int_\Omega |Du|. \tag{4.10}$$

As in the study of minimal surfaces (De Giorgi [17]), the suitable companion condition becomes

$$u\big|_{\Omega\setminus K} = u_0\big|_{\Omega\setminus K}, \tag{4.11}$$

where Ω denotes the entire image domain, generally assumed to be a bounded Lipschitz domain. As in the study of minimal surfaces, Ω can be replaced by any open neighborhood of $\bar{K}$ (see [8]). We call the combination of (4.10) and (4.11) the *noise-free TV inpainting model*.

THEOREM 4.4 (existence of a noise-free TV inpainting). *Suppose that the original complete image u_0 lies in $BV(\Omega)$ and takes gray values between 0 (black) and 1 (white). Then the noise-free TV inpainting model* (4.10) *and* (4.11), *together with the gray value constraint $u \in [0,1]$, has at least one optimal inpainting.*

Proof. Since the original complete image u_0 is admissible, we can find a minimizing sequence of admissible inpaintings $(u_n)_n$. Then both

$$\int_\Omega |Du_n| \qquad \text{and} \qquad \int_\Omega |u_n(x)|dx$$

are bounded because Ω is, and all u_n's take values from $[0,1]$. By the weak compactness of BV functions, there is a subsequence, still denoted by $(u_n)_n$ for convenience, that strongly converges to some $u_{\mathrm{tv}} \in L^1(\Omega)$ in the L^1 norm. Apparently u_{tv} still meets the constraints

$$u_{\mathrm{tv}}\big|_{\Omega\setminus K} = u_0\big|_{\Omega\setminus K} \qquad \text{and} \qquad u_{\mathrm{tv}}(x) \in [0,1].$$

By the lower semicontinuity of the TV measure with respect to the L^1 convergence,

$$\int_\Omega |Du_{\mathrm{tv}}| \le \liminf_{n\to\infty} \int_\Omega |Du_n| = \min_u \int_\Omega |Du|.$$

Thus u_{tv} must be a minimizer. ∎

Another important issue for inpainting is how to deal with noise, since in applications (such as the restoration of degraded photos or films), the available part of the image $u_0\big|_{\Omega\setminus K}$ is often noisy or contains misleading outliers (especially along the damaged boundary of K). Chan and Shen [8] therefore modified the above TV inpainting model by having the constraint (4.11) replaced by the denoising term

$$\frac{1}{\mathrm{Area}(\Omega\setminus K)} \int_{\Omega\setminus K} (u-u_0)^2 dx = \sigma^2, \tag{4.12}$$

where σ^2 is the variance of the homogeneous white noise, which can be estimated from $u_0\big|_{\Omega\setminus K}$ by suitable statistical estimators. By such a constraint, we are assuming that embedded within u_0 is a clean image u_c such that

$$u_0(x) = u_c(x) + n(x),$$

where the white noise $n(x)$ is independent of u_c. Since we pay attention only to its second-order statistics, implicitly we are assuming that $n(x)$ can be well approximated by the Gaussian $N(0, \sigma^2)$.

THEOREM 4.5 (existence of a TV inpainting for a noisy image). *Given an image observation $u_0 : u_0(x) \in [0,1]$ on $\Omega \setminus K$, assume that there exists at least one image $u_c : u_c(x) \in [0,1]$ in $BV(\Omega)$ that meets the denoising constraint* (4.12). *Then there exists at least one optimal TV inpainting in $BV(\Omega)$ that satisfies both the denoising requirement* (4.12) *and the gray scale constraint $u \in [0,1]$.*

Proof. From the assumption on u_c, there exists a minimizing sequence of admissible inpainting $(u_n)_n$ that meets both the denoising constraint and the gray scale constraint. Thanks to the gray scale constraint, $(u_n)_n$ must be bounded in $\mathrm{BV}(\Omega)$. Thus there is subsequence, still denoted by $(u_n)_n$ for convenience, which converges in the L^1 norm to some $u_{\mathrm{tv}} \in L^1(\Omega)$. Then by the lower semicontinuity property [18],

$$\int_\Omega |Du_{\mathrm{tv}}| \le \liminf_{n\to\infty} \int_\Omega |Du_n| = \min_u \int_\Omega |Du|.$$

We can further refine the subsequence so that

$$u_n \to u_{\mathrm{tv}} \qquad \text{a.e. on } \Omega.$$

Thus u_{tv} also meets the gray scale constraint $u_{\mathrm{tv}} \in [0,1]$, and more importantly, by the Lebesgue dominated convergence theorem,

$$\int_{\Omega\setminus K} (u_{\mathrm{tv}} - u_0)^2 dx = \lim_n \int_{\Omega\setminus K} (u_n - u_0)^2 dx.$$

Therefore, u_{tv} is indeed an optimal TV inpainting, subject to the two constraints imposed. ∎

Remark 2. If we drop the assumption that $u_0 \in [0,1]$, and thus remove the gray scale constraint on u, then under the natural assumption (see [7]) that

$$\sigma^2 \le \frac{1}{\mathrm{area}(\Omega \setminus K)} \int_{\Omega\setminus K} (u_0 - \langle u_0 \rangle)^2 dx < \infty,$$

one can still establish the existence theorem by applying Friedrich's trace inequality [24] and Fatou's lemma. (Here $\langle u_0 \rangle$ is the mean value of u_0 over the integration domain.)

The solutions to both TV and elastica inpaintings can be nonunique. In our opinion, such nonuniqueness of the models should not be cursed but appreciated, since it is an important characteristic of the inpainting problem itself. Take, for example, the image in Figure 4.1, whose middle square patch has been encrypted by a random image. We now try to inpaint the square to restore the original complete image.

It seems that we have a black ($u = 0$) bar and a white ($u = 1$) one against a gray background ($u = 1/2$). A perceptually meaningful inpainting would be to fill in either

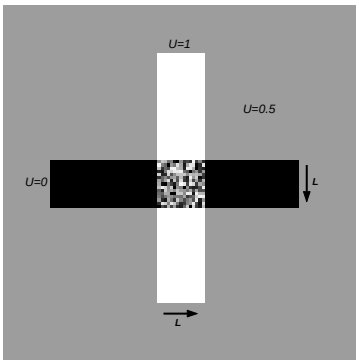


FIG. 4.1. *Nonuniqueness: which is to blame, the model or the problem itself?*

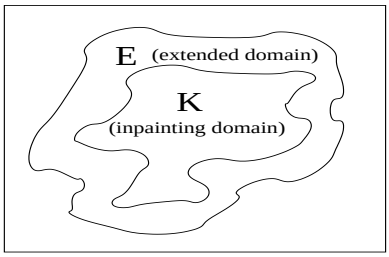


FIG. 4.2. *Inpainting Domain K.*

the black color so that the image shows a black bar occludes a farther white bar, or the white color for the opposite occlusion. Which one is more likely? The available part of the image (both the geometry and colors) is perfectly symmetric. Thus it appears as a half-half situation to human perception. In our opinion, such perceptual uncertainty is the foundation of the nonuniqueness of inpainting models. In terms of the Bayesian decision theory in vision analysis, nonuniqueness corresponds to the situation in which the risk or cost function has many similar valleys which compete with each other.

Fortunately, in many real applications, as Chan and Shen demonstrated in [8, 9], the outputs from the TV inpainting model do always seem to be meaningful to human vision. This is because, in most applications, the location, shape, and size of inpainting domains are often randomly distributed.

4.4. Relaxation of the constraints. As we have seen, the original formulation of the elastica inpainting model (4.5) and (4.6) has mainly been inspired by the image processing point of view. But it is unclear whether the formulation is mathematically feasible. In fact, as well practiced in the theory of BV functions and minimal surfaces [18], it is often more manageable to formulate the problem on a larger domain than the original one. The TV inpainting model for noisy images mentioned above has imparted this idea. The same can be done for general elastica inpaintings.

Let E be a subset contained in $\Omega \setminus K$ such that $E \cup K$ is open and contains the closure of K (see Figure 4.2). For example, depending on the situation, one could

simply take $E = \Omega \setminus K$. Suppose for the original image u_0,

$$J_2[u_0] = \int_{E\cup K} (a + b\tilde{\kappa}^2)\ |Du_0|$$

is finite. We inpaint $u_0\big|_K$ by the minimizer of

$$J_2[u] = \int_{E\cup K} (a + b\tilde{\kappa}^2)\ |Du|, \tag{4.13}$$

with the condition that

$$u\big|_E = u_0\big|_E. \tag{4.14}$$

In this way, the two other original boundary conditions in (4.6) are approximately built into the energy itself.

If the available part of the image in the vicinity of K has been corrupted by homogeneous white noise with variance σ^2, then the condition (4.14) is further replaced by the fitting constraint

$$\frac{1}{\text{area}(E)} \int_E |u - u_0|^2\ dx = \sigma^2. \tag{4.15}$$

PROPOSITION 4.6. *Let $u \in BV(E \cup K)$ be a minimizer of the elastica inpainting* (4.13) *and* (4.15). *Then u automatically satisfies the mean value constraint*

$$\langle u \rangle = \langle u_0 \rangle,$$

where $\langle f \rangle$ denotes the mean value of f over E.

Proof. The technique is similar to that used by Chambolle and Lions [7]. Assume the contrary: $\langle u \rangle \neq \langle u_0 \rangle$. Define $v = u - \langle u - u_0 \rangle$. Then

$$\int_E |v - u_0|^2 dx = \int_E |u - u_0 - \langle u - u_0 \rangle|^2 dx < \int_E |u - u_0|^2 dx = \sigma^2\ \text{area}(E),$$

where the strict inequality is due to the fact that, among all constants, the mean is the best L^2 fitting to a given signal. On the other hand, by the natural assumption of u_0 on E in Remark 2 of section 4.3,

$$\int_E |\langle v \rangle - u_0|^2 dx = \int_E |\langle u_0 \rangle - u_0|^2 dx \geq \sigma^2\ \text{area}(E).$$

Therefore, there must exist some $s \in [0, 1)$ such that

$$\int_E |sv + (1 - s)\langle v \rangle - u_0|^2 dx = \sigma^2\ \text{area}(E).$$

Define $w = sv + (1 - s)\langle v \rangle$. By the invariant property of the weak curvature in Proposition 4.3, we have

$$J_2[w] = \int_{E\cup K} (a + b\tilde{\kappa}^2)|Dw| = s \int_{E\cup K} (a + b\tilde{\kappa}^2)|Du| < \int_{E\cup K} (a + b\tilde{\kappa}^2)|Du| = J_2[u],$$

where the strict inequality is because u cannot be a constant (otherwise the constant must be $\langle u_0 \rangle$ due to the fitting constraint, which is impossible since the whole

argument starts with $\langle u \rangle \neq \langle u_0 \rangle$). This eventually contradicts the fact that u is a minimizer. ∎

Finally, even the fitting constraint (4.15) can be built into the energy functional by minimizing

$$J_2^\lambda[u] = \int_{E \cup K} (a + b\tilde{\kappa}^2)|Du| + \frac{\lambda}{2} \int_E (u - u_0)^2 dx. \tag{4.16}$$

It can also easily be shown that a minimizer of $J_2^\lambda[u]$ automatically satisfies the mean value constraint.

The last formulation bears a formal Bayesian interpretation as Mumford did for various segmentation models [30]. From the probability point of view, the conditional probability

$$P(u_0|u) = \text{const.} \exp\left(-\frac{\lambda}{2} \int_E (u - u_0)^2 dx\right)$$

is the *generative data model*, and the probability

$$P(u) = \text{const.} \exp(-J_2[u])$$

is the *prior model*. Together, the minimization of $J_2^\lambda[u]$ corresponds to the method of MAP, or *maximum a posteriori probability*, in the theory of statistical inference and decision making.

4.5. Local analysis near a generic stationary point: The effect of the curvature power p and $p = 3$. Generally, for any $p \geq 1$, one could consider the p-elastica energy

$$J_p[u] = \int_\Omega (a + b|\kappa|^p)|Du|,$$

and, if necessary, $|\kappa|$ is replaced by the weak absolute curvature $\tilde{\kappa}$. This general form of elasticity energy was also mentioned in Masnou and Morel [27]. So the question arises naturally: is there any essential difference among all the different choices of p's? We claim that, indeed, in some sense $p = 3$ is the threshold.

THEOREM 4.7. *Suppose that u is C^2 near a* generic *stationary pixel z, i.e.,*

$$\nabla u(z) = 0 \quad \textit{but the Hessian } H_u(z) \textit{ is nonsingular.}$$

Then for all $p \geq 3$, $J_p[u] = \infty$.

Therefore, generic stationary points are forbidden by the p-elasticity energy when $p \geq 3$.

Proof. Without loss of generality, assume that $z = (0, 0)$ and $u(z) = 0$. Since curvature is a second-order feature, we can assume that u coincides with its second-order Taylor expansion at z:

$$u(x) = u(x_1, x_2) = (x_1, x_2) A (x_1, x_2)^T,$$

where A is the nonsingular Hessian $H_u(z)$. Thus A must be either elliptic or hyperbolic. Take the elliptic case, for example. Since both $|\nabla u|$ and κ are invariant under Euclidean transforms, we can assume that

$$A = \text{diag}(\sigma_1^2, \sigma_2^2),$$

with $\sigma_1 \geq \sigma_2 > 0$. First we consider the case in which $\sigma_1 = \sigma_2$. For convenience, we can assume that $\sigma_1 = 1$. Then $u = r^2 = x_1^2 + x_2^2$, and by section 4.1,

$$\begin{aligned}\int_{B_1} \kappa^p |\nabla u| dx &= \int_0^1 d\lambda \left(\int_{u=\lambda} \kappa^p ds \right) \\ &= \int_0^1 2r dr \left(2\pi r \left(\frac{1}{r} \right)^p \right) \\ &= 4\pi \int_0^1 \left(\frac{1}{r} \right)^{p-2} dr.\end{aligned}$$

Here, for convenience, we have assumed that the C^2 neighborhood encloses B_1. Thus J_p on B_1 is finite if and only if $p < 3$, and as a result, when $p \geq 3$, J_p on Ω must blow up. The general case follows easily from the fact that along any ellipse isophote

$$\sigma_1^2 x_1^2 + \sigma_2^2 x_2^2 = \lambda = r^2,$$

the curvature is bounded by

$$\frac{\sigma_1}{r} \geq \kappa \geq \frac{\sigma_2}{r}.$$

This completes the proof. □

The $p = 3$ threshold has also appeared in the theoretical work by Bellettini, Dal Maso, and Paolini [2] on boundary elastica energies. It is quite controversial whether we should really encourage generic stationary points for general image interpolation problems. On the one hand, if the p-elastica energy is also imposed outside the inpainting domain K for denoising purposes (as in (4.16) below), then generic stationary pixels should be allowed, and consequently $p < 3$ is required. On the other hand, existing interpolation mechanisms have seemed to generally discourage the emergence of local minima or maxima over the missing domain K. For example, the axiomatic interpolation approach of Caselles, Morel, and Sbert [6] has emphasized the constant gradient condition (i.e., $|\nabla u| = \text{const.}$) along any normal integral line, while the harmonic inpainting model discussed in Chan and Shen [8] is also generically against the existence of a local minimum or maximum, due to the maximum principle of harmonic functions. This issue needs more study.

5. The Euler–Lagrange equation. For the general elastica inpainting model (4.13) or (4.16), the direct method is difficult due to the involvement of the geometric quantity, curvature, which has no linear structure. That is, one cannot say much about the curvature κ_{u+v} of the summation $u + v$, even when precise information on κ_u and κ_v is available. Such an obstacle invalidates the classical linear approach based on Sobolev spaces or the BV space. For example, it is unclear how to prove that the minimizer to (4.16) actually exists.

In this situation, as is well practiced in the PDE method in image processing, one is led to the study of the formal Euler–Lagrange equation. Often the PDEs can handle geometry more explicitly than the variational formulation itself, as in the case of *mean curvature motions* [28].

In this section, we first derive the formal Euler–Lagrange equation for the fitted elastica inpainting model (4.16). We then show that the geometric meaning of the equation unifies the early method of Bertalmio, et al. [3] based on transport PDEs, and that of Chan and Shen [9] based on CDD. We conjecture that transport and CDD are the two universal mechanisms for any low-level nontexture inpainting.

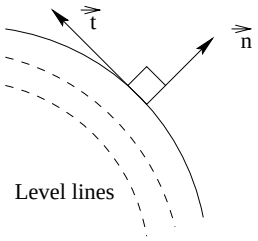


FIG. 5.1. *The normal* $\vec{n} = \nabla u/|\nabla u|$ *and the tangent* $\vec{t}$.

5.1. Derivation of the Euler–Lagrange equation. In the formal derivation, we shall always assume that the image is smooth enough and that the curvature is well defined.

THEOREM 5.1. *Let* $\phi \in C^1(R, (0,\infty))$ *be given and*

$$R[u] = \int_{E\cup K} \phi(\kappa)|\nabla u|\, dx.$$

Then the gradient descent time marching is given by

$$\frac{\partial u(x,t)}{\partial t} = \nabla \cdot \vec{V}(x,t), \qquad x \in E \cup K, \quad t > 0,$$

with the boundary conditions along $\partial(E \cup K)$

$$\frac{\partial u}{\partial \vec{\nu}} = 0 \qquad \textit{and} \qquad \frac{\partial(\phi'(\kappa)|\nabla u|)}{\partial \vec{\nu}} = 0 \tag{5.1}$$

($\vec{\nu}$ *denotes the outer normal of the boundary). The flux field* $\vec{V}$ *is given by*

$$\vec{V} = \phi(\kappa)\, \vec{n} - \frac{\vec{t}}{|\nabla u|} \frac{\partial\, (\phi'(\kappa)|\nabla u|)}{\partial \vec{t}}. \tag{5.2}$$

Here $\vec{n}$ *is the ascending normal field* $\nabla u/|\nabla u|$, *and* $\vec{t}$ *the tangent field (whose exact orientation does not matter due to the parity of* $\vec{t}$ *in the expression) (see Figure* 5.1*).*

Proof. For later convenience, we write $\langle f\rangle = \int_{E\cup K} f dx$, and the boundary integral $\langle f\rangle_\partial = \int_{\partial(E\cup K)} f ds$, with ds denoting the Euclidean arc-length element. Then the variation of $R \to R + \delta R$ is by

$$\begin{aligned}
\delta R &= \langle \delta(\phi(\kappa)|\nabla u|)\rangle \\
&= \langle \phi(\kappa)\ \delta|\nabla u|\rangle + \langle |\nabla u|\ \delta\phi(\kappa)\rangle \\
&= \left\langle \phi(\kappa)\frac{\nabla u}{|\nabla u|}\cdot \delta\nabla u \right\rangle + \langle \phi'(\kappa)|\nabla u|\ \delta\kappa\rangle \\
&= -\langle \nabla\cdot(\phi(\kappa)\vec{n})\ \delta u\rangle + \langle \phi'(\kappa)|\nabla u|\ \delta\kappa\rangle.
\end{aligned}$$

Here in the last line, in order to justify the drop of the boundary integral from the divergence theorem (or integration by parts)

$$\left\langle \frac{\phi(\kappa)}{|\nabla u|}\frac{\partial u}{\partial \vec{\nu}}\ \delta u \right\rangle_\partial,$$

we impose the first boundary condition by noticing that $\phi(\kappa) > 0$:

$$\frac{\partial u}{\partial \vec{\nu}} = 0 \qquad \text{along} \quad \partial(E \cup K). \tag{5.3}$$

Next, we need to work out the variation of the curvature:

$$\begin{aligned}
\delta\kappa = \delta\left(\nabla \cdot \frac{\nabla u}{|\nabla u|}\right) &= \nabla \cdot \left[\frac{1}{|\nabla u|}\,\nabla(\delta u) + \nabla u\;\delta\left(\frac{1}{|\nabla u|}\right)\right] \\
&= \nabla \cdot \left[\frac{1}{|\nabla u|}\nabla(\delta u) - \frac{\nabla u}{|\nabla u|^3}\,(\nabla u \cdot \nabla(\delta u))\right] \\
&= \nabla \cdot \left[\frac{1}{|\nabla u|}\{I - \vec{n} \otimes \vec{n}\}\nabla(\delta u)\right].
\end{aligned}$$

Here I denotes the identity transform, and $P = \vec{n} \otimes \vec{n}$ the orthogonal projection onto the normal direction. Noticing that $I = \vec{n} \otimes \vec{n} + \vec{t} \otimes \vec{t}$, we have

$$\begin{aligned}
\langle \phi'(\kappa)|\nabla u|\;\delta\kappa\rangle &= \left\langle \phi'(\kappa)|\nabla u|\;\nabla \cdot \left[\frac{1}{|\nabla u|}\{\vec{t} \otimes \vec{t}\}\nabla(\delta u)\right]\right\rangle \\
&= \left\langle -\nabla(\phi'(\kappa)|\nabla u|) \cdot \left[\frac{1}{|\nabla u|}\{\vec{t} \otimes \vec{t}\}\nabla(\delta u)\right]\right\rangle.
\end{aligned}$$

We have dropped the boundary integral after applying the divergence theorem:

$$\left\langle \phi'(\kappa)\;\vec{\nu} \cdot \{\vec{t} \otimes \vec{t}\}\nabla(\delta u)\right\rangle_{\partial}.$$

This is well justified under the Neumann boundary condition (5.3) because, from (5.3), $\vec{\nu} \cdot \nabla u = 0$, or equivalently, $\vec{\nu} = \pm\vec{t}$, along $\partial(E \cup K)$. Therefore,

$$\begin{aligned}
\left\langle \phi'(\kappa)\;\vec{\nu} \cdot \{\vec{t} \otimes \vec{t}\}\nabla(\delta u)\;\right\rangle_{\partial} &= \langle \phi'(\kappa)\;\vec{\nu} \cdot \{\vec{\nu} \otimes \vec{\nu}\}\nabla(\delta u)\;\rangle_{\partial} \\
&= \langle \phi'(\kappa)\;\vec{\nu} \cdot \nabla(\delta u)\;\rangle_{\partial} \\
&= \left\langle \phi'(\kappa)\frac{\partial(\delta u)}{\partial \vec{\nu}}\right\rangle_{\partial} = \left\langle \phi'(\kappa)\delta\left(\frac{\partial u}{\partial \vec{\nu}}\right)\right\rangle_{\partial} \\
&= \langle \phi'(\kappa)\;\delta(0)\;\rangle_{\partial} = 0.
\end{aligned}$$

Let us now come back to the calculus of variation. Noticing that $\vec{t} \otimes \vec{t}$ is symmetric, we have

$$\begin{aligned}
\langle \phi'(\kappa)|\nabla u|\delta\kappa\rangle &= \left\langle -\{\vec{t} \otimes \vec{t}\}\left[\frac{1}{|\nabla u|}\nabla(\phi'(\kappa)|\nabla u|)\right] \cdot \nabla(\delta u)\right\rangle \\
&= \left\langle \nabla \cdot \{\vec{t} \otimes \vec{t}\}\left[\frac{1}{|\nabla u|}\nabla(\phi'(\kappa)|\nabla u|)\right]\;\delta u\right\rangle.
\end{aligned}$$

Here, to justify the drop of the boundary integral

$$\left\langle -\vec{\nu} \cdot \{\vec{t} \otimes \vec{t}\}\left[\frac{1}{|\nabla u|}\nabla(\phi'(\kappa)|\nabla u|)\right]\;\delta u\right\rangle_{\partial},$$

we require that, along $\partial(E \cup K)$,

$$\vec{\nu} \cdot \{\vec{t} \otimes \vec{t}\}\left[\frac{1}{|\nabla u|}\nabla(\phi'(\kappa)|\nabla u|)\right] = 0.$$

Since $\vec{\nu} = \pm\vec{t}$ along $\partial(E \cup K)$, we get the second boundary condition:

$$\frac{\partial(\phi'(\kappa)|\nabla u|)}{\partial\vec{\nu}} = \vec{\nu}\cdot\nabla(\phi'(\kappa)|\nabla u|) = 0 \qquad \text{along} \quad \partial(E\cup K). \tag{5.4}$$

Eventually, under the boundary conditions (5.3) and (5.4), we have

$$\begin{aligned}\delta R &= \left\langle -\delta u\ \nabla\cdot\left[\phi(\kappa)\ \vec{n} - \frac{1}{|\nabla u|}\{\vec{t}\otimes\vec{t}\}\nabla(\phi'(\kappa)|\nabla u|)\right]\ \right\rangle \\ &= \left\langle -\delta u\ \nabla\cdot\left[\phi(\kappa)\ \vec{n} - \frac{1}{|\nabla u|}\frac{\partial(\phi'(\kappa)|\nabla u|)}{\partial\vec{t}}\ \vec{t}\right]\ \right\rangle \\ &= \langle -\delta u\ \nabla\cdot\vec{V}\rangle.\end{aligned}$$

This completes the proof. □

Therefore, the gradient of $R[u]$ is in the divergence form. The vector field $\vec{V}$ shall be called the *flux field* in this paper. The theorem shows that the flux field has a natural decomposition in the normal and tangent fields. Moreover, it is *morphologically* invariant.

PROPOSITION 5.2. *The flux field $\vec{V}$ is morphologically invariant.*

Proof. Let g be any smooth morphological transform of gray scales:

$$u \to g(u), \qquad g'(u) > 0.$$

We show that for any image u the fluxes $\vec{V}_u = \vec{V}_{g(u)}$. Notice that κ, $\vec{n}$, and $\vec{t}$ are already morphologically invariant. Furthermore,

$$\begin{aligned}\frac{1}{|\nabla g(u)|}\frac{\partial(\phi'(\kappa)|\nabla g(u)|)}{\partial\vec{t}}\ \vec{t} &= \frac{1}{g'(u)|\nabla u|}\frac{\partial(g'(u)\phi'(\kappa)|\nabla u|)}{\partial\vec{t}}\ \vec{t} \\ &= \frac{g'(u)}{g'(u)|\nabla u|}\frac{\partial(\phi'(\kappa)|\nabla u|)}{\partial\vec{t}}\ \vec{t} \\ &= \frac{1}{|\nabla u|}\frac{\partial(\phi'(\kappa)|\nabla u|)}{\partial\vec{t}}\ \vec{t},\end{aligned}$$

where we have applied the fact that u, and therefore $g'(u)$, are both constant in the tangent direction. Thus $\vec{V}_{g(u)} = \vec{V}_u$. □

Masnou and Morel (private communication) have also worked out the Euler–Lagrange equation (5.2), although it is not expressed in the above geometric form.

COROLLARY 5.3. *For the elastica inpainting model* (4.16)*, the gradient descent marching is given by*

$$\frac{\partial u(x,t)}{\partial t} = \nabla\cdot\vec{V}(x,t) - \lambda_E(x)\ (u(x) - u_0(x)), \qquad x\in E\cup K,\ t>0,$$

with the boundary conditions (5.1) *and*

$$\vec{V} = (a + b\kappa^2)\ \vec{n} - \frac{2b}{|\nabla u|}\frac{\partial(\kappa|\nabla u|)}{\partial\vec{t}}\ \vec{t}, \tag{5.5}$$

$$\lambda_E(x) = \lambda\cdot 1_E(x) \qquad (1_E \text{ is the indicator of } E). \tag{5.6}$$

In numerical computation, as Marquina and Osher [25] proposed, the weighted gradient descent method generally converges faster than the original one:

$$\frac{\partial u}{\partial t} = |\nabla u|\ \nabla\cdot\vec{V} - |\nabla u|\ \lambda_E\ (u - u_0). \tag{5.7}$$

By such modification, without the fitting term, the evolution equation is morphologically invariant since the $g'(u)$ factors (associated with a morphological transform g) cancel each other out in $\partial u/\partial t$ and $|\nabla u|$, while the flux field $\vec{V}$ is already morphologically invariant. If $b = 0$, we have the well-known *mean curvature motion* [28].

Our numerical PDE scheme in the coming section is applied to the modified gradient descent equation (5.7).

5.2. The inpainting mechanisms of transport and diffusion. We now show that the flux field $\vec{V}$ beautifully offers a unified viewpoint on the earlier work of Bertalmio et al. on *transport*-based inpainting [3] and on that of Chan and Shen [8] on CDD-based inpainting. In return, these earlier works help reveal the fine structure of elastica inpainting from the PDE point of view.

The first high-order PDE-based inpainting model of Bertalmio et al. [3] is based on the beautiful intuition of smoothness transport along isophotes:

$$\frac{\partial u}{\partial t} = \nabla^{\perp} u \cdot \nabla L(u), \tag{5.8}$$

where $\nabla^{\perp} u = (-u_y, u_x) = |\nabla u|\ \vec{t}$, and $L(u)$ can be any smoothness measure of the image u. For example, in the numerical experiment of [3], L is chosen to be the Laplacian Δu. The model carries the transport nature since, as the evolution approaches its equilibrium state, we have

$$\vec{t} \cdot \nabla L(u) = 0 \qquad \text{and} \qquad \frac{\partial L(u)}{\partial \vec{t}} = 0,$$

which means that the smoothness measure remains constant along any completed isophote. Thus, in terms of the available boundary data, the image evolves as though transporting the boundary smoothness information along the restored isophotes into the inpainting domain.

However, due to the lack of communication among the isophotes, the transport may result in kinks inside the inpainting domain, just as shocks may develop in traffic models. Thus in [3], (5.8) is implemented with the help of intermediate steps of anisotropic diffusions. As we shall see below, such intuition is well supported by the elastica inpainting.

On the other hand, in [9], in order to realize the so-called *connectivity principle* in perceptual disocclusion, Chan and Shen proposed the CDD inpainting model

$$\frac{\partial u}{\partial t} = \nabla \cdot \left(\frac{g(\kappa)}{|\nabla u|} \nabla u \right), \tag{5.9}$$

where $g : R \to [0, +\infty)$ is a continuous function satisfying $g(0) = 0$ and $g(\pm\infty) = +\infty$. If g is replaced by 1, then this is the classical TV anisotropic diffusion. Here $g(\kappa)$ has been introduced to penalize large curvatures and encourage small ones, since physically $D = g(\kappa)/|\nabla u|$ denotes the diffusivity coefficient. With this action, the reconnection of objects which were broken by the inpainting domain is generally encouraged.

It has remained a mystery why we can have two seemingly *orthogonal* inpainting mechanisms: the model in [3] transports information *along* isophotes, while the CDD inpainting model [9] diffuses information *across*. We now explain that the elastica inpainting model makes a unification by incorporating both mechanisms.

We have established in Theorem 5.1 that the flux $\vec{V}$ for the inpainting energy $R[u]$ consists of two components: the normal part $\vec{V}_n = \phi(\kappa)\vec{n}$ and the tangential part

$$\vec{V}_t = -\frac{1}{|\nabla u|}\frac{\partial(\phi'(\kappa)|\nabla u|)}{\partial \vec{t}}\vec{t}.$$

The normal flux $\vec{V}_n$ precisely corresponds to Chan and Shen's CDD program (5.9) with $g(\kappa) = \phi(\kappa)$.

On the other hand, the tangential component can be written as

$$\vec{V}_t = -\left(\frac{1}{|\nabla u|^2}\frac{\partial(\phi'(\kappa)|\nabla u|)}{\partial \vec{t}}\right)\nabla^{\perp}u,$$

and its divergence as

$$\nabla \cdot \vec{V}_t = \nabla^{\perp}u \cdot \nabla\left(\frac{-1}{|\nabla u|^2}\frac{\partial(\phi'(\kappa)|\nabla u|)}{\partial \vec{t}}\right)$$

because $\nabla^{\perp}u$ is divergence-free. It corresponds to the scheme of Bertalmio et al. in the form (5.8), with the smoothness measure

$$L_\phi = \frac{-1}{|\nabla u|^2}\frac{\partial(\phi'(\kappa)|\nabla u|)}{\partial \vec{t}}.$$

We can further work out the expression to

$$L_\phi = \frac{-1}{|\nabla u|^2}\left(|\nabla u|\ \phi''(\kappa)\frac{\partial \kappa}{\partial \vec{t}} + \phi'(\kappa)\ [\nabla \otimes \nabla u](\vec{n}, \vec{t}\,)\right).$$

Here $[\nabla \otimes \nabla u](\bullet, \bullet)$ denotes the Hessian bilinear form. Thus in a simple case such as $\phi(s) = |s|$ and $\kappa \neq 0$, it simplifies to

$$L_\phi = \frac{\pm 1}{|\nabla u|^2}[\nabla \otimes \nabla u](\vec{n}, \vec{t}\,),$$

which more closely resembles the experimental choice in Bertalmio et al. [3] of the Laplacian

$$\Delta u = \text{trace}(\nabla \otimes \nabla u) = [\nabla \otimes \nabla u](\vec{n}, \vec{n}\,) + [\nabla \otimes \nabla u](\vec{t}, \vec{t}\,).$$

In summary, the elastica inpainting scheme combines both the transport mechanism of the Bertalmio group's model and the CDD mechanism of Chan and Shen's model. It thus provides a theoretical foundation for these two earlier empirical works. In return, the earlier works also shed light on the meaning of the flux field $\vec{V}$ and the PDE interpretation of elastica inpainting.

6. Computation and examples.

6.1. Numerical implementation. In this section, we explain the numerical scheme for the evolution equation (5.7):

$$\frac{\partial u}{\partial t} = |\nabla u|\ \nabla \cdot \vec{V} - |\nabla u|\ \lambda_E\ (u - u_0),$$

where the flux $\vec{V}$ and λ_E are as given in Corollary 5.3. We remind our readers that the factor $|\nabla u|$, as suggested by Marquina and Osher in [25], is for accelerating the time marching.

Let (i,j) denote a general pixel, and time be digitized to $n = 0, 1, \ldots$, with a small time step h. Thus $u^n_{(i,j)}$ denotes the value of u at pixel (i,j) at time nh. Then

$$u^{n+1}_{(i,j)} = u^n_{(i,j)} + h\left(|\nabla u^n_{(i,j)}|\, F(u^n_{(i,j)}) - |\nabla u^n_{(i,j)}|\, \lambda_{E,(i,j)}\, (u^n_{(i,j)} - u_{0,(i,j)})\right),$$

where $F(u^n_{(i,j)}) = \nabla \cdot \vec{V}^n_{(i,j)}$ and u_0 is the available image.

We now focus on the spatial digitization of the right-hand side at a fixed time nh. Thus we shall conveniently leave out the superscript n. Following Marquina and Osher [25], the accelerating factor $|\nabla u_{(i,j)}|$ in front of $F(u_{(i,j)})$ is approximated by the central differencing

$$|\nabla u_{(i,j)}| = \frac{1}{2}\sqrt{(u_{(i+1,j} - u_{(i-1,j)})^2 + (u_{(i,j+1)} - u_{(i,j-1)})^2}, \tag{6.1}$$

and the factor $|\nabla u_{(i,j)}|$ in front of $\lambda_{E,(i,j)}\, (u_{(i,j)} - u_{0,(i,j)})$ by the upwind scheme

$$|\nabla u_{(i,j)}| = \sqrt{(\text{upwind } D_x u_{(i,j)})^2 + (\text{upwind } D_y u_{(i,j)})^2},$$

$$\text{upwind } D_x u_{(i,j)} = \begin{cases} u_{(i,j)} - u_{(i-1,j)} & \text{if} \quad (u_{(i+1,j)} - u_{(i-1,j)})(u_{(i,j)} - u_{0,(i,j)}) > 0, \\ u_{(i+1,j)} - u_{(i,j)} & \text{if} \quad (u_{(i+1,j)} - u_{(i-1,j)})(u_{(i,j)} - u_{0,(i,j)}) < 0. \end{cases}$$

The upwinding on y is similar.

Now we focus on the discretization of $F(u_{(i,j)}) = \nabla \cdot \vec{V}_{(i,j)}$. Write $\vec{V} = (V^1, V^2)$ and

$$\vec{n} = (n^1, n^2) = \left(\frac{u_x}{|\nabla u|}, \frac{u_y}{|\nabla u|}\right), \qquad \vec{t} = (t^1, t^2) = \left(-\frac{u_y}{|\nabla u|}, \frac{u_x}{|\nabla u|}\right).$$

Then

$$\begin{aligned} V^1 &= (a + b\kappa^2)\, n^1 - \frac{2b}{|\nabla u|}\left(t^1 D_x(\kappa|\nabla u|) + t^2 D_y(\kappa|\nabla u|)\right) t^1 \\ &= (a + b\kappa^2)\, \frac{D_x u}{|\nabla u|} + \frac{2b}{|\nabla u|^3}\left(-D_y u\, D_x(\kappa|\nabla u|) + D_x u\, D_y(\kappa|\nabla u|)\right)\, D_y u, \end{aligned}$$

where the partial derivative symbols D_x and D_y are introduced to ease the placement of subscripts. The expression of V^2 can be worked out similarly. Based on the half-point central differencing, we have

$$\begin{aligned} F(u_{(i,j)}) = \nabla \cdot \vec{V}_{(i,j)} &= D_x V^1_{(i,j)} + D_y V^2_{(i,j)} \\ &= \left(V^1_{(i+\frac{1}{2},j)} - V^1_{(i-\frac{1}{2},j)}\right) + \left(V^2_{(i,j+\frac{1}{2})} - V^2_{(i,j-\frac{1}{2})}\right). \end{aligned}$$

Thus we need to specify the *half-point* values for all the quantities involved. Take the x-half-point $(i+1/2, j)$, for example. For the curvature, we take the min-mod [33] between the *whole* pixels:

$$\kappa_{(i+\frac{1}{2},j)} = \text{minmod}(\kappa_{(i+1,j)}, \kappa_{(i,j)}), \qquad \text{minmod}(\alpha, \beta) = \frac{\text{sign}(\alpha) + \text{sign}(\beta)}{2} \min(|\alpha|, |\beta|).$$

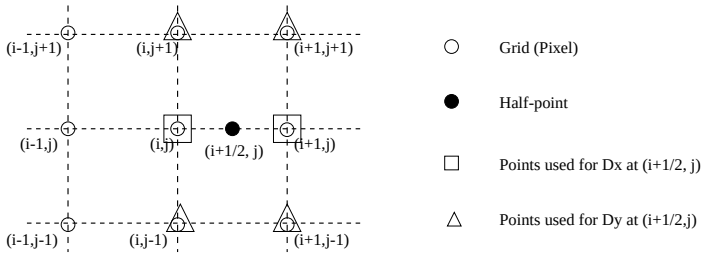


FIG. 6.1. *Grid description for the finite difference schemes.*

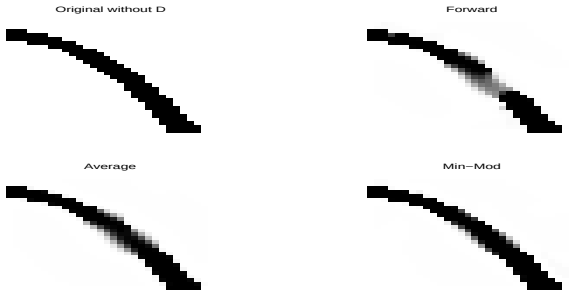


FIG. 6.2. *Experimental results show the advantage of the min-mod discretization in* (6.2). *The upper left panel is the original complete image of a ribbon. The inpainting domain is a square covering the middle part. The other three images are the outputs of the numerical inpainting schemes based on the* forward *substitution, the* average *discretization, and the* min-mod *discretization, separately. (See the text for more details.) The min-mod scheme seems to yield better edge sharpness, as expected from the shock wave computation in computational fluid dynamics.*

D_x's at an x-half-point $(i+1/2, j)$ are approximated by the central differencing of the two adjacent whole pixels $(i+1, j)$ and (i, j). For examples (see Figure 6.1),

$$D_x u_{(i+1/2,j)} = u_{(i+1,j)} - u_{(i,j)},$$
$$D_x(\kappa|\nabla u|)_{(i+1/2,j)} = \kappa_{(i+1,j)}\ |\nabla u|_{(i+1,j)} - \kappa_{(i,j)}\ |\nabla u|_{(i,j)}.$$

Here $|\nabla u|_{(i,j)}$ is as in (6.1). The D_y's at an x-half-point $(i+1/2, j)$ are approximated by the min-mod of the D_y's at the two adjacent whole pixels $(i+1, j)$ and (i, j) (see Figure 6.1). For instance, for $D_y u_{(i+1/2,j)}$,

$$D_y u_{(i+1/2,j)} = \text{minmod}\left(\frac{1}{2}(u_{(i+1,j+1)} - u_{(i+1,j-1)}),\ \ \frac{1}{2}(u_{i,j+1)} - u_{(i,j-1)})\right). \tag{6.2}$$

The same can be done for $D_y(\kappa|\nabla u|)$ at $(i+1/2, j)$. Then $|\nabla u|^2$ at $(i+1/2, j)$ is naturally defined as the sum of squares of $D_x u_{(i+1/2,j)}$ and $D_y u_{(i+1/2,j)}$. Therefore, eventually, all quantities involved are expressed by the gray levels $u_{(i,j)}$ at whole pixels.

Numerical experiments in Figure 6.2 have shown the advantage of the min-mod discretization for D_y at x-half-points, compared with two other competing methods ($w = u$ or $\kappa|\nabla u|$): the *forward* substitution of $D_y w_{(i+1/2,j)}$ by $D_y w_{(i,j)}$, and the *average* substitution by

$$\frac{1}{2}\left(\frac{u_{(i+1,j+1)} - u_{(i+1,j-1)}}{2} + \frac{u_{(i,j+1)} - u_{(i,j-1)}}{2}\right).$$

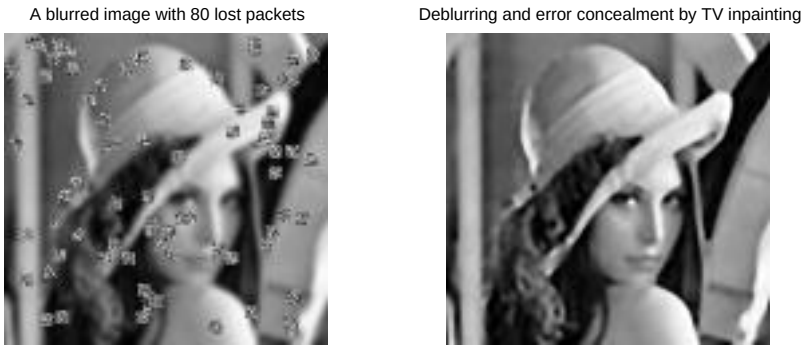


FIG. 6.3. *TV inpainting for the error concealment of a blurry image with packets randomly lost when transmitted through a wireless network* [10].

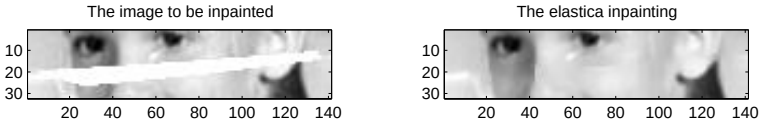


FIG. 6.4. *An example of elastica inpainting for scratch removal.*

As in computational fluid dynamics, the min-mod method seems to catch sharper edges (or *shocks*) more effectively.

6.2. Examples. We provide several numerical examples of elastica inpainting.

Figure 6.3 shows an application of the TV inpainting model (with $b = 0$ in the elastica model (4.16)) for the error concealment of wireless image transmissions where packets are randomly lost. (To be applicable to a blurry image with a linear blurring kernel H, model (4.16) should have the least square data model replaced by (choosing $E = \Omega \setminus K$)

$$\frac{\lambda}{2} \int_{\Omega \setminus K} (Hu - u_0)^2 dx.$$

See the recent paper by Chan and Shen [10] for more detail.) The TV inpainting model works well for most *local* inpainting problems but becomes problematic when applied to the inpainting of incomplete images with large-scale missing domains (see Chan and Shen [8]). For the latter, the curvature term in the elastica model becomes necessary.

Figure 6.4 shows the output of the elastica inpainting model when applied to the digital restoration of an old scratched photograph (image source: [3]). The example shows what an inpainting model must be able to accomplish: consistently reconnecting all the broken isophotes, including broken edges with low contrasts (like the shadow of the nose).

The next two figures demonstrate two universal effects of curvature-based inpainting. The example in Figure 6.5 shows that if more weights are put against the curvature term in the elastica model, the inpainted isophotes and edges become smoother and perceptually better. The second example in Figure 6.6 explains that as more weights are put against the curvature term, the model tends to favor the

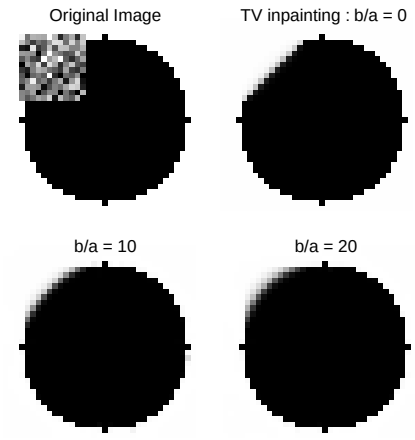


FIG. 6.5. *Effect* I *of elastica inpainting: a larger weight b against the curvature term produces smoother isophotes and edges and better visual results.*

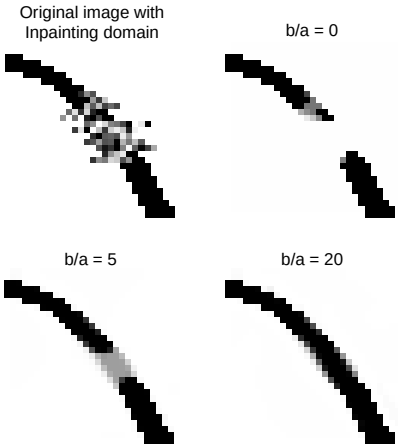


FIG. 6.6. *Effect* II *of elastica inpainting: a large weight b against the curvature term favors the* connectivity principle*: the model encourages the connection of separated parts.*

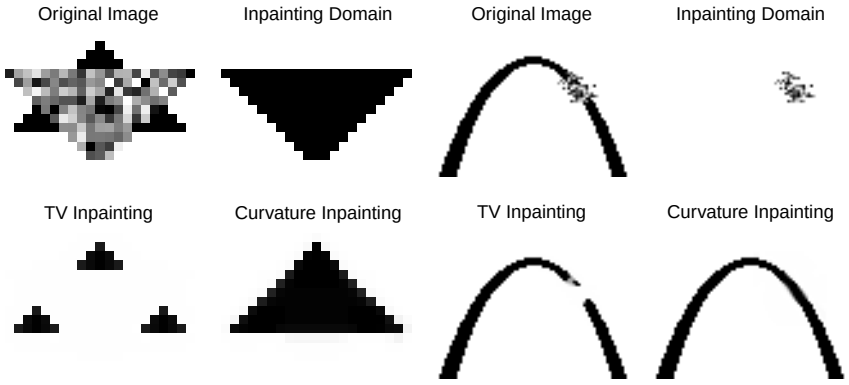


FIG. 6.7. *Two more examples of elastica inpainting.*

connectivity principle in perception [9, 20, 32]. That is, unlike the extreme case of TV inpainting, the model encourages connection.

The last figure (Figure 6.7) shows two more examples of elastica inpainting, where one can further appreciate the power of the elastica inpainting model and the numerical PDE approach. Large scale "communication" among the separated parts is made possible simply because of a good image or curve prior model—Euler's elastica.

7. A remark on the inpainting model of Ballester et al. [1]. Before concluding, we would like to direct our readers to a recent paper by Ballester, Bertalmio, Caselles, Sapiro, and Vergera [1] that is closely connected to our own. We shall conveniently call it the BBCSV model below.

The BBCSV variational inpainting model inpaints both the normal field $\vec{n} = \nabla u/|\nabla u|$ and the gray value image u simultaneously over the extended inpainting domain $K \cup E$, based on the boundary data u_0 and $\vec{n}_0$. (To be consistent and more readable, we have followed our own notations in the current paper.) The BBCSV

model tries to minimize the energy

$$J[u,\vec{n}] = \int_{K\cup E} |\nabla\cdot\vec{n}|^p\ (a+b|\nabla u|)dx + \alpha\int_{K\cup E}(|\nabla u| - \vec{n}\cdot\nabla u)dx \tag{7.1}$$

in a suitably defined admissible space (see [1]).

The second fitting term is used to enforce the ideal meaning of $\vec{n}$: $\vec{n} = \nabla u/|\nabla u|$. Therefore, ideally, the BBCSV energy $J[u,\vec{n}]$ is essentially reduced to

$$\int_{K\cup E}\left|\nabla\cdot\left[\frac{\nabla u}{|\nabla u|}\right]\right|^p\ (a+b|\nabla u|)dx, \tag{7.2}$$

which is different from the p-elastica energy in the current paper:

$$\int_{K\cup E}\left(a+b\left|\nabla\cdot\left[\frac{\nabla u}{|\nabla u|}\right]\right|^p\right)|\nabla u|dx.$$

The major difference between the two is that the elastica model leads to morphologically invariant flows, as shown in Theorem 5.1 and Proposition 5.2.

We refer to [1] for further interesting discussion.

Acknowledgments. We thank the reviewers for all their valuable and constructive comments. We are grateful to Professor Guillermo Sapiro's group for their help on the inpainting topic, and Professors Stanley Osher, Luminita Vese, Li-Tien Cheng, Simon Masnou, Jean-Michel Morel, Peter Olver, Robert Gulliver, Fadil Santosa, Selim Esedoglu, and Rachid Deriche for their helpful interactions and suggestions during this project. Jianhong (Jackie) Shen would also like to thank Professors Gil Strang, Stu Geman, David Mumford, Jayant Shah, and Dan Kersten for their teaching and inspirations on the subject.

REFERENCES

[1] C. Ballester, M. Bertalmio, V. Caselles, G. Sapiro, and J. Verdera, *Filling-in by joint interpolation of vector fields and grey levels,* IEEE Trans. Image Process., 10 (2001), pp. 1200–1211.

[2] G. Bellettini, G. Dal Maso, and M. Paolini, *Semicontinuity and relaxation properties of a curvature depending functional in* 2*D,* Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 20 (1993), pp. 247–297.

[3] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, *Image inpainting,* in Proceedings of the 27th International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH 2000), New Orleans, LA, ACM Press, New York, 2000.

[4] G. Birkhoff and C. R. De Boor, *Piecewise polynomial interpolation and approximation*, in Approximation of Functions, H. Garabedian, ed., Elsevier, New York, 1965, pp. 164–190.

[5] E. J. Candés and D. L. Donoho, *Curvelets and reconstruction of images from noisy radon data*, in Wavelet Applications in Signal and Image Processing VIII, A. Aldroubi, A. F. Laine, and M. A. Unser, eds., Proc. SPIE 4119, 2000.

[6] V. Caselles, J.-M. Morel, and C. Sbert, *An axiomatic approach to image interpolation,* IEEE Trans. Image Process., 7 (1998), pp. 376–386.

[7] A. Chambolle and P. L. Lions, *Image recovery via total variational minimization and related problems,* Numer. Math., 76 (1997), pp. 167–188.

[8] T. F. Chan and J. Shen, *Mathematical models for local nontexture inpaintings*, SIAM J. Appl. Math., 62 (2001), pp. 1019–1043.

[9] T. F. Chan and J. Shen, *Nontexture inpainting by curvature driven diffusions (CDD)*, J. Visual Comm. Image Rep., 12 (2001), pp. 436–449.

[10] T. F. Chan and J. Shen, *On the Role of the BV Image Model in Image Restoration*, CAM Report 02-14, Mathematics Department, UCLA, Los Angeles, CA, 2000; in Contemp. Math., AMS, Providence, RI, 2002, to appear.

[11] S. S. Chern, W. H. Chen, and K. S. Lam, *Lectures on Differential Geometry*, World Scientific, River Edge, NJ, 1998.
[12] A. Cohen, R. DeVore, P. Petrushev, and H. Xu, *Nonlinear approximation and the space $BV(R^2)$*, Amer. J. Math., 121 (1999), pp. 587–628.
[13] I. Daubechies, *Ten Lectures on Wavelets*, CBMS-NSF Reg. Conf. Ser. in Appl. Math. 61, SIAM, Philadelphia, 1992.
[14] G. Emile-Male, *The Restorer's Handbook of Easel Painting*, Van Nostrand Reinhold, New York, 1976.
[15] S. Esedoglu and J. Shen, *Digital inpainting based on the Mumford-Shah-Euler image model*, European J. Appl. Math., 2002, to appear.
[16] S. Geman and D. Geman, *Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images*, IEEE Trans. Pattern Anal. Machine Intell., 6 (1984), pp. 721–741.
[17] E. De Giorgi, *Frontiere orientate di misura minima*, lecture notes, Sem. Mat. Scuola Norm. Sup. Pisa, 1960–61.
[18] E. Giusti, *Minimal Surfaces and Functions of Bounded Variation*, Birkhäuser Boston, Cambridge, MA, 1984.
[19] G. H. Golub and J. M. Ortega, *Scientific Computing and Differential Equations*, Academic Press, New York, San Diego, 1992.
[20] G. Kanizsa, *Organization in Vision*, Praeger, New York, 1979.
[21] D. C. Knill and W. Richards, *Perception as Bayesian Inference*, Cambridge University Press, London, 1996.
[22] J. Langer and D. A. Singer, *The total squared curvature of closed curves*, J. Differential Geom., 20 (1984), pp. 1–22.
[23] A. E. H. Love, *A Treatise on the Mathematical Theory of Elasticity*, 4th ed., Dover, New York, 1927.
[24] R. March, *Visual reconstruction with discontinuities using variational methods*, Image Vision Comput., 10 (1992), pp. 30–38.
[25] A. Marquina and S. Osher, *Explicit algorithms for a new time dependent model based on level set motion for nonlinear deblurring and noise removal*, SIAM J. Sci. Comput., 22 (2000), pp. 387–405.
[26] D. Marr and E. Hildreth, *Theory of edge detection*, Proc. Roy. Soc. London Sect. B, 207 (1980), pp. 187–217.
[27] S. Masnou and J.-M. Morel, *Level-lines based disocclusion*, in Proceedings of 5th IEEE International Conference on Image Processing (ICIP), Chicago, 3 (1998), pp. 259–263.
[28] J.-M. Morel and S. Solimini, *Variational Methods in Image Segmentation*, Progr. Nonlinear Differential Equations Appl. 14, Birkhäuser Boston, Cambridge, MA, 1995.
[29] D. Mumford, *Elastica and computer vision*, in Algebraic Geometry and Its Applications, C. L. Bajaj, ed., Springer-Verlag, New York, 1994, pp. 491–506.
[30] D. Mumford, *The Bayesian rationale for energy functionals*, in Geometry Driven Diffusion in Computer Vision, B. M. ter Haar Romeny, ed., Kluwer Academic Publishers, Norwell, MA, 1994, pp. 141–153.
[31] D. Mumford and J. Shah, *Optimal approximations by piecewise smooth functions and associated variational problems*, Comm. Pure Appl. Math., 42 (1989), pp. 577–685.
[32] M. Nitzberg, D. Mumford, and T. Shiota, *Filtering, Segmentation, and Depth*, Lecture Notes in Comp. Sci. 662, Springer-Verlag, Berlin, 1993.
[33] S. Osher and L. I. Rudin, *Feature-oriented image enhancement using shock filters*, SIAM J. Numer. Anal., 27 (1990), pp. 919–940.
[34] S. Osher and J. A. Sethian, *Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton-Jacobi formulations*, J. Comput. Phys., 79 (1988), pp. 12–49.
[35] E. L. Pennec and S. Mallat, *Image compression with geometrical wavelets*, in Proceedings of the 7th IEEE International Conference on Image Processing (ICIP), 1 (2000), pp. 661–664.
[36] L. Rudin and S. Osher, *Total variation based image restoration with free local constraints*, Proc. 1st IEEE International Conference on Image Processing (ICIP), 1 (1994), pp. 31–35.
[37] L. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Phys. D, 60 (1992), pp. 259–268.
[38] J. Shah, *Elastica with hinges*, J. Visual Comm. Image Rep., 13 (2002), pp. 36–43.
[39] G. Strang, *Introduction to Applied Mathematics*, Wellesley–Cambridge Press, Wellesley, MA, 1993.
[40] S. Walden, *The Ravished Image*, St. Martin's Press, New York, 1985.

SIAM J. APPL. MATH.
Vol. 63, No. 2, pp. 593–618

THE APEX METHOD IN IMAGE SHARPENING AND THE USE OF LOW EXPONENT LÉVY STABLE LAWS*

ALFRED S. CARASSO†

Abstract. The APEX method is an FFT-based direct blind deconvolution technique that can process complex high resolution imagery in seconds or minutes on current desktop platforms. The method is predicated on a restricted class of shift-invariant blurs that can be expressed as finite convolution products of two-dimensional radially symmetric Lévy stable probability density functions. This class generalizes Gaussian and Lorentzian densities but excludes defocus and motion blurs. Not all images can be enhanced with the APEX method. However, it is shown that the method can be usefully applied to a wide variety of *real blurred images*, including astronomical, Landsat, and aerial images, MRI and PET brain scans, and scanning electron microscope images. APEX processing of these images enhances contrast and sharpens structural detail, leading to noticeable improvements in visual quality. The discussion includes a documented example of nonuniqueness, in which distinct point spread functions produce high-quality restorations of the same blurred image. Significantly, *low exponent* Lévy point spread functions were detected and used in all the above examples. Such low exponents are exceptional in physical applications where symmetric stable laws appear. In the present case, the physical meaning of these Lévy exponents is uncertain.

Key words. image deblurring; blind deconvolution; direct methods; electronic imaging systems; heavy-tailed distributions; low exponent stable laws; APEX method; SECB method; nonuniqueness; astronomical, Landsat, and SEM images; MRI and PET brain scans

AMS subject classifications. 35R25, 35B60, 60E07, 68U10

PII. S0036139901389318

1. Introduction.

The APEX method is an FFT-based direct blind deconvolution technique introduced by the author in [9]. The significance of the present paper lies in the successful use of that method in sharpening a wide variety of *real blurred images*, as opposed to the synthetically blurred images discussed in [9]. The reasons behind these successful applications are not fully understood. Not all images can be usefully enhanced with the APEX method. The present paper is essentially self-contained and may be read independently of [9].

Blind deconvolution seeks to deblur an image without knowing the point spread function (psf) describing the blur. Most approaches to that problem are iterative in nature. Because nonuniqueness is compounded with discontinuous dependence on data, such iterative procedures are not always well-behaved. When the iterative process is stable, several thousand iterations may be necessary to achieve useful reconstructions. However, as shown in [9], by limiting the class of blurs, noniterative direct procedures can be devised that accomplish blind deconvolution of 512×512 images in seconds on current desktop platforms.

The APEX method assumes the image $g(x, y)$ to have been blurred by a restricted type of shift-invariant psf $h(x, y)$, one that can be expressed as a finite convolution

*Received by the editors May 11, 2001; accepted for publication (in revised form) June 5, 2002; published electronically December 11, 2002.

http://www.siam.org/journals/siap/63-2/38931.html

†Mathematical and Computational Sciences Division, National Institute of Standards and Technology, Gaithersburg, MD 20899 (alfred.carasso@nist.gov).

product of two-dimensional (2-D) radially symmetric Lévy stable probability density functions. Such so-called class **G** psfs include Gaussians, Lorentzians, and their convolutions. However, the class **G** also excludes defocus and motion blurs and convolutions of such blurs with Gaussians and Lorentzians.

The synthetically blurred images $g(x, y)$ used in [9] were created by numerical convolution of sharp images $f(x, y)$ with class **G** psfs $h(x, y)$. Such blurred images necessarily obey the convolutional model $g(x, y) = h(x, y) \otimes f(x, y) + noise$, on which the APEX method is predicated. In a real image, the blur need not be radially symmetric nor shift-invariant and may otherwise be poorly approximated by an element of **G**. More fundamentally, the blurring operator may not even be linear. Applicability of the APEX method to a given real image is far from obvious. Therefore, useful sharpening of any such image with an APEX-detected psf is always instructive.

Stable distributions are the natural generalization of the Gaussian distribution. Their theory was developed by Paul Lévy in the 1930s in connection with his work on the central limit theorem (see [17]). In the simplest radially symmetric case, these distributions are characterized by an exponent β, $0 < \beta \leq 1$, with $\beta = 1$ corresponding to the Gaussian distribution, and $\beta = 1/2$ corresponding to the Cauchy or Lorentzian distribution. Because stable distributions have infinite variance when $\beta < 1$, their appearance in physical contexts sometimes poses philosophical difficulties. In the present case, use of such heavy-tailed psfs as the framework for the APEX method is motivated by the important role Lévy densities appear to play in numerous imaging systems. This is documented in section 2. When the APEX method is applied to a given image in the manner described below, a Lévy psf with a specific value of β is necessarily detected. That value of β may not be indicative of the actual physical process that created the image. This is true even if deblurring with the detected psf significantly improves the image. As shown in section 4, there are in general infinitely many distinct values of β that can produce useful reconstructions from the same blurred image. In some cases, the usefully enhanced image may not have been blurred by a class **G** psf to begin with. In other cases, APEX processing does not significantly improve the image.

Below, we exhibit ten images where APEX processing provided noticeable improvement. These examples encompass such diverse imaging applications as astronomical, Landsat, and aerial images, MRI and PET brain scans, a scanning electron microscope image, a face image, and other types of interesting images. In some cases, the improvement is due primarily to an increase in contrast. In other cases, there is demonstrable sharpening of structural detail in addition to increased contrast. In all cases, the change in image quality is more than cosmetic, as APEX processing significantly alters the image Fourier transform. It is noteworthy that *low exponent* stable laws, with $\beta \ll 1/2$, were detected and used to deblur all of the images shown below. Such β-values are exceptional in physical contexts where radially symmetric Lévy densities appear. Whether or not these values have a physical origin cannot be ascertained in the present case. Moreover, the APEX detection procedure may not be well founded. Nevertheless, the fact remains that the use of such psfs produced valuable restoration of real imagery from important fields of science and technology. To the author's knowledge, this application of sub-Cauchy stable laws in image processing is new and unanticipated.

In recent years, there has been considerable interest in image processing techniques that can be formulated as initial value problems in nonlinear PDEs. An instructive survey of these developments may be found in [11]. In particular, novel approaches to image deblurring have been devised, based on integrating well-posed

nonlinear anisotropic diffusion equations [33], [38]. In contrast, the APEX method centers around ill-posed continuation in linear fractional diffusion equations. As noted in section 7, for the type of finely textured imagery considered in the present paper, APEX processing compares favorably with what is feasible with nonlinear methods. This indicates that the APEX method can be a useful addition to PDE-based image analysis.

2. Imaging systems, Lévy processes, and the class G. The occurrence and analysis of Lévy processes in the physical sciences are subjects of significant current interest; see [1], [2], [4], [32], [39], [40], [41], [44], and the references therein. An important special case involves 2-D radially symmetric Lévy stable densities $h(x, y)$, implicitly defined in terms of their Fourier transforms by

$$(1) \qquad \hat{h}(\xi,\eta) \equiv \int_{R^2} h(x,y) e^{-2\pi i(\xi x+\eta y)} dxdy = e^{-\alpha(\xi^2+\eta^2)^\beta}, \qquad \alpha > 0, \quad 0 < \beta \leq 1.$$

The cases $\beta = 1$ and $\beta = 1/2$ correspond to Gaussian and Lorentzian (or Cauchy) densities, respectively. For other values of β, $h(x, y)$ in (1) is not known in closed form. When $\beta = 1$, $h(x, y)$ has slim tails and finite variance. For $0 < \beta < 1$, $h(x, y)$ has fat tails and infinite variance. As noted in [44], there are examples in science where the occurrence of a stable law can be deduced from "first principles" in terms of physical mechanisms that do not explicitly involve the parameter β. One such instance is the Holtsmark distribution describing the gravitational field of stars (see [17]). There, mathematical analysis reveals the value $\beta = 3/4$. Such cases must be distinguished from the many other cases in which empirically obtained data with fat tails are *fitted* to a Lévy law, and the exponent β is *inferred* from these data. Given the limitations of physical measurements, such empirically established Lévy processes do not have the degree of scientific legitimacy that attends the Holtsmark distribution. The considerations of the present paper generally lie in this weaker scientific realm. Nevertheless, as will be seen below, techniques derived from such considerations turn out to be effective.

Image intensifiers, charge-coupled devices, and numerous other electronic devices are used in a wide variety of astronomical, industrial, biomedical, military, and surveillance imaging systems; see [3], [14], [15], [18], [31]. Each such device has a psf $h(x, y)$ characterizing that device's imaging properties. The psf is a probability density function since it is nonnegative and integrates to unity. Use of such a device to image an object $f(x, y)$ produces a blurred image $g(x, y) = h(x, y) \otimes f(x, y)$, where $\otimes$ denotes convolution. An ideal device would have $h(x, y) = \delta(x, y)$. The Fourier transform $\hat{h}(\xi, \eta)$ of the psf is generally complex-valued and is called the *optical transfer function* (otf). The absolute value of the otf is the *modulation transfer function* (mtf).

In [42], it is noted that electron optical mtfs are often nearly Gaussian in shape, and that this should be expected from the central limit theorem, since the process of converting incoming signal photons into the final image that is observed on a screen involves many intermediate stages. However, it is also observed in [42] that when such mtfs are fitted with Gaussians, the fitted curves often have *slimmer* tails than is the case for the true mtfs.

A systematic study of electron optical mtf *measurements* is the subject of [22], [24], and [27]. There, the author claims the empirical discovery that a wide variety of electronic imaging devices, including phosphor screens and some types of photographic film, have otfs $\hat{h}(\xi, \eta)$ that are well described by (1) with $1/2 \leq \beta \leq 1$. For any given device, the values of α and β can be determined using specialized graph paper [28].

Other instances of electron optical stable laws are mentioned in [23], [26], and [30]. Analysis of the physical mechanisms responsible for such non-Gaussian behavior is not included in these works. An understanding of such mechanisms may lead to the design of imaging devices with *low* values of β. The latter parameter affects the attenuation of high frequency information in the recorded image. Deconvolution of that image in the presence of noise is generally better behaved at low values of β than it is at high values of β.

The characterization (1) is useful in other areas of optics. The otf for long-exposure imaging through atmospheric turbulence [21] is known to be given by (1), with $\beta = 5/6$ and α determined by atmospheric conditions. Also, as shown in [25], the analytically known diffraction-limited otf for a perfect lens [43, p. 154] can be approximated over a wide frequency range by (1), with $\beta = 3/4$ and α a properly chosen function of the cutoff frequency.

The range of β values discussed above, namely, $1/2 \leq \beta \leq 1$, mirrors that found in most other physical contexts where *symmetric* stable laws appear or are surmised. Values of $\beta \ll 1/2$ seem to be relatively rare in applications. Examples of such β values occur in [29], where mtf data for 56 different kinds of *photographic film* are analyzed. Good agreement is found when these data are fitted with (1) and the pairs (α, β) characterizing each of these 56 mtfs are identified. It is found that 36 types of film have mtfs where $1/2 \leq \beta \leq 1$. The remaining 20 types have mtfs with values of β in the range $0.265 \leq \beta \leq 0.475$.

We now consider imaging systems composed of various elements satisfying (1). Such systems might be used to image objects through a turbulent atmosphere or through other distorting media whose otfs obey (1). The resulting composite otf has the form

$$\hat{h}(\xi, \eta) = e^{-\sum_{i=1}^{J} \alpha_i (\xi^2+\eta^2)^{\beta_i}}, \qquad \alpha_i \geq 0, \quad 0 < \beta_i \leq 1. \tag{2}$$

Such an object corresponds to a *multifractal law* in [4]. We define the class **G** to be the class of all psfs $h(x, y)$ with Fourier transforms satisfying (2). We shall be interested in image deblurring problems

$$Hf \equiv \int_{R^2} h(x-u, y-v) f(u, v) du dv \equiv h(x, y) \otimes f(x, y) = g(x, y), \tag{3}$$

where $g(x, y)$ is the recorded blurred image, $f(x, y)$ is the desired unblurred image, and $h(x, y)$ is a known psf in class **G**. The blurred image $g(x, y)$ includes (possibly multiplicative) noise, which is viewed as a separate additional degradation,

$$g(x, y) = g_e(x, y) + n(x, y). \tag{4}$$

Here, $g_e(x, y)$ is the blurred image that would have been recorded in the absence of any noise, and $n(x, y)$ represents the cumulative effects of all errors affecting final acquisition of the digitized array $g(x, y)$. Neither $g_e(x, y)$ nor $n(x, y)$ are known, only their sum $g(x, y)$. The unique solution of (3) when the right-hand side is $g_e(x, y)$ is the exact sharp image denoted by $f_e(x, y)$. Thus

$$h(x, y) \otimes f_e(x, y) = g_e(x, y). \tag{5}$$

3. Deblurring with the SECB method. The SECB method is a direct FFT-based image deblurring technique designed for equations of the form (3), when $h(x, y)$ is known and belongs to **G**. The method is based on inverse diffusion equations, and

features an important new *slow evolution* regularizing constraint. Such regularization leads to smaller error bounds for the reconstructed image $f(x,y)$, as a function of the noise level ϵ in the blurred image $g(x,y)$, than is mathematically possible with the basic Tikhonov–Miller method. Significantly, the method does not impose smoothness constraints on the unknown image $f(x,y)$, nor does it require knowledge of the noise statistics other than an L^2 upper bound ϵ. Naturally, the method works best when ϵ is small. The above important theoretical advantages, coupled with the practical advantages of fast computation through FFT algorithms, render the SECB method a valuable tool in blind deconvolution. Theoretical analysis of the SECB method, along with error bounds and documented comparisons with the Tikhonov–Miller method, may be found in [5] and [6]. Comparisons with other widely used nonlinear probabilistic algorithms, including the Lucy–Richardson and maximum entropy methods, may be found in [7]. Image deblurring with class **G** psfs is just one example of an extensive class of ill-posed PDE problems [8]. That class includes problems ranging from analytic continuation in the unit disc to the time-reversed Navier–Stokes equations. As shown explicitly in [8], use of the "slow evolution" constraint in that class of problems leads to stronger stability estimates in terms of ϵ than previously known "Hölder-continuity" estimates.

For class **G** psfs, we may define fractional powers H^t, $0 \le t \le 1$, of the convolution integral operator H in (3) as follows:

$$H^t f \equiv \mathcal{F}^{-1}\left\{\hat{h}^t(\xi,\eta)\hat{f}(\xi,\eta)\right\}, \qquad 0 \le t \le 1. \tag{6}$$

Class **G** psfs are intimately related to diffusion processes, and solving (3) is mathematically equivalent to finding the initial value $u(x,y,0) = f(x,y)$ in the *backwards-in-time* problem for the generalized diffusion equation

$$\begin{aligned} &u_t = -\sum_{i=1}^{J} \lambda_i (-\Delta)^{\beta_i} u, \quad \lambda_i = \alpha_i (4\pi^2)^{-\beta_i}, \qquad 0 < t \le 1, \\ &u(x,y,1) \approx g(x,y). \end{aligned} \tag{7}$$

When $f(x,y)$ is known, $u(x,y,t) = H^t f$ is the solution of (7) at time t. The SECB method is a regularization method for solving the ill-posed problem (7) that takes into account the presence of noise in the blurred image data $g(x,y)$ at $t = 1$. With f, g, and n as in (3) and (4), and $u(t)$ the solution of (7), let ϵ, M be known positive constants such that

$$\|u(0)\|_2 = \|f\|_2 \le M, \quad \|u(1) - g\|_2 = \|n\|_2 \le \epsilon, \qquad \epsilon \ll M, \tag{8}$$

where $\|\ \|_2$ denotes the L^2 norm. For any constant $K > 0$ such that $K \ll M/\epsilon$, define $s^*\ (\epsilon, M, K)$ by

$$s^* = \frac{\log\{M/(M - K\epsilon)\}}{\log(M/\epsilon)}. \tag{9}$$

The "slow evolution" constraint applied to the backwards-in-time solution of (7) requires that there exist a known small constant $K > 0$ and a known fixed small $s > 0$, with $s/s^* \gg 1$, such that

$$\|u(s) - u(0)\|_2 \le K\epsilon. \tag{10}$$

Knowledge of the regularization parameters K and s represents a priori information about the solution of (3). As is well known, some form of a priori information is *always* necessary in the solution of ill-posed problems. Given K and s, the SECB solution of the backwards problem for (7) is defined to be that initial value $u^\dagger(0)$ which *minimizes*

$$\|u(1) - g\|_2^2 + K^{-2} \|u(s) - u(0)\|_2^2 \tag{11}$$

over all choices of initial values $u(0)$ in L^2. The SECB deblurred image $f^\dagger(x,y) \equiv u^\dagger(0)$ can be obtained in closed form in Fourier space. With $\overline{z}$ denoting the complex conjugate of z,

$$\hat{f}^\dagger(\xi,\eta) = \frac{\overline{\hat{h}}(\xi,\eta)\hat{g}(\xi,\eta)}{|\hat{h}(\xi,\eta)|^2 + K^{-2}|1 - \hat{h}^s(\xi,\eta)|^2}, \tag{12}$$

leading to $f^\dagger(x,y)$ upon inverse transformation. In practice, FFT algorithms are used to obtain $f^\dagger(x,y)$. This may result in individual pixel values that are negative or that exceed 255, the maximum value in an 8-bit image. Accordingly, all negative values are reset to the value zero, and all values exceeding 255 are reset to the value 255. One way of obtaining initial estimates for K and s in (12) is as follows. With ϵ, M, and the psf $h(x,y)$ known, fix $s > 0$ in the range $0.001 \le s \le 0.01$ and construct the operator H^s as in (6). If $f^\pi(x,y)$ is a prototype image for the class of images under consideration, we can estimate K in (10) by evaluating $\|H^s f^\pi - f^\pi\| / \epsilon$. We may then compute s^* in (9) and verify that $s/s^* \gg 1$. This is usually the case, as s^* is infinitesimally small, provided that $K\epsilon \ll M$. This initial choice of K can be refined interactively when the reconstructed image is a recognizable object. With s fixed as above, increasing K increases resolution until a threshold value is reached. Further increases in K bring out noise. Conversely, if the initial choice of K brings out noise, K must be decreased. Note that for 512×512 images, 20 trial SECB restorations, each with a different value of K, can be obtained simultaneously in about 10 seconds of cpu time on an MIPS R12000 (400MHz) workstation. A visually optimal value of K for fixed small s is usually easily found. We may also form and display

$$u^\dagger(x,y,t) = H^t f^\dagger(x,y) \tag{13}$$

for selected *decreasing* values of t lying between 1 and 0. This simulates *marching backwards in time* in (7) and allows *monitoring* the gradual deblurring of the image. As $t \downarrow 0$, the partial restorations $u^\dagger(x,y,t)$ become sharper. However, noise and other artifacts typically become more noticeable as $t \downarrow 0$. Marching backwards from $t = 0.2$ to $t = 0$, say, may allow detection of features in the image before they become obscured by noise or ringing artifacts.

The above discussion assumed that the psf $h(x,y)$ was known. As shown in sections 5 and 6, such marching backwards in time becomes much more vital in the blind deconvolution problem, where the initial APEX-detected psf may erroneously be *too wide*. Theoretically, use of too wide a psf all the way to $t = 0$ implies sharpening features that may have already become infinitely sharp at some $t_0 > 0$. In practice, this leads to severe ringing and other undesirable artifacts at $t = 0$. Here, it is often advisable to start marching backwards from $t = 1$.

It should be noted that the class **G** is only a small subclass of the class of *infinitely divisible* densities [17]. The latter class includes multimodal nonsymmetric psfs associated with linear diffusion equations more complex than (7). Detection of

such psfs from blurred image data would require considerable extension of the APEX method discussed below.

4. Nonuniqueness in blind deconvolution. Blind deconvolution of images is a mathematical problem that is not fully understood. Well-documented examples of the kinds of behavior that may occur are of particular interest. In this section, we highlight important nonuniqueness aspects of that problem that are helpful in understanding the results of the APEX method. Let $f_e(x,y)$ be a given exact sharp image, let $h(x,y)$ be a Lévy point spread function, and let $g_e(x,y) = h(x,y) \otimes f_e(x,y)$. We shall show that, given the blurred image $g_e(x,y)$, there are in general *many* point spread functions $h_i(x,y) \neq h(x,y)$ that deblur $g_e(x,y)$, producing *high quality* reconstructions $f_i(x,y) \neq f_e(x,y)$, with $h_i(x,y) \otimes f_i(x,y) \approx g_e(x,y)$.

The sharp 512×512 Sydney image $f_e(x,y)$ in Figure 1(a) was synthetically blurred by convolution with a Cauchy density $h(x,y)$ with $\alpha_0 = 0.075$, $\beta_0 = 0.5$. This produced the blurred image $g_e(x,y)$ in Figure 1(b). To avoid distractions caused by noise, the blurred image $g_e(x,y)$ in this experiment was computed and stored in 64-bit precision. Deblurring this noiseless image with the correct psf values $\alpha = 0.075$, $\beta = 0.5$, produces Figure 1(c). This is in excellent visual agreement with $f_e(x,y)$ in Figure 1(a), as expected. However, the visual quality in Figures 1(d)–(f) is generally as good as that in Figure 1(c); the latter three images were deblurred with Lévy densities with values (α,β), where $\alpha > \alpha_0$, $\beta < \beta_0$, and they differ from Figure 1(a) in contrast and brightness. All deblurred images were obtained using the SECB method with $s = 0.001$ and $K = 10000$. One-dimensional (1-D) cross sections of the four distinct psfs used in Figure 1 are displayed in Figure 2. These psfs also exhibit distinct heavy tail behavior not shown in Figure 2.

One can imagine four photographers, simultaneously photographing the identical scene depicted in Figure 1(a), yet producing the four distinct images shown in Figures 1(c)–(f) through use of different lenses, film, filters, exposures, printing, and the like. In practice, given only the blurred image in Figure 1(b), any one of these four restorations would be considered highly successful. Convolution of each reconstruction with its corresponding psf in Figure 2 reproduces the blurred image in Figure 1(b).

For any restoration $f(x,y)$ of the exact image $f_e(x,y)$ in Figure 1(a) and any norm $\|\ \|$, we can evaluate the relative error $\|f - f_e\| / \|f_e\|$. Define the discrete L^1, L^2, and H^m norms as follows:

$$\|f\|_1 = N^{-2} \sum_{x,y=1}^{N} |f(x,y)|,$$

$$\|f\|_2 = \left\{ N^{-2} \sum_{x,y=1}^{N} |f(x,y)|^2 \right\}^{1/2}, \tag{14}$$

$$\|f\|_{H^m} = \left\{ N^{-2} \sum_{\xi,\eta=1}^{N} \left(1 + \xi^2 + \eta^2\right)^m |\hat{f}(\xi,\eta)|^2 \right\}^{1/2}.$$

The relative errors in the L^1, L^2, H^1, and H^5 norms for each of the four restorations in Figure 1 are shown in Table 1. As might be expected, image (c) is the closest to image (a) in each of these norms, since the correct psf values were used to obtain image (c) from image (b). It is also evident from Table 1 that the four restorations are distinct from one another, since they differ from image (a) by different amounts.

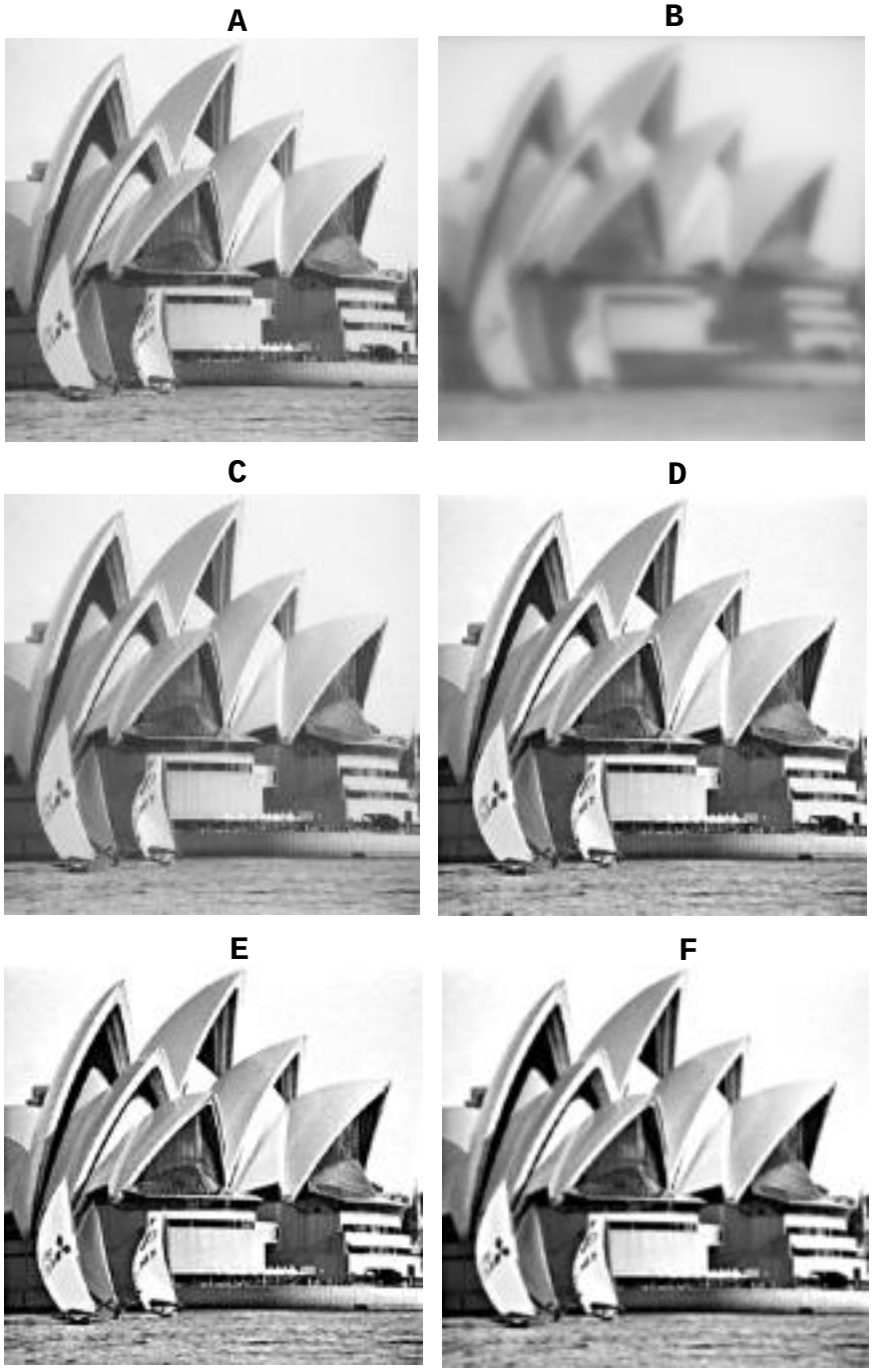


FIG. 1. *Nonuniqueness in blind deconvolution. Distinct psfs exist that produce high quality reconstructions from the same blurred image.* (a) *Original sharp* 512×512 *Sydney image.* (b) *Synthetically blurred Sydney image created by convolution with Lorentzian density with* $\alpha_0 = 0.075$, $\beta_0 = 0.5$. *Blurred image computed and stored in* 64*-bit precision.* (c) *Deblurring of image* (b) *using correct parameters* $\alpha = 0.075$, $\beta = 0.5$. (d) *Deblurring of image* (b) *using* $\alpha = 0.1301264$, $\beta = 0.44298$. (e) *Deblurring of image* (b) *using* $\alpha = 0.1950345$, $\beta = 0.403889$. (f) *Deblurring of image* (b) *using* $\alpha = 0.2360994$, $\beta = 0.369666$. *Notice that images* (d), (e), *and* (f) *were found using specific pairs* (α, β), *where* $\alpha > \alpha_0$ *and* $\beta < \beta_0$. *All deblurred images were obtained using the SECB procedure with* $s = 0.001$ *and* $K = 10000$.

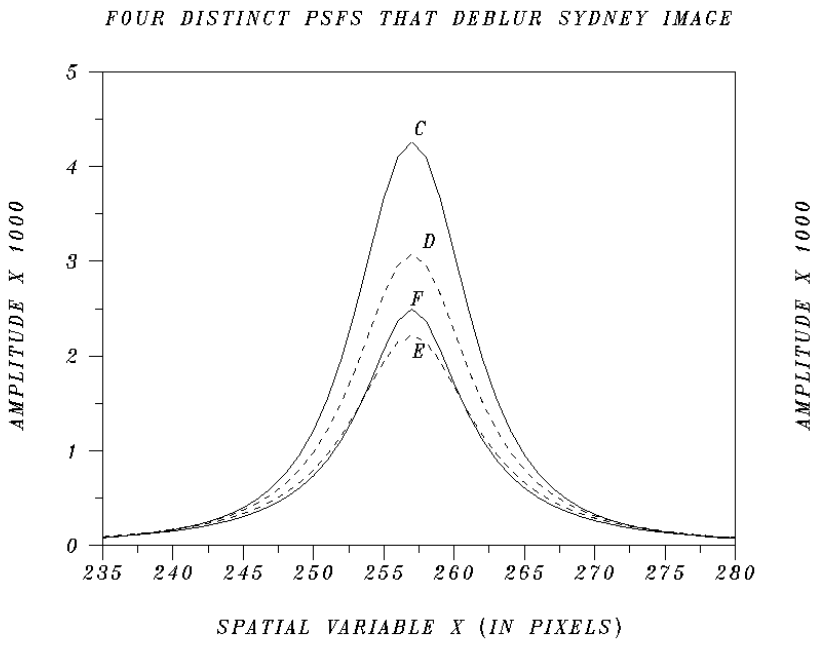


FIG. 2. *Four distinct psfs that deblur image* (b) *in Figure* 1. *Curves C, D, E, and F are* 1-*D cross sections of the* 512×512 *psfs that respectively produced images* (c), (d), (e), *and* (f) *in Figure* 1. *These psfs also exhibit distinct heavy tail behavior.*

TABLE 1
Relative errors in various norms for the four deblurred images in Figure 1.

Restoration	*Parameters* $\alpha,\ \beta$	L^1	L^2	H^1	H^5
Image (c)	$\alpha = 0.075,\ \beta = 0.500$	2.13 %	3.52 %	4.13 %	19.66 %
Image (d)	$\alpha = 0.130,\ \beta = 0.443$	6.63 %	8.37 %	8.67 %	21.11 %
Image (e)	$\alpha = 0.195,\ \beta = 0.404$	12.64 %	15.53 %	15.75 %	25.52 %
Image (f)	$\alpha = 0.236,\ \beta = 0.370$	12.54 %	15.08 %	15.31 %	26.17 %

Most important, the fact that image (e) is a significantly poorer approximation to image (a) in these norms than is image (c) *does not imply* that image (e) is an inaccurate representation of the visual scene depicted in image (a). Notice also that image (f) is not as sharp as image (e), although it is closer to image (a) in three of the four norms.

Iterative algorithms are the most common approach to blind deconvolution. Convergence proofs for such iterative procedures are seldom available. The above example illustrates some of the difficulties underlying any analysis of convergence. Such analysis should allow for the possibility of *infinitely many* useful limit points, while the mathematical characterization of such limit points is not obvious. Moreover, as is evident from Table 1 and has been known for some time, the use of L^p or H^m norms in assessing the visual quality of a reconstruction can be misleading.

5. Marching backwards in time and the APEX method. The APEX method is a blind deconvolution technique based on detecting class **G** psf signatures by appropriate 1-D Fourier analysis of the blurred image $g(x, y)$. The detected

psf parameters are then input into the SECB algorithm to deblur the image. Let $f_e(x,y)$ be an exact sharp image as in (5). Since $f_e(x,y) \geq 0$,

$$|\hat{f}_e(\xi,\eta)| \leq \int_{R^2} f_e(x,y)dxdy = \hat{f}_e(0,0) = \sigma > 0. \tag{15}$$

Also, since $g_e(x,y) = h(x,y) \otimes f_e(x,y)$ and $h(x,y)$ is a probability density,

$$\hat{g_e}(0,0) = \int_{R^2} g_e(x,y)dxdy = \int_{R^2} f_e(x,y)dxdy = \hat{f}_e(0,0) = \sigma > 0. \tag{16}$$

Using σ as a normalizing constant, we may normalize Fourier transform quantities $\hat{q}(\xi,\eta)$ by dividing by σ. Let

$$\hat{q}^*(\xi,\eta) = \frac{\hat{q}(\xi,\eta)}{\sigma} \tag{17}$$

denote the normalized quantity. The function $|\hat{f_e}^*(\xi,\eta)|$ is highly oscillatory, and $0 \leq |\hat{f_e}^*| \leq 1$. Since $f_e(x,y)$ is real, its Fourier transform is conjugate symmetric. Therefore, the function $|\hat{f_e}^*(\xi,\eta)|$ is symmetric about the origin on any line through the origin in the (ξ,η) plane. The same is true for the blurred image data $|\hat{g}^*(\xi,\eta)|$.

All blurred images in this and the next section are of size 512×512 and quantized at 8 bits per pixel. For any blurred image $g(x,y)$, the discrete Fourier transform is a 512×512 array of complex numbers, which we again denote by $\hat{g}(\xi,\eta)$ for simplicity. The "frequencies" ξ,η are now integers lying between -256 and 256, and the zero frequency is at the center of the transform array. This ordering is achieved by premultiplying $g(x,y)$ by $(-1)^{x+y}$. We shall be interested in the values of such transforms along single lines through the origin. The discrete transforms $|\hat{g}^*(\xi,0)|$ and $|\hat{g}^*(0,\eta)|$ are immediately available. Image rotation may be used to obtain discrete transforms along other directions. All 1-D Fourier domain plots shown in this paper are taken along the axis $\eta = 0$ in the (ξ,η) plane. In these plots, the zero frequency is at the center of the horizontal axis, and the graphs are necessarily symmetric about the vertical line $\xi = 0$. Examples of such plots are shown in Figures 3, 5, and 10.

The class of blurred images $g(x,y)$ considered in the present paper can be described in terms of the behavior of $\log|\hat{g}^*(\xi,\eta)|$ along lines through the origin in the (ξ,η) plane. While local behavior is highly oscillatory, global behavior is generally monotone decreasing and *convex*. This is shown in Figure 3 for two typical images along the line $\eta = 0$. In [9], a large class of images with that property was exhibited, the class **W**. The blurred images considered here may be loosely characterized as being in class **W**. Not all blurred images may be so characterized. For example, if the Cindy Crawford image $g(x,y)$ in Figure 3(a) were convolved with a wide Gaussian psf to form a new blurred image $g_1(x,y)$, global behavior in $\log|\hat{g_1}^*(\xi,0)|$, away from the origin, would be monotone decreasing and *concave*. Application of the APEX method to several concave examples is discussed in [9]. Convolution of Figure 3(a) with a defocus psf produces a different kind of blurred image $g_2(x,y)$, and global behavior in $\log|\hat{g_2}^*(\xi,0)|$ is neither concave nor convex. Instead, there is a regular pattern of sharp singularities corresponding to successive zeroes of the defocus otf. Use of the APEX method in the manner to be described below is intended only for blurred images with Fourier behavior analogous to that shown in Figure 3.

The APEX method is based on the following observations. In the basic relation

$$g(x,y) = h(x,y) \otimes f_e(x,y) + n(x,y), \tag{18}$$

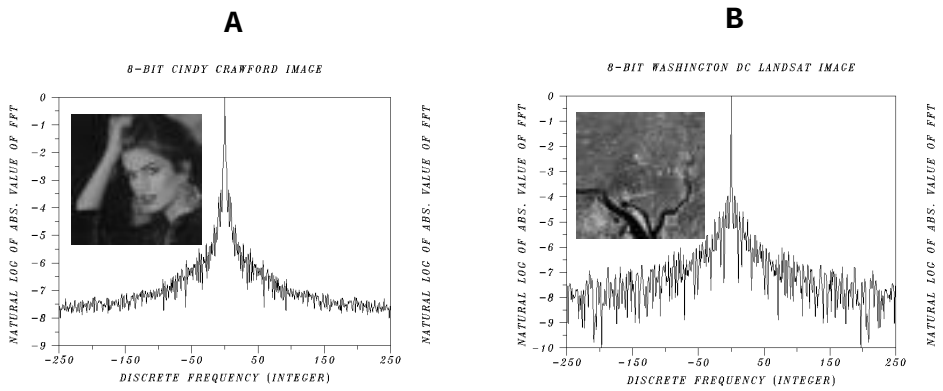


FIG. 3. *Behavior of a normalized Fourier transform in types of blurred images $g(x,y)$ considered in the present paper.* (a) $\log|\hat{g}^*(\xi,0)|$ *in an image of Cindy Crawford.* (b) $\log|\hat{g}^*(\xi,0)|$ *in a Washington, DC Landsat image. While local behavior is highly oscillatory, global behavior is generally monotone decreasing and convex.*

we may safely assume that the noise $n(x,y)$ satisfies

$$\int_{R^2} |n(x,y)|dxdy \ll \int_{R^2} f_e(x,y)dxdy = \sigma > 0, \tag{19}$$

so that

$$|\hat{n}^*(\xi,\eta)| \ll 1. \tag{20}$$

Consider the case in which the otf is a pure Lévy density $\hat{h}(\xi,\eta) = e^{-\alpha(\xi^2+\eta^2)^\beta}$. Since $g = g_e + n$,

$$\log|\hat{g}^*(\xi,\eta)| = \log|e^{-\alpha(\xi^2+\eta^2)^\beta}\hat{f_e}^*(\xi,\eta) + \hat{n}^*(\xi,\eta)|. \tag{21}$$

Let $\Omega = \{(\xi,\eta) \mid \xi^2+\eta^2 \le \omega^2\}$ be a neighborhood of the origin, where

$$e^{-\alpha(\xi^2+\eta^2)^\beta}|\hat{f_e}^*(\xi,\eta)| \gg |\hat{n}^*(\xi,\eta)|. \tag{22}$$

Such an Ω exists since (22) is true for $\xi = \eta = 0$, in view of (20). The radius $\omega > 0$ of Ω decreases as α and n increase. For $(\xi,\eta) \in \Omega$ we have

$$\log|\hat{g}^*(\xi,\eta)| \approx -\alpha(\xi^2+\eta^2)^\beta + \log|\hat{f_e}^*(\xi,\eta)|. \tag{23}$$

Because of the radial symmetry in the psf, it is sufficient to consider (23) along a single line through the origin in the (ξ,η) plane. Choosing the line $\eta = 0$, we have

$$\log|\hat{g}^*(\xi,0)| \approx -\alpha|\xi|^{2\beta} + \log|\hat{f_e}^*(\xi,0)|, \qquad |\xi| \le \omega. \tag{24}$$

Some type of a priori information about $f_e(x,y)$ is necessary for blind deconvolution. In (24), knowledge of $\log|\hat{f_e}^*(\xi,0)|$ on $|\xi| \le \omega$ would immediately yield $\alpha|\xi|^{2\beta}$ on that interval. Moreover, any other line through the origin could have been used in (23). However, such detailed knowledge is unlikely in practice. The APEX method seeks to identify a useful psf from (24) without prior knowledge of $\log|\hat{f_e}^*(\xi,0)|$. The

method assumes instead that $f_e(x,y)$ is a recognizable object, and typically requires several interactive trials before locating a suitable psf. As previously noted, such trial SECB restorations are easily obtained. Here, prior information about $f_e(x,y)$ is disguised in the form of user recognition or rejection of the restored image, and that *constraint* is applied at the end of the reconstruction phase, rather than at the beginning of the detection phase.

In the absence of information about $\log|\hat{f}_e^{\,*}(\xi,0)|$, we replace it by a negative constant $-A$ in (24). For any $A>0$, the approximation

$$\log|\hat{g}^*(\xi,0)| \approx -\alpha|\xi|^{2\beta} - A \tag{25}$$

is not valid near $\xi = 0$, since the curve $u(\xi) = -\alpha|\xi|^{2\beta} - A$ has $-A$ as its apex. Choosing a value of $A>0$, we best fit $\log|\hat{g}^*(\xi,0)|$ with $u(\xi) = -\alpha|\xi|^{2\beta} - A$ on the interval $|\xi| \le \omega$, using nonlinear least squares algorithms. The resulting fit is close only for ξ away from the origin. The returned values for α and β are then used in the SECB deblurring algorithm. Different values of A return different pairs (α,β). Experience indicates that useful values of A generally lie in the interval $2 \le A \le 6$. Increasing the value of A decreases the curvature of $u(\xi)$ at $\xi = 0$, resulting in a larger value of β together with a smaller value of α. A value of $A>0$ that returns $\beta > 1$ is clearly too large, as $\beta > 1$ is impossible for probability density functions [17]. Decreasing A has the opposite effect, producing lower values of β and higher values of α. As a rule, deconvolution is better behaved at lower values of β than it is when $\beta \approx 1$. A significant observation is that *an image blurred with a pair* (α_0,β_0) *can often be successfully deblurred with an appropriate pair* (α,β)*, where* $\alpha > \alpha_0$ *and* $\beta < \beta_0$. Examples of this phenomenon were shown in Figure 1 in connection with the blurred Sydney image. An effective interactive framework for performing the above least squares fitting is the *fit* command in *DATAPLOT* [20]. This is a high-level English-syntax graphics and analysis software package developed at the National Institute of Standards and Technology. This software tool was used throughout this paper.

The following version of the APEX method, using the SECB *marching backwards in time* option (13), has been found useful in a variety of image enhancement problems where the image $g(x,y)$ is such that $\log|\hat{g}^*(\xi,0)|$ is generally globally monotone decreasing and convex, as shown in Figure 3. Choose a value of $A>2$ in (25) such that the least squares fit develops a slight *cusp* at $\xi = 0$. Using the returned pair (α,β) in the SECB method, obtain a sequence $u^\dagger(x,y,t)$ of partial restorations as t decreases from $t=1$, as illustrated in the Cindy Crawford sequence[1] in Figure 4. Often, the initial choice of A results in a psf that is *too wide* in physical space, i.e., wider than the unknown psf that might have blurred the image. Use of that psf all the way to $t=0$ will result in *oversharpening*. Typically, high quality restorations will be found at positive values of t, and these will gradually deteriorate as $t \downarrow 0$. At $t=0$, the restoration may exhibit severe ringing and other undesirable artifacts [9, Figure 13], indicating that continuation backwards in time has proceeded *too far* in (7). Terminating the continuation at some appropriate $t = t_1 > 0$ is equivalent to rescaling the value of α without changing the value of β. If the pair (α,β) produces a high quality restoration at $t = t_1 > 0$, the pair (α_1,β), where $\alpha_1 = (1-t_1)\alpha$, will produce the same quality results at $t=0$. Thus, marching backwards in time is equivalent to simultaneously sampling numerous values of α while keeping β fixed. This process

[1] Given a 512×512 blurred image as input, the APEX procedure computes and displays a time marching sequence of 10 partial restorations in about 10 seconds on an MIPS R12000 (400MHz) workstation.

FIG. 4. *Enhancement of a Cindy Crawford image by marching backwards from $t = 1$ with an APEX-detected psf. Image sequence shows a gradual increase in contrast as t decreases. Undesirable artifacts at $t = 0$ indicate that continuation backwards in time has proceeded too far. Best results are highly subjective in this case, but probably occur at some $t > 0.5$. Note the sharpness of the earrings near $t = 0.5$.*

can be repeated with a different choice of A, resulting in a different value of β. In general, there will be many values of A in (25) returning pairs (α, β) that produce good reconstructions at some $t_{\alpha\beta} > 0$. A large number of distinct pairs (α^*, β^*) can thus be found that produce useful, but distinct, results at $t = 0$. Indeed, this is the process that was used to obtain the four psfs shown in Figure 2.

We have been assuming $\hat{h}(\xi, \eta)$ to be a pure Lévy otf in (18). For more general class **G** otfs (2), we may still use the approximation $\log|\hat{g}^*(\xi, 0)| \approx -\alpha|\xi|^{2\beta} - A$ and apply the same technique to extract a suitable pair (α, β) from the blurred image. Here, the returned APEX values may be considered representative values for the α_i, β_i in (2), producing a single pure Lévy otf approximating the composite otf.

6. Application to real images. The developments in sections 2 through 5 are predicated on two assumptions. The first assumption is that the blurred image $g(x, y)$ obeys the simple convolution equation (3) rather than a more general, possibly nonlinear, integral equation

$$Hf = \int_{R^2} h(x, y, u, v, f(u, v))dudv = g(x, y). \tag{26}$$

In addition to linearity, (3) implies that the blur is isoplanatic. The second assumption is that the point spread function $h(x, y)$ belongs to a restricted class of unimodal, radially symmetric, probability density functions, the class **G** defined in (2). In [9], successful blind deconvolution of *synthetically blurred* images, with added noise, was demonstrated. Such synthetically blurred images necessarily obey (2) and (3).

The applicability of the preceding theory to real blurred images is by no means assured. Deviations from linearity, isoplanatism, unimodality, and radial symmetry

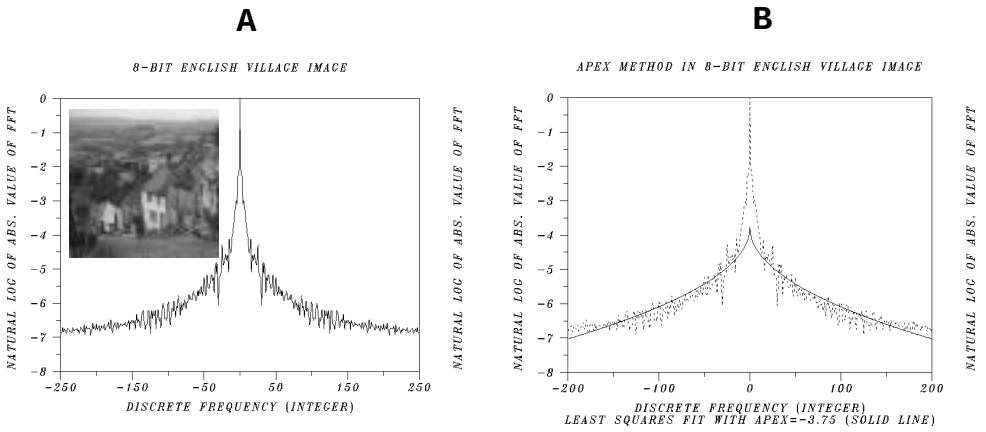


FIG. 5. *The APEX method of psf detection.* (a) $\log|\hat{g}^*(\xi,0)|$ *on* $|\xi| \le 250$ *in the* 8*-bit English village image.* (b) *A least squares fit of* $\log|\hat{g}^*(\xi,0)|$*, with* $u(\xi) = -\alpha\ |\xi|^{2\beta} - 3.75$ *on* $|\xi| \le 200$*, develops a cusp at* $\xi = 0$ *and returns* $\alpha = 0.251274$*,* $\beta = 0.242246$*.*

are possible. Moreover, the class **G** excludes motion and defocus blurs. In addition, the types and intensities of noise processes in real images may differ fundamentally from the noise models typically used in numerical experiments. Therefore, only limited success on a narrow class of images can be expected in real applications.

The examples discussed below involve images obtained from multiple sources using diverse imaging modalities. Some of these images have been used as test images in the literature. In this paper, each of these images is assumed to have been blurred by some unknown process, and we seek to improve visual quality by APEX processing. All images are of size 512×512 and are quantized at 8 bits per pixel.

Our first example is a well-known English village image denoted by $g(x, y)$ and shown in Figure 5(a) together with $\log|\hat{g}^*(\xi,0)|$ on $|\xi| \le 250$. The plot displays globally convex monotone behavior. In Figure 5(b), the APEX fit of $\log|\hat{g}^*(\xi,0)|$ with $u(\xi) = -\alpha|\xi|^{2\beta} - A$ on the interval $|\xi| \le 200$ is shown. With $A = 3.75$, the fit develops a cusp at $\xi = 0$ and returns $\alpha = 0.251274$, $\beta = 0.242246$. With these psf parameters, SECB deblurring using $s = 0.01$, $K = 1300$, and continuation backwards in time terminated at $t = 0.5$ produces Figure 6(b). This is compared with the original in Figure 6(a).

The extent of sharpening in Figure 6(b) becomes evident when zooming in on selected parts of the image. In Figure 7, roof lines on the first three houses are compared before and after APEX processing. There is noticeable enhancement of structural detail in the roof shingles and stone fronts of the three houses in Figure 7(b). In Figure 8(b), Holstein cows grazing in the meadow, not previously identifiable, are clearly visible. So are individual chimney bricks. In Figure 9(b), buildings in the distance, not readily noticed in Figure 9(a), become well defined.

It should be noted that the use of a different value of A, and/or a different neighborhood of the origin Ω in Figure 5(b), may return a different psf pair (α, β). In that case, backwards continuation in the SECB method may need to be terminated at some other value of t to obtain the best image. However, with good choices of A and Ω, the new image would again be a high quality representation of the visual scene in Figure 6(b), while differing from Figure 6(b) at individual pixels. This is

FIG. 6. *Enhancement of the English village image.* (a) *Original* 8*-bit image.* (b) *SECB deblurred image using* $s = 0.01$*,* $K = 1300$*, with APEX-detected values* $\alpha = 0.251274$*,* $\beta = 0.242246$*, and with continuation backwards in time terminated at* $t = 0.5$*.*

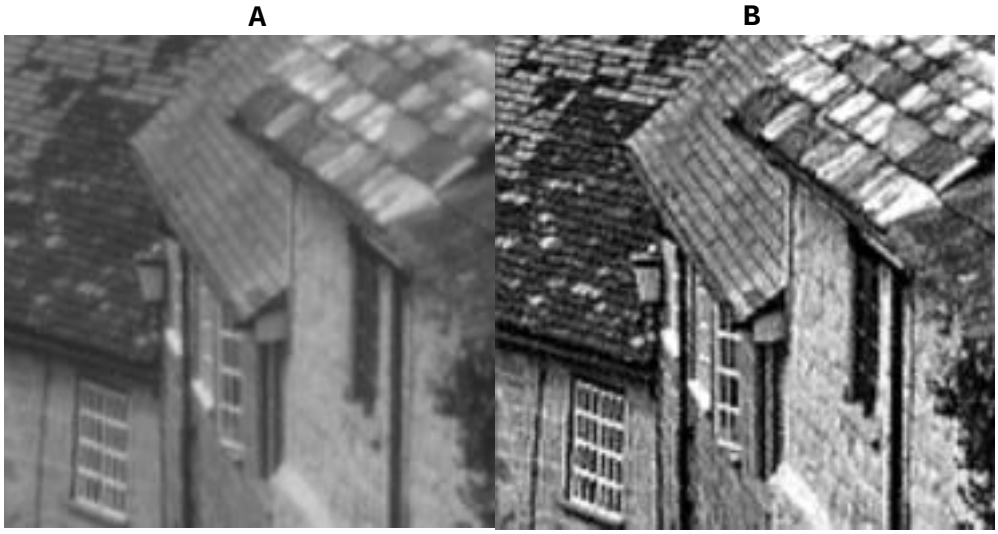


FIG. 7. *Extent of sharpening in the English village scene becomes evident when zooming in on selected parts of the image.* (a) *Roof lines in the original image.* (b) *Roof lines in the enhanced image.*

the nonuniqueness phenomenon previously discussed in connection with the Sydney image in Figure 1.

Deconvolution of Figure 6(a) with the above APEX-detected psf significantly alters its Fourier transform. As shown in Figure 10(a), the Fourier transform in Figure 6(b) (dashed curve) decays less rapidly as $|\xi|$ increases than was the case in the original Figure 6(a) (solid curve). Evidently, APEX processing amplifies high frequency image components in a stable coherent fashion, resulting in the overall

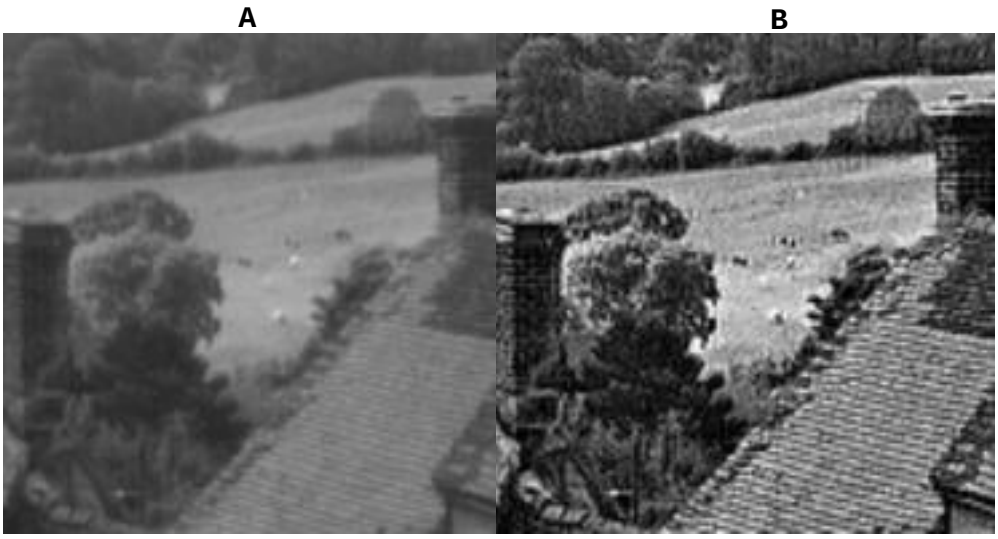


FIG. 8. *Extent of sharpening in the English village scene becomes more evident with zooming. Holstein cows grazing in the meadow in image* (b) *are not readily identifiable in image* (a).

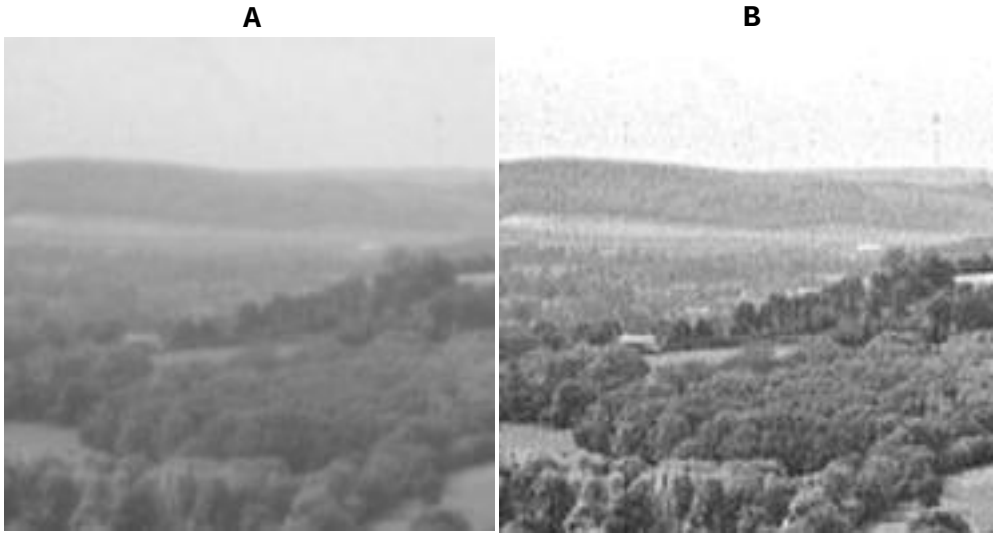


FIG. 9. *Extent of sharpening in the English village image becomes more evident with zooming. The enhanced image* (b) *shows buildings in the distance not immediately apparent in the original image* (a).

improvements visible in Figures 6 through 9. The "before and after" Fourier transform pattern shown in Figure 10(a) occurs in every example discussed in this paper, with the exception of the F15 image in Figure 12; the anomalous behavior in that case is shown in Figure 10(b).

The next example is the boat image in Figure 11(a). With $A = 4.0$, the APEX fit on $|\xi| \leq 250$ returned $\alpha = 0.518155$, $\beta = 0.215083$. Using these values in the SECB method, with $s = 0.01$, $K = 1300$, and continuation terminated at $t = 0.5$, produced

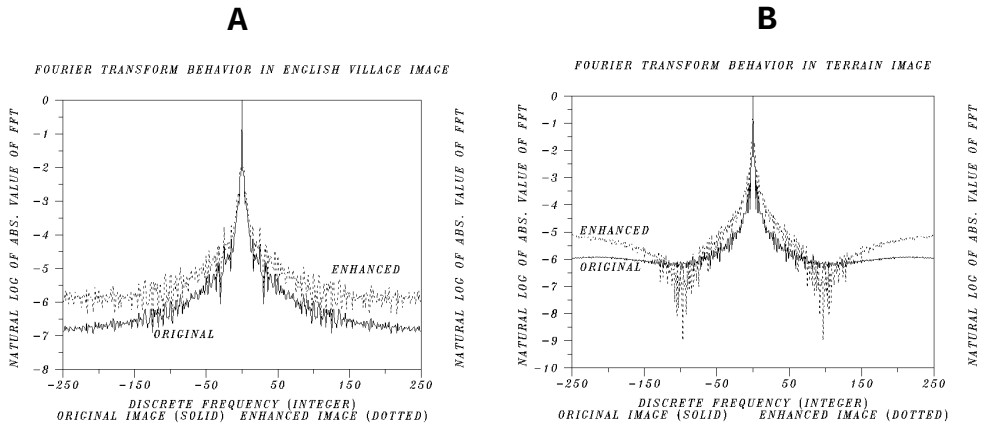


FIG. 10. *APEX processing significantly alters Fourier transform behavior.* (a) *English village image before and after processing.* (b) *F*15 *terrain image in Figure* 12 *before and after processing. The behavior shown in* (b) *is exceptional; all other examples in the present paper conform with the behavior shown in* (a).

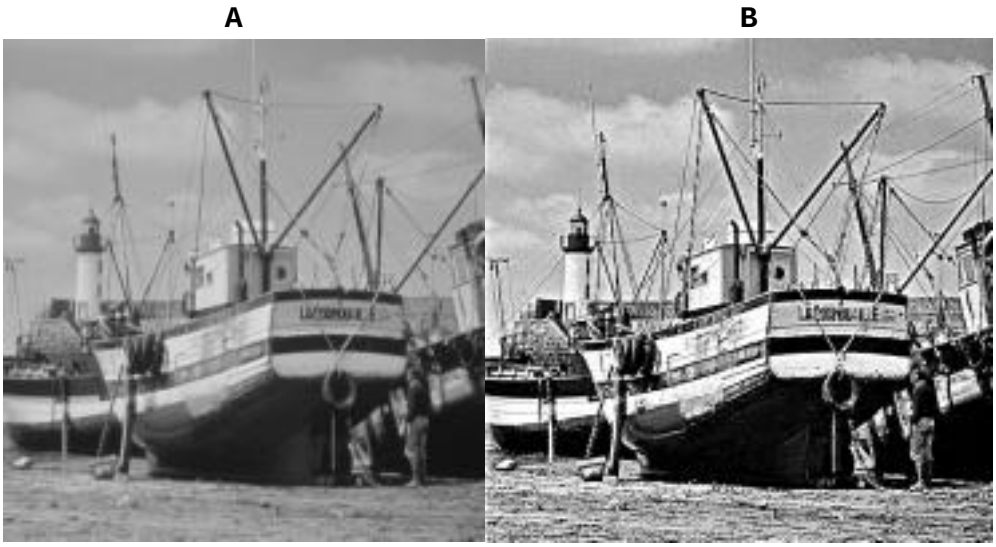


FIG. 11. *Enhancement of the boat image. The APEX method with* $A = 4.0$ *on* $|\xi| \leq 250$ *yields* $\alpha = 0.518155$, $\beta = 0.215083$. *Using these parameters, with* $s = 0.01$, $K = 1300$, *and backwards continuation terminated at* $t = 0.5$, *the SECB method applied to image* (a) *produces image* (b). *The number* 7 2 7 *on the side of the boat in image* (b) *was not easily identifiable in image* (a).

Figure 11(b). Enhancement has now rendered visible the number 7 2 7 on the left side of the boat. Other identifiable details include the stripe along the left trouser leg of the man on the ground, the lettering on the sign hanging from the boat to his right, and part of the stone work and stairway to the left of the lighthouse.

The F15 plane image in Figure 12(a) is another interesting example. The aim here is to enhance the background terrain. With $A = 3.5$, the APEX fit on $|\xi| \leq 250$ develops a cusp at $\xi = 0$ and returns $\alpha = 0.856096$, $\beta = 0.107289$. Using these

FIG. 12. *Striking enhancement of terrain features in an F*15 *image. The APEX method with* $A = 3.5$ *on* $|\xi| \leq 250$ *yields* $\alpha = 0.856096$, $\beta = 0.107289$. *Using these parameters, with* $s = 0.01$, $K = 1000$, *and backwards continuation terminated at* $t = 0.25$, *the SECB method applied to image* (a) *produced image* (b). *Condensation trails behind the aircraft in image* (b) *were not immediately evident in image* (a).

values in the SECB method, with $s = 0.01$, $K = 1000$, and backwards continuation terminated at $t = 0.25$, produces rather striking enhancement of the ground features in Figure 12(b). This example is noteworthy on two counts: the exceptionally low value of β detected by the APEX method and the previously mentioned unexpected Fourier behavior shown in Figure 10(b).

Beginning with Figure 1, all of the examples discussed so far involve images of familiar objects. This allows for relatively easy evaluation of the results of APEX processing. The next five examples involve less familiar objects. Moreover, fine details visible on a modern high resolution computer screen are sometimes lost in the printing process. Consequently, improvements in image quality in some of the next examples may seem less obvious than in previous examples. At the same time, the performance of the APEX method in reconstructing real details of familiar objects provides a measure of confidence in the results obtained when that method is applied to unfamiliar objects.

Figure 13(a) is a Landsat image of the Washington, DC area. With $A = 4.25$, the APEX fit on $|\xi| \leq 250$ returns $\alpha = 0.540825$, $\beta = 0.182410$. Using these parameters in the SECB method, with $s = 0.01$, $K = 1300$, and continuation terminated at $t = 0.5$, produces Figure 13(b). There is a significant increase in resolution in Figure 13(b), which improves definition of several landmarks and thoroughfares. The Washington Monument, the bridges over the Potomac, Pennsylvania and Maryland Avenues radiating from the Capitol, Massachusetts Avenue to the north, and Virginia Avenue and the Southeast Freeway to the south are some of the features that are more easily identified in the enhanced image.

Figure 14(a) is a scanning electron microscope image of a mosquito's head. A prominent feature is the insect's compound eye. With $A = 4.0$, the APEX fit on $|\xi| \leq 250$ yields $\alpha = 0.734259$, $\beta = 0.156963$. Using these values in the SECB method, with

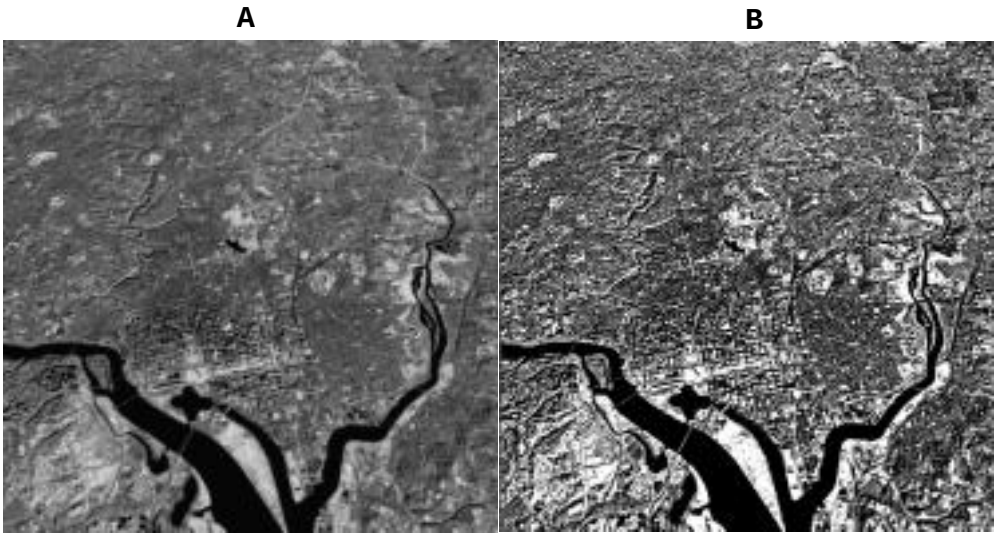


FIG. 13. *Enhancement of a Washington, DC Landsat image. The APEX method with* $A = 4.25$ *on* $|\xi| \leq 250$ *yields* $\alpha = 0.540825$, $\beta = 0.182410$. *Using these parameters, with* $s = 0.01$, $K = 1300$, *and backwards continuation terminated at* $t = 0.5$, *the SECB method applied to image* (a) *produced image* (b). *Increased resolution in image* (b) *improves definition of several landmarks and thoroughfares.*

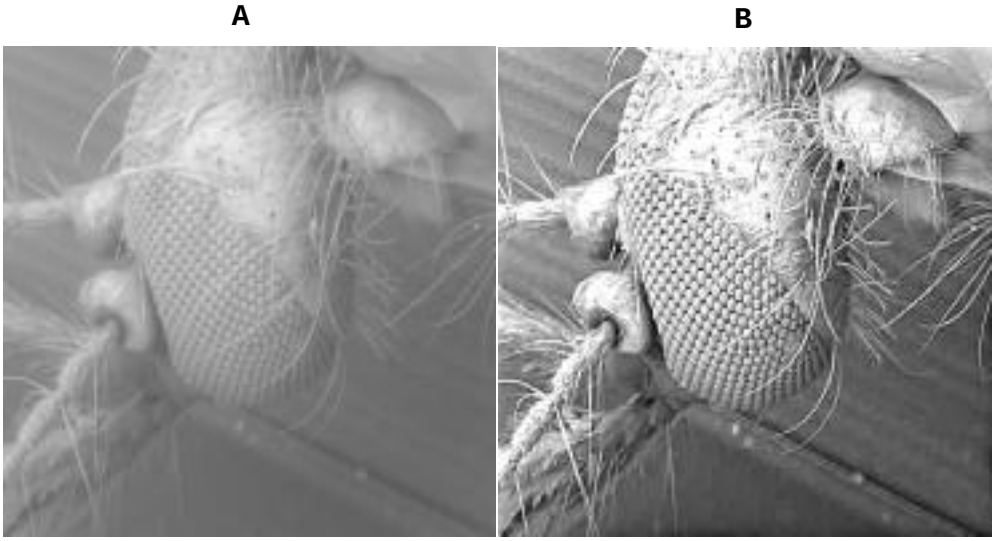


FIG. 14. *Enhancement of a scanning electron microscope image of a mosquito's head showing the compound eye. The APEX method with* $A = 4.0$ *on* $|\xi| \leq 250$ *yields* $\alpha = 0.734259$, $\beta = 0.156963$. *Using these parameters, with* $s = 0.001$, $K = 10$, *and backwards continuation terminated at* $t = 0.4$, *the SECB method applied to image* (a) *produced image* (b). *APEX processing enhances contrast and brings the eye into sharper focus. Further applications in electron microscopy are discussed in* [10].

$s = 0.001$, $K = 10.0$, and backwards continuation terminated at $t = 0.4$, produces Figure 14(b). Evidently, APEX processing results in significant overall improvement. In particular, the eye appears in much sharper focus. Further applications to electron

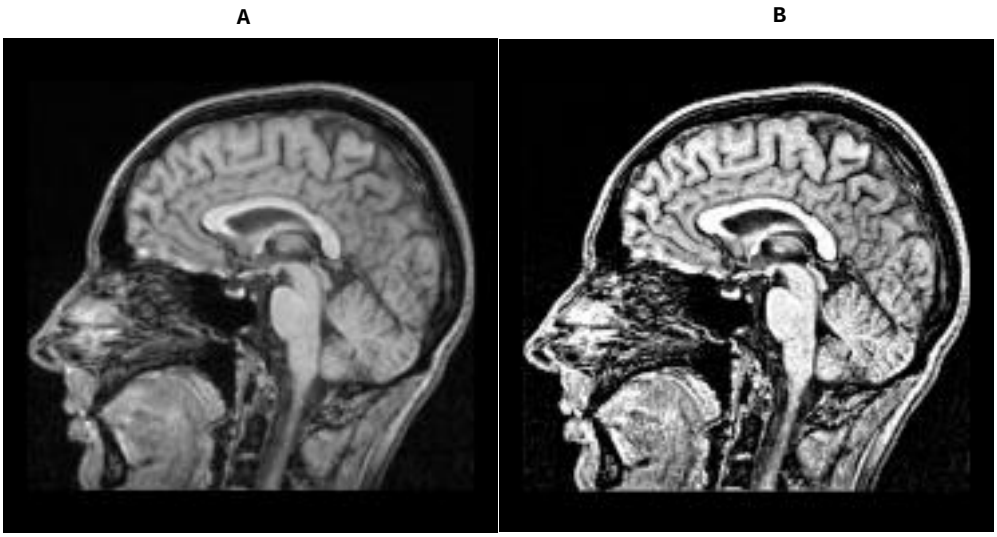


FIG. 15. *Enhancement of a sagittal MRI brain image. The APEX method with* $A = 4.0$ *on* $|\xi| \leq 250$ *yields* $\alpha = 0.333267$, $\beta = 0.209416$. *Using these parameters, with* $s = 0.01$, $K = 1300$, *and backwards continuation terminated at* $t = 0.35$, *the SECB procedure applied to image* (a) *produced image* (b). *APEX processing noticeably improves feature definition in areas between two and four o'clock.*

microscopy are discussed in [10].

The sagittal MRI (magnetic resonance imaging) brain image in Figure 15(a) has been used as a test *sharp* image in previous publications. In [5] and [7], synthetically blurred versions of that sharp image were used in a comparative evaluation of restoration algorithms when the psf is *known*. Here, we consider further sharpening the sharp image by blind deconvolution. With $A = 4.0$, the APEX fit on $|\xi| \leq 250$ returns $\alpha = 0.333267$, $\beta = 0.209416$. Using these parameters in the SECB procedure, with $s = 0.01$, $K = 1300$, and continuation terminated at $t = 0.35$, produced the image in Figure 15(b). Substantial improvement is apparent over the whole image. In the sector between two and four o'clock, in particular, sharpening of structural detail significantly improves feature definition.

In PET (positron emission tomography) imaging, a positron emitting radionuclide is injected into the patient and used to tag glucose molecules in their course through the brain. The metabolic rate of glucose is a key parameter that reflects cerebral function and indicates the extent to which regions of the brain are active. Performing specific mental tasks activates various parts of the brain, causing increased glucose uptake and hence increased positron emission. Centers of activity translate into relatively bright spots in the PET image. However, blurring by the scanner psf tends to average out such relative differences, resulting in loss of contrast. Figure 16(a) is a PET image of a transverse (horizontal) slice through the brain. Blind deconvolution is used to enhance that image. With $A = 5.0$, the APEX fit on $|\xi| \leq 250$ returns $\alpha = 0.198931$, $\beta = 0.284449$. Using these parameters in the SECB method, with $s = 0.001$, $K = 5.0$, and backwards continuation terminated at $t = 0.6$, produces Figure 16(b). Note that both images in Figure 16 show identical features, but the contrast has been increased in the APEX-processed image, with some regions becoming darker while others have become lighter. In particular, several bright spots appear

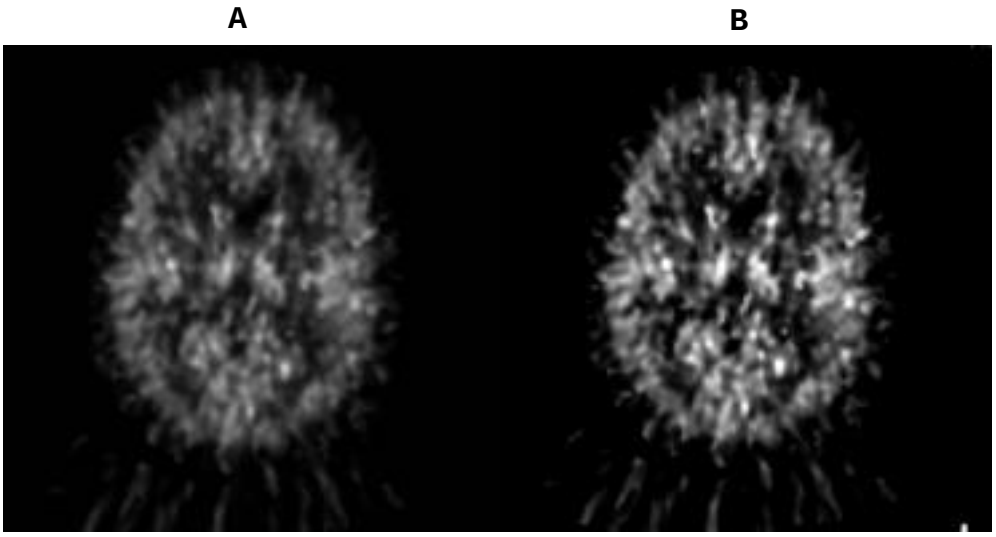


FIG. 16. *Enhancement of a transverse PET brain image. The APEX method with* $A = 5.0$ *on* $|\xi| \leq 250$ *yields* $\alpha = 0.198931$, $\beta = 0.284449$. *Using these parameters, with* $s = 0.001$, $K = 5.0$, *and backwards continuation terminated at* $t = 0.6$, *the SECB procedure applied to image* (a) *produced image* (b). *Bright spots in the enhanced image* (b), *indicating areas of the brain responding to applied external stimuli, are more clearly defined than in the original image* (a).

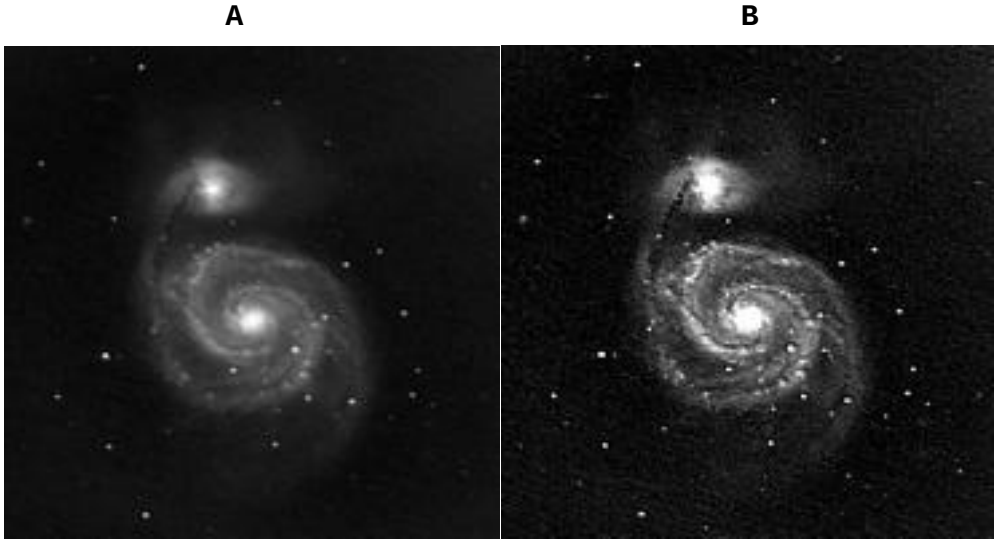


FIG. 17. *Enhancement of the Whirlpool galaxy (M*51*) image. The APEX method with* $A = 4.0$ *on* $|\xi| \leq 250$ *yields* $\alpha = 0.451615$, $\beta = 0.221955$. *Using these parameters, with* $s = 0.001$, $K = 5.0$, *and backwards continuation terminated at* $t = 0.5$, *the SECB applied to image* (a) *produced image* (b). *APEX processing increases resolution and enhances luminosity in the spiral arms and galactic cores.*

in Figure 16(b) that were not as readily apparent in the original image.

Our last example is the Whirlpool galaxy (M51) in Figure 17(a). With $A = 4.0$, the APEX fit on $|\xi| \leq 250$ yields $\alpha = 0.451615$, $\beta = 0.221955$. Using these values in

the SECB method, with $s = 0.001$, $K = 5.0$, and backwards continuation terminated at $t = 0.5$, produced Figure 17(b). In the enhanced image, the spiral arms are more luminous and better defined, and the luminous cores are larger in both the spiral galaxy and its companion. The dark connecting arm between the two galaxies is also more clearly defined. These enhancements are due to a change in Fourier transform behavior brought about by deconvolution with the APEX-detected psf. This change in Fourier behavior is similar to that shown in Figure 10(a), although it is more pronounced. A concomitant effect of deconvolution is amplification of data noise, which now becomes visible against the dark background in Figure 17(b).

Clearly, in this galaxy image as in the preceding PET image, there is no way of knowing whether the enhanced image conforms with reality. Conceivably, the increased luminosity in Figure 17(b) may be exaggerated. However, the bright areas along the galactic arms in Figure 17(b), as well as the bright spots in Figure 16(b), did not materialize spontaneously. These areas must have been just below some brightness threshold in the original image, and APEX processing has served the very useful purpose of revealing their presence. If such areas appear overenhanced, this can be corrected by repeating the SECB procedure and terminating continuation at *higher* values of t.

7. Anisotropic diffusion, total variation deblurring, and the "staircase effect." As is evident from the survey [11], there is considerable interest in the use of anisotropic diffusion equations to perform various tasks in image processing. In pure denoising applications, such methods have been found to be effective at removing high levels of noise while preserving edges in an image. An important related idea is the use of the *total variation* norm for regularizing the image restoration problem [12], [13], [16], [33], [37], [38]. The Euler–Lagrange problem for minimizing the total variation can be written as a nonlinear anisotropic diffusion equation, with a forcing term that describes convolution of the unknown image with the known psf. This is supplemented by homogeneous Neumann boundary conditions together with the blurred image as initial data; see [33]. Deblurring the image is equivalent to stepwise numerical computation of this nonlinear initial value problem until a steady state is reached.

Total variation deblurring is especially useful for recovering "blocky" images, i.e., images that are nearly piecewise constant and have many edges [12], [16]. For this reason, the total variation blind deconvolution approach in [13] aims primarily at recovering blocky images that had been blurred by psfs with *sharp edges*. This is the case with defocus and motion blurs; a defocused satellite image is the example used in [13]. The authors observe that their algorithm is more effective on defocused images than it is on Gaussian blurred images. In a complementary role, the APEX method can also handle blocky images, but it is based on detecting class **G** blurs, a class that includes heavy-tailed psfs but excludes defocus and motion blurs.

A major drawback of the total variation approach is the so-called "staircase effect," whereby the deblurred image can develop spurious piecewise constant regions. This often produces an "oil painting" appearance that does not correspond to the true image and prevents identification of fine detail. For this reason, the authors in [12] and [16] conclude that total variation deblurring is not useful for images that are not nearly piecewise constant. In [34], [35], it is proved that total variation restoration necessarily leads to the staircase effect. In [19], the mathematical premise of minimizing image total variation is questioned, and the authors prove that because of their fine texture, most natural images are *not* of bounded variation. Therefore, in images

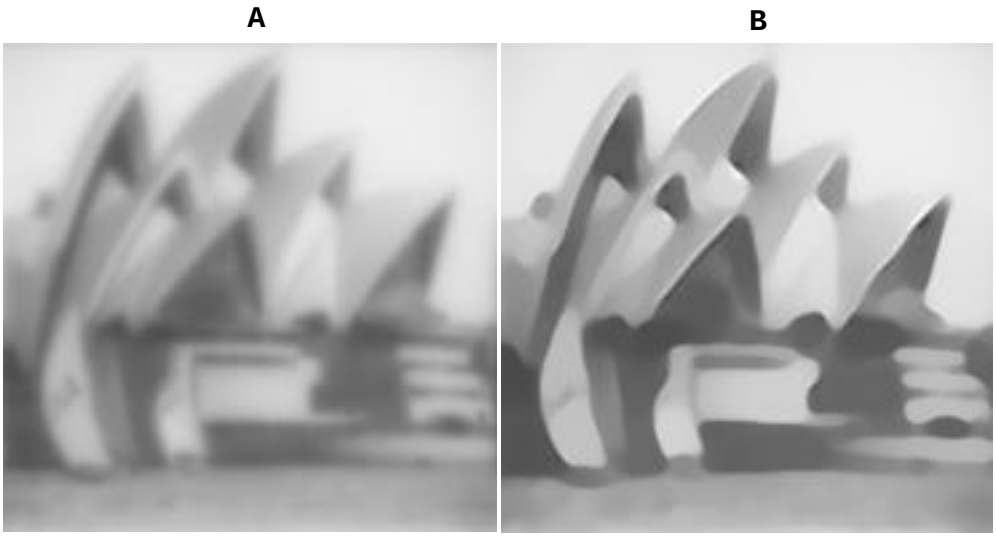


FIG. 18. *Staircase effect in total variation deblurring of the Sydney image.* (a) *Synthetically blurred* 512×512 *Sydney image, previously used in Figure* 1(b)*, was computed in* 64*-bit precision.* (b) *Deblurring of image* (a) *using the known psf and the total variation scheme in* [33]*, with noise variance parameter* $\beta^{\Sigma} = 0.001$*, Lagrange multiplier* $\lambda = 50$*, CFL restriction* $\Delta t = 0.1(\Delta x)^2$*, and stepwise integration to time* $T = 100\Delta t$*. The strong "oil painting" effect in image* (b) *impairs recognition and occurs with other choices of* $\beta^{\Sigma} \leq 0.001$ *and* $\lambda \geq 50$*. Compare with the SECB deblurred image in Figure* 1(c).

with fine texture, total variation deblurring must inevitably smooth out texture.

The following example illustrates why total variation deblurring is typically not useful for the type of textured imagery considered in this paper. The blurred noiseless 512×512 Sydney image, previously used in Figure 1(b), was deblurred using the total variation scheme described in [33, section 5]. This is a pure deblurring problem in which the synthetically blurred input image, Figure 18(a), was computed in 64-bit precision. Moreover, the precisely known psf was used. The aim here is to evaluate the reconstructive ability of the total variation scheme under the most favorable circumstances. As recommended in [33], in this noiseless case the noise variance parameter β^{Σ} should be chosen small, while the Lagrange multiplier λ should be chosen large. Here, several values of β^{Σ} in the range $0.00001 \leq \beta^{\Sigma} \leq 0.01$ were tried, together with several values of λ in the range $1 \leq \lambda \leq 100$. The CFL restriction $\Delta t = 0.1(\Delta x)^2$ was applied with all these choices, and no sign of computational instability was detected. Figure 18(b) is the result of stepwise numerical computation of the nonlinear diffusion problem in [33, section 5] up to time $T = 100\Delta t$, using $\beta^{\Sigma} = 0.001$ and $\lambda = 50$. The "oil painting" effect in Figure 18(b) occurs with other choices of $\beta^{\Sigma} \leq 0.001$ and $\lambda \geq 50$, and the deblurred image does not improve if more time steps are taken. SECB deblurring of the same image is shown in Figure 1(c). In the presence of noise, the SECB deblurred image is less sharp, but maintains its strong qualitative edge over Figure 18(b). It should be noted that the authors in [33] did not intend their scheme to be used for images as seriously blurred as Figure 18(a). However, the staircase effect is still pronounced, even with more mildly blurred images.

8. Concluding remarks. Setting aside all theoretical considerations, APEX processing is a practical enhancement technique that can sharpen significant classes

of images originating from diverse imaging modalities. One important feature of this approach is its fast implementation on desktop platforms. Even with large image sizes, numerous trial restorations can be accomplished in seconds or minutes of cpu time. This makes for easy fine tuning of parameters and a quick determination of whether the APEX method significantly improves a given image. Once improvement is detected, fine tuning must be used to obtain optimal results. Here, another important feature of the APEX method plays a useful role. This is the marching-backwards-in-time option characteristic of class **G** psfs, which allows for deconvolution to be performed in *slow motion*. Robustness is a third important property of the APEX method, allowing detection of multiple psfs capable of significant sharpening. This substantially increases the probability of finding a useful candidate.

On the theoretical side, this paper raises new questions. The first of these is the existence of several useful psfs, as demonstrated for the Sydney image in Figure 1. This phenomenon warrants further investigation. A second question concerns the important role Lévy psfs appear to play in numerous imaging systems. The discussion in section 2 has surveyed *inferences* of stable laws that have been made from mtf measurements. Development of methods of analyzing imaging systems that can rigorously establish such laws, and predict the Lévy exponent β, would be a major advance.

Reconciling the results of section 2 with the behavior of large classes of images raises additional questions. Electronic imaging psfs $h(x,y)$ are found to have Lévy exponents $\beta > 0.5$ in most cases, so that $\log \hat{h}(\xi,0) = -\alpha|\xi|^{2\beta}$ is a monotone decreasing concave function on $\xi > 0$. However, as illustrated in Figure 3, all images $g(x,y)$ used in this paper are such that global behavior in $\log|\hat{g}^*(\xi,0)|$ is generally monotone decreasing and *convex*. Another large class of images with this convexity property, the class **W**, was described in [9]. When such images are APEX-fitted with a Lévy psf in the manner shown in Figure 5(b), a value of $\beta \leq 0.5$ is inevitably detected. An average value of $\beta = 0.23$ was found for the six images in Figures 4, 6, 11, 15, 16, and 17, and significantly lower values were found for the three images in Figures 12, 13, and 14. A possible partial explanation for this discrepancy is provided by the Sydney experiment in Figure 1. There, the APEX method detected several useful psfs with values of β *smaller* than the value that was used to blur the image. The detected β-values in the above nine images may likewise underestimate the true imaging system β-values. An entirely different scenario may be that the APEX method provides generic low exponent Lévy psfs capable of enhancing a wide variety of images, independently of the imaging physics that created them. Other generic enhancement techniques have been used for some time in image processing (see [36, Chapter 10]). More recent approaches based on nonlinear diffusion equations are also intended as generic enhancement methods [11]. However, nonlinear methods generally require large numbers of iterations and may not be well suited for real-time processing of complex high resolution imagery.

Whatever may be the reasons behind it, the effectiveness of the APEX method on many types of images is undeniable, and the method is a useful addition to the image processing toolbox.

REFERENCES

[1] O. Barndorff-Nielsen, T. Mikosch, and S. Resnick, eds., *Lévy Processes—Theory and Applications*, Birkhäuser Boston, Cambridge, MA, 2001.

[2] J. Bertouin, *Lévy Processes*, Cambridge Tracts in Math. 121, Cambridge University Press, Cambridge, UK, 1998.
[3] L. M. Biberman and S. Nudelman, *Photoelectronic Imaging Devices*, Plenum Press, New York, 1971.
[4] P. Biler, G. Karch, and W. A. Woyczyński, *Multifractal and Lévy conservation laws*, C. R. Acad. Sci. Paris Sér. I Math., 330 (2000), pp. 343–348.
[5] A. S. Carasso, *Overcoming Hölder continuity in ill-posed continuation problems*, SIAM J. Numer. Anal., 31 (1994), pp. 1535–1557.
[6] A. S. Carasso, *Error bounds in nonsmooth image deblurring*, SIAM J. Math. Anal., 28 (1997), pp. 656–668.
[7] A. S. Carasso, *Linear and nonlinear image deblurring: A documented study*, SIAM J. Numer. Anal., 36 (1999), pp. 1659–1689.
[8] A. S. Carasso, *Logarithmic convexity and the "slow evolution" constraint in ill-posed initial value problems*, SIAM J. Math. Anal., 30 (1999), pp. 479–496.
[9] A. S. Carasso, *Direct blind deconvolution*, SIAM J. Appl. Math., 61 (2001), pp. 1980–2007.
[10] A. S. Carasso, D. S. Bright, and A. E. Vladár, *The APEX method and real-time blind deconvolution of scanning electron microscope imagery*, Optical Engineering, 41 (2002), pp. 2499–2514.
[11] V. Caselles, J. M. Morel, G. Sapiro, and A. Tannenbaum, eds., *Special Issue on Partial Differential Equations and Geometry-Driven Diffusion in Image Processing*, IEEE Trans. Image Process., 7 (1998).
[12] A. Chambolle and P. L. Lions, *Image recovery via total variation minimization and related problems*, Numer. Math., 76 (1997), pp. 167–188.
[13] T. F. Chan and C. K. Wong, *Total variation blind deconvolution*, IEEE Trans. Image Process., 7 (1998), pp. 370–375.
[14] I. P. Csorba, ed., *Electron Image Tubes and Image Intensifiers*, Proc. SPIE 1243, 1990.
[15] I. P. Csorba, ed., *Electron Image Tubes and Image Intensifiers* II, Proc. SPIE 1449, 1991.
[16] D. C. Dobson and F. Santosa, *Recovery of blocky images from noisy and blurred data*, SIAM J. Appl. Math., 56 (1996), pp. 1181–1198.
[17] W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 2, 2nd ed., Wiley, New York, 1971.
[18] R. E. Franseen and D. K. Schroder, eds., *Applications of Electronic Imaging Systems*, Proc. SPIE 143, 1978.
[19] Y. Gousseau and J.-M. Morel, *Are natural images of bounded variation?*, SIAM J. Math. Anal., 33 (2001), pp. 634–648.
[20] A. Heckert and J. J. Filliben, *DATAPLOT Reference Manual*, http://www.itl.nist.gov/div898/software/dataplot/document.htm.
[21] R. E. Hufnagel and N. R. Stanley, *Modulation transfer function associated with image transmission through turbulent media*, J. Opt. Soc. Amer., 54 (1964), pp. 52–61.
[22] C. B. Johnson, *A method for characterizing electro-optical device modulation transfer functions*, Photographic Science and Engineering, 14 (1970), pp. 413–415.
[23] C. B. Johnson, C. E. Catchpole, and C. C. Matle, *Microchannel plate inverter image intensifiers*, IEEE Trans. Electron Devices, 18 (1971), pp. 1113–1116.
[24] C. B. Johnson, *Classification of electron-optical device modulation transfer function*, Adv. Electronics and Electron Phys., 33B (1972), pp. 579–584.
[25] C. B. Johnson, *Circular aperture diffraction limited MTF: Approximate expressions*, Applied Optics, 11 (1972), pp. 1875–1876.
[26] C. B. Johnson, *MTFs: A simplified approach*, Electro-Optical Systems Design, 4 (1972), pp. 22–26.
[27] C. B. Johnson, *Point-spread functions, line-spread functions, and edge-response functions associated with mtfs of the form* $exp[-(\omega/\omega_c)^n]$, Applied Optics, 12 (1973), pp. 1031–1033.
[28] C. B. Johnson, *A convenient form of graph paper for determination of electro-optical device modulation transfer function parameters*, IEEE Trans. Electron Devices, 20 (1973), pp. 80–81.
[29] C. B. Johnson, *MTF parameters for all photographic films listed in Kodak pamphlet P-49*, Applied Optics, 15 (1976), p. 1130.
[30] C. B. Johnson, S. B. Patton, and E. Bender, *High-resolution microchannel plate image tube development*, in Proc. SPIE 1449, 1991, pp. 2–12.
[31] C. B. Johnson and B. N. Laprade, eds., *Electron Tubes and Image Intensifiers*, Proc. SPIE 1655, 1992.
[32] R. Kuske and J. B. Keller, *Rate of convergence to a stable law*, SIAM J. Appl. Math., 61 (2000), pp. 1308–1323.

[33] A. Marquina and S. Osher, *Explicit algorithms for a new time dependent model based on level set motion for nonlinear deblurring and noise removal*, SIAM J. Sci. Comput., 22 (2000), pp. 387–405.

[34] M. Nikolova, *Local strong homogeneity of a regularized estimator*, SIAM J. Appl. Math., 61 (2001), pp. 633–658.

[35] M. Nikolova, *Image restoration by minimizing objective functions with nonsmooth data-fidelity terms*, in Proceedings of the IEEE Workshop on Variational and Level Set Methods (VLSM01), International Conference on Computer Vision (ICCV 2001), Vancouver, IEEE, Piscataway, NJ, 2001.

[36] W. K. Pratt, *Digital Image Processing*, 2nd ed., Wiley, New York, 1991.

[37] L. Rudin and S. Osher, *Total variation based image restoration with free local constraints*, in Proceedings of the IEEE International Conference on Image Processing, Austin, TX, 1994, Vol. 1, pp. 31–35.

[38] L. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Phys. D, 60 (1992), pp. 259–268.

[39] G. Samorodnitsky and M. S. Taqqu, *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*, Chapman and Hall, New York, 1994.

[40] M. F. Shlesinger, G. M. Zaslavsky, and U. Frisch, eds., *Lévy Flights and Related Topics in Physics*, Lecture Notes in Phys. 450, Springer-Verlag, New York, 1995.

[41] C. Tsallis, S. V. Levy, A. M. Souza, and R. Maynard, *Statistical-mechanical foundation of the ubiquity of Lévy distributions in nature*, Phys. Rev. Lett., 75 (1995), pp. 3589–3593.

[42] R. Weber, *The ground-based electro-optical detection of deep-space satellites*, in Proc. SPIE 143, 1978, pp. 59–69.

[43] C. S. Williams and O. A. Becklund, *Introduction to the Optical Transfer Function*, Wiley, New York, 1989.

[44] W. A. Woyczyński, *Lévy processes in the physical sciences*, in Lévy Processes—Theory and Applications, O. Barndorff-Nielsen, T. Mikosch, and S. Resnick, eds., Birkhäuser Boston, Cambridge, MA, 2001.

SIAM J. APPL. MATH.
Vol. 63, No. 2, pp. 619–635

PULSE PROPAGATION IN DISCRETE SYSTEMS OF COUPLED EXCITABLE CELLS*

A. CARPIO† AND L. L. BONILLA‡

Abstract. Propagation of pulses in myelinated fibers may be described by appropriate solutions of spatially discrete FitzHugh–Nagumo systems. In these systems, propagation failure may occur if either the coupling between nodes is not strong enough or the recovery is too fast. We give an asymptotic construction of pulses for spatially discrete FitzHugh–Nagumo systems, which agrees well with numerical simulations, and discuss the evolution of initial data into pulses and pulse generation at a boundary. Formulas for the speed and length of pulses are also obtained.

Key words. discrete reaction-diffusion equations, traveling wave pulses, propagation failure, spatially discrete FitzHugh–Nagumo system

AMS subject classifications. 34E15, 92C30

PII. S0036139901391732

1. Introduction.

Effects of spatial discreteness are important in many physical and biological systems comprising interacting smaller components such as atoms, quantum wells, cells, etc. Examples include the motion of dislocations [13], crystal growth and interface motion in crystalline materials [6], the motion of domain walls in semiconductor superlattices [4, 8], sliding of charge density waves [14], and pulse propagation through myelinated nerves [2]. The mathematical study of spatially discrete models is challenging because of special and poorly understood phenomena occurring in them that are absent if the continuum limit of these models is taken. Paramount among these phenomena is the pinning or propagation failure of wave fronts in spatially discrete equations. Physically, the pinning of wave fronts is related to the existence of Peierls stresses in continuum mechanics [17], relocation of electric field domains [1] and self-sustained oscillations of the current in semiconductor superlattices [18, 4], electric current due to the sliding of charge density waves [14], saltatory propagation of impulses in myelinated fibers and its failure [2], etc.

Mathematical understanding of the propagation failure of wave fronts in spatially discrete equations experienced significant progress after a paper by Keener [20]. In [20], Keener used comparison principles to characterize the pinning of wave fronts and their motion for spatially discrete reaction-diffusion equations of the form

$$u_{n,t} = d(u_{n+1} - 2u_n + u_{n-1}) + f(u_n), \tag{1.1}$$

where f is a bistable source term and d measures the strength of the coupling. Models described by (1.1) include the spatially discrete Nagumo equation for nerve conduction

*Received by the editors November 26, 2001; accepted for publication (in revised form) May 30, 2002; published electronically December 11, 2002. This research was supported by the DGES grant PB98-0142-C04-01, by the MCyT grant BFM2002-04127-C02, by the Third Regional Research Program of the Autonomous Region of Madrid (Strategic Groups Action), and by the European Union under grant RTN2-2001-00349.

http://www.siam.org/journals/siap/63-2/39173.html

†Departamento de Matemática Aplicada, Universidad Complutense de Madrid, 28040 Madrid, Spain (carpio@mat.ucm.es).

‡Departamento de Matemáticas, Escuela Politécnica Superior, Universidad Carlos III de Madrid, Avenida de la Universidad 30, 28911 Leganés, Spain, and Unidad Asociada al Instituto de Ciencia de Materiales de Madrid (CSIC), 28049 Cantoblanco, Spain, (bonilla@ing.uc3m.es).

[12, 26] or crystal growth [6], and the Frenkel–Kontorova model for motion of dislocations [13]. More recently, a number of results on the existence of traveling wave fronts $u_n(t) = v(n - ct)$ with smooth profiles v have been established [31, 24, 10]. These papers do not precisely characterize the propagation failure of a wave front for critical values of the control parameters. For a piecewise linear source function, an explicit description has been given by Fáth, extensively using the properties of special functions [11]. In the general case, we have found that propagation failure can be explained as a loss of continuity of the wave front profile in critical values of the control parameter [7]. Furthermore, the smoothness of the wave front profile just before the propagation failure occurs can be exploited to obtain an analytical description of wave fronts and their speed near critical parameter values [7, 9]. This theory has been extended to spatially discrete reaction-diffusion-convection equations describing dynamics of domain walls in semiconductor superlattices [8].

These advances in the mathematical understanding of propagation phenomena have occurred for spatially discrete scalar reaction-diffusion equations. Similar phenomena occur in models of calcium release at discrete sites [22]. The latter consist of scalar reaction-diffusion equations with spatially inhomogeneous source terms that are close to $f(u)$ times a series of delta functions centered at spatially periodic sites. Comparatively little progress has been made in understanding wave propagation and failure in spatially discrete systems. Anderson and Sleeman [2] have extended Keener's techniques to discrete reaction-diffusion systems modelled by the FitzHugh–Nagumo (FHN) dynamics [12, 26]. Hastings and Chen [16] have proved the existence of pulse traveling waves for a myelinated nerve model with a Morris–Lecar type of dynamics. They also comment on the difficulties of extending their results to the FHN system. An attempt to understand the mechanisms of propagation failure in the FHN system has been carried out by Booth and Erneux [5]. They consider slow recovery and very special limiting (small) values of the parameters characterizing the bistable source and the spatial diffusivity in the FHN system. Furthermore, they also impose particular boundary and initial conditions. With these restrictions, they could study how a specific disturbance localized in one cell propagated to neighboring ones until the resulting front failed to propagate. No construction of pulses or formulas for their velocity were given.

In this paper, we asymptotically construct pulse solutions of the spatially discrete FHN system describing nerve conduction through myelinated fibers. We also discuss how the pulses may fail to propagate. Our ideas could be extended to spatially discrete systems whose cell dynamics contain widely separated time scales corresponding to fast excitation and slow recovery variables. Among these systems, let us cite models for bursting behavior in pancreatic β cells [30] or the much more difficult case of front propagation in voltage-biased semiconductor superlattices [4]. In the latter, a separation of time scales exists, but it is not obviously included as a small parameter in the equations. In our presentation, we have chosen the FHN dynamics for its simplicity. This model has been widely used to understand issues that are obscured by technical complications in more realistic models of nerve conduction. We consider the following system of dimensionless equations:

$$\epsilon \frac{du_n}{dt} = d\,(u_{n+1} - 2u_n + u_{n-1}) + Au_n(2 - u_n)(u_n - a) - v_n, \tag{1.2}$$

$$\frac{dv_n}{dt} = u_n - Bv_n, \tag{1.3}$$

$n = 0, \pm 1, \ldots$. Here u_n and v_n are the membrane potential and the recovery variable

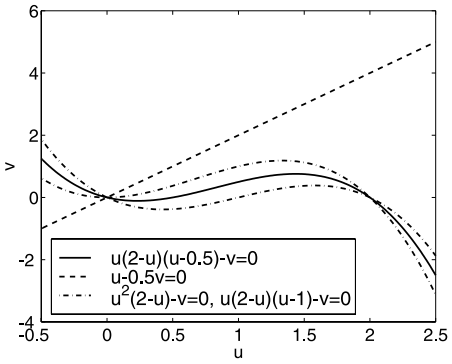


FIG. 1.1. *Nullclines for the space-independent FHN model with different values of a.*

(which acts as an outward ion current) at the nth excitable membrane site (node of Ranvier). The cubic source term is an ionic current, and the discrete diffusive term is proportional to the difference in internodal currents through a given site. The constants A and B are selected so that the source terms in the FHN system are $O(1)$ for u_n and v_n of order 1, that the only stationary uniform solution is $u_n = 0 = v_n$, and that the FHN system has excitable dynamics ($A = 1$, $B = 0.5$ is a good choice; see Figure 1.1). The constant $\epsilon > 0$ is the ratio between the characteristic time scales of both variables. We assume $\epsilon \ll 1$, that is, fast excitation and slow recovery. A dimensional version of (1.2) and (1.3) was derived in the appendix of [3] from an equivalent-circuit model of myelinated nerves. For background on similar models, see [28, 29, 21, 25].

We shall study pulse propagation in the spatially discrete FHN system (1.2) and (1.3) by asymptotic methods. At first sight, such a task is hopeless: asymptotic methods require a degree of smoothness at appropriate time or length scales, and the spatial variable n in these systems is discrete. However, we can use the separation between time scales in the FHN system to show that a pulse is made out of two "sharp" wave fronts separating regions of slow spatial variation. Wave fronts are *smooth* solutions of the *continuous* variable $z = n - ct/\epsilon$, and perturbative arguments apply straightforwardly to them. Thus the theory of wave front propagation for spatially discrete scalar reaction-diffusion equations plays an important role in our construction of pulses.

Let $U_1 < U_2 < U_3$ denote the three zeros of the cubic nonlinearity $f(u)$ in (1.1). U_1 and U_3 are stable solutions for $d = 0$. A wave front is a solution of (1.1) with a smooth profile $u_n(t) = u(n - ct)$ moving at a speed c such that $u(\mp\infty) = U_1$ and $u(\pm\infty) = U_3$. If $f(u)$ is odd about $u = U_2$ and d is sufficiently small, a stationary solution of (1.1) exists and therefore no wave fronts can propagate (see [10]). As the source term departs from this symmetric form, front propagation is made easier. In [7], we selected $d = 1$ and $f(u) = F - Ag(u)$, where $g(u)$ is odd about its middle zero and F is an external force that quantifies departure from symmetry. Notice that we can obtain (1.1) with $d = 1/A$ and $f = (F/A) - g(u)$ after rescaling time. We found that wave fronts propagate for $|F| > F_c$, where $F_c > 0$ depends on A and the specific $g(u)$ that we adopt. Equivalently, we could set $f(u) = -u(u-a)(u-2)$ and use $a - 1$ as a control parameter instead of F. After the "external force" a surpasses a critical value sufficiently far from the symmetry point $a = 1$, stationary fronts may cease to exist and propagating wave fronts may appear. See Figure 1.2(a). Notice that the limiting

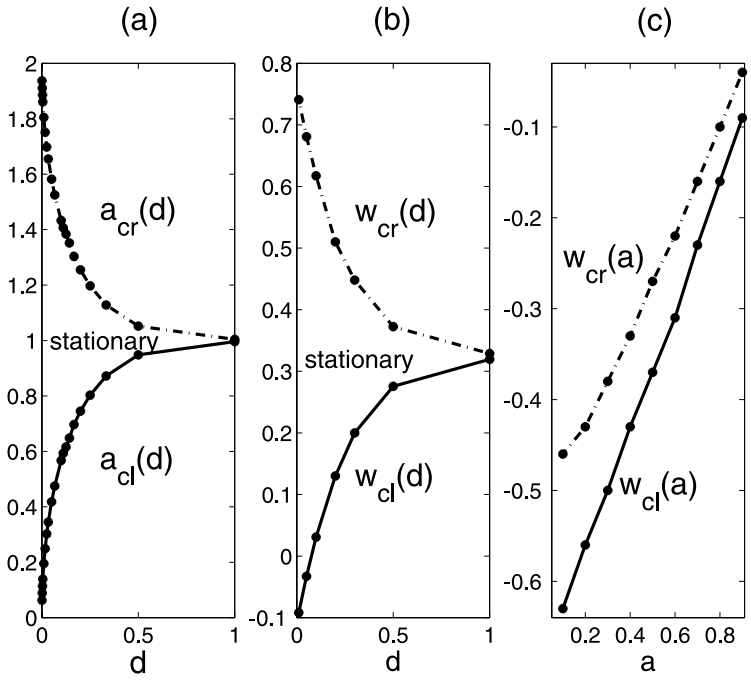


FIG. 1.2. (a) *Critical values* a_{cl} *and* a_{cr} *as functions of* d. (b) *Critical* $w_{cl}(a,d)$ *and* $w_{cr}(a,d)$ *for* $a = 0.5$. (c) *Critical* $w_{cl}(a,d)$ *and* $w_{cr}(a,d)$ *for* $d = 0.5$.

case considered by Booth and Erneux ("slow capture near a limit point"), $d = O(a^2)$, $a \to 0+$, corresponds to the parameter region in the lowest corner in this figure. In this region, propagation failure is the normal situation. We could, alternatively, fix a and set $f(u) = -w - u(u-a)(u-2)$, using w as the control parameter. Then there are critical values $w_{cl}(a,d)$ and $w_{cr}(a,d)$ such that wave fronts fail to propagate if $w_{cl} < w < w_{cr}$; see Figure 1.2(b). How does the parameter a affect the critical values of w? Assume that $a < a_{cl}(d)$ for a fixed value of d, so that wave fronts propagate for $w = 0$. To compensate for this effect, we need a small critical value $w_{cl}(a,d)$ of the parameter w. As a departs more and more from $a_{cl}(d)$, larger and larger critical values $w_{cl}(a,d)$ are needed to return to the situation of propagation failure. A similar situation occurs with $w_{cr}(a,d)$. Thus the critical values of w increase (in absolute value) as $|a - a_c|$ increases; see Figure 1.2(c). The effect of increasing the diffusivity d is to shrink the parameter range in which stationary fronts exist. In fact, as $d \to \infty$ (the continuum limit), the width of the pinning interval is conjectured to decrease exponentially quickly to zero for certain nonlinearities [6, 23]. Propagation failure can be understood as a loss of continuity of the moving front as appropriate critical parameter values are approached. Increasing the discrete diffusivity and deforming the source term sufficiently far from odd symmetry about its middle zero both facilitate the propagation of wave fronts [7, 9]. Reciprocally, weakening the coupling between cells and diminishing the "external force" $a - 1$ helps induce propagation failure.

For the spatially discrete FHN system, the description of wave propagation is more complicated. This also happens for the spatially continuous FHN system ($D\,\partial^2 u/\partial x^2$ instead of discrete diffusion). Depending on the initial condition, stable wave trains or pulses may be approached as time elapses [27, 15, 19]. Pulses cannot be obtained

for ϵ larger than a critical value. For discrete diffusion, we can construct pulses, provided that ϵ is smaller than a critical value $\epsilon_c(a, d)$, a is outside a certain interval (corresponding to propagation failure in the scalar case), and the initial condition is chosen appropriately. Our construction combines the theory of front depinning developed in [7] with Keener's asymptotic ideas [19] developed for the FHN model with spatially continuous diffusion. Our results agree very well with direct numerical solutions of (1.2) and (1.3).

The key ideas of an asymptotic theory for (1.2) and (1.3) in the limit as $\epsilon \to 0$ are simple. First, pulses consist of regions in which $u_n(t)$ vary smoothly with n, separated by moving sharp interfaces (fronts). In the first type of region, we may set $\epsilon = d = 0$ and obtain a description of slow recovery. The sharp interfaces are wave fronts with *smooth profiles*, $u_n(t) = u(z)$, $v_n(t) = v(z)$, with $z = n - ct/\epsilon$. Then v is constant at each side of a front, and the excitation variable u obeys the spatially discrete Nagumo equation, whose fronts we can characterize [7, 9]. A stable pulse is obtained when the velocity of the leading front is equal to that of the trailing front [19]. This condition fixes the pulse width. Its violation or propagation failure of any of the fronts bounding the pulse result in propagation failure thereof. Notice that we use an analytic expression for the wave front velocity of the Nagumo equation, valid as a is near its critical value for propagation failure. For small values of d, the front propagation range is narrow, and the formula for wave front velocity holds for all appropriate values of a; see Figure 1.2(a). For larger values of d, the interval where propagation occurs is wide, and we can use our approximation only for a close to its critical values a_{cl} and a_{cr}. Outside these parameter ranges, the velocity of the Nagumo wave fronts should be calculated numerically.

The rest of the paper is organized as follows. In section 2, we recall certain needed results on wave front propagation and failure for the spatially discrete scalar reaction-diffusion (Nagumo) equation. Section 3 contains the main theoretical ideas of this paper, with the asymptotic construction of pulses for the discrete FHN system. These ideas and our results are tested by numerically solving the FHN system with appropriate boundary conditions in section 4. Comments on propagation failure of pulses in the FHN system are made in section 5. Section 6 briefly discusses how a pulse may be generated by applying a temporary stimulus at one end of a fiber with finitely many nodes. The last section contains our conclusions.

2. The spatially discrete Nagumo equation. We consider the equation

$$\frac{du_n}{ds} = d\,(u_{n+1} - 2u_n + u_{n-1}) + u_n(2 - u_n)(u_n - a) - w \tag{2.1}$$

for some constant w and denote $h(u, w, a) = u(2 - u)(u - a) - w$. As long as $\min h(u, 0, a) < w < \max h(u, a, w)$, this is a "cubic" source having three zeroes $U_i(w, a)$, $i = 1, 2, 3$, $U_1 < U_2 < U_3$. Wave front solutions joining U_1 and U_3 (the two stable zeros) exist. A theory of the pinning and propagation of fronts for this type of equation has been developed in [7, 8]. We sketch its implications for (2.1) below.

First assume $w = 0$, so that the asymmetry of the source is controlled by the parameter a. For d fixed, there are values $a_{cl}(d)$ and $a_{cr}(d)$ such that the following hold:

- The fronts joining $u = 0$ and $u = 2$ are stationary if $a_{cl}(d) \le a \le a_{cr}(d)$. No front propagation is possible.
- Outside this interval, there exist traveling wave fronts $u_n(s) = u(n - cs)$ joining 0 and 2. For $a > a_{cr}(d)$, increasing fronts move to the right and de-

creasing fronts move to the left. For $a < a_{cl}(d)$, fronts move in the opposite way: decreasing fronts move to the right and increasing fronts move to the left. The values $a_{cl}(d)$ and $a_{cr}(d)$ can be approximately calculated as follows. In a large lattice, we decrease or increase a from 1 until we obtain a stationary solution $u_n(a)$ whose linear stability problem has a zero eigenvalue; see Figure 1.2.

Now we fix a and vary w. The asymmetry of the source is controlled by a and w. For fixed d and a, critical values $w_{cl}(a,d)$ and $w_{cr}(a,d)$ are found such that the following hold:

- The fronts joining $U_1(w,a)$ and $U_3(w,a)$ are stationary if $w_{cl}(a,d) \le w \le w_{cr}(a,d)$.
- Outside this interval, there exist traveling wave fronts $u_n(s) = u(n - cs)$ joining $U_1(w,a)$ and $U_3(w,a)$. For $w < w_{cl}$, these fronts move to the left if they increase from U_1 to U_3, and to the right if they decrease from U_3 to U_1. For $w > w_{cr}$, fronts decreasing from U_3 to U_1 move to the left, and increasing fronts move to the right.

To calculate w_{cl} and w_{cr}, we start by fixing a and finding a value $w = w_0$ at which stationary solutions exist for a large lattice. We now decrease or increase w from this value until we obtain a stationary solution $u_n(w)$ whose linear stability problem has a zero eigenvalue; see Figure 1.2.

For w near any of its critical values, we can use the following formula to predict the speed of the fronts for $|w| > |w_c|$:

$$c(a,d,w) \sim \operatorname{sign}(w - w_c)\frac{\sqrt{\alpha\beta(w - w_c)}}{\pi}. \tag{2.2}$$

The parameters α and β, given by $\alpha = \sum \phi_n$, $\beta = \frac{1}{2}\sum[-6u_n(w_c) + 2(2+a)]\phi_n^3$ (see [7, 9]), are functions of a, d, and the critical value of w. In these formulas, ϕ is a positive eigenfunction of the linear stability problem for $u_n(w_c)$ with $\sum \phi_n^2 = 1$, and $u_n(w_c)$ is a stationary solution of (2.1) with $w = w_c$ [9]. If w is not close to its critical values, the speed $c(a,d,w)$ should be calculated numerically.

A peculiarity of the Nagumo equation is the scenario for front propagation failure. As we approach the critical values for a, w, or any other appropriate pinning control parameter, the front profiles become less smooth and a number of steps appear. In the limit as the control parameter tends to its critical value, the transition regions between steps become infinitely steep, the front profile becomes discontinuous, and its velocity vanishes [7, 9].

3. Asymptotic construction of pulses. As we will discuss below, an appropriate initial condition evolves towards a pulse. In particular, we need to fix the parameters $d > 0$, $a < a_{cl}(d)$ (the case $a > a_{cr}(d)$ follows by symmetry), and ϵ smaller than a certain critical value, $\epsilon_c(a,d)$. This last condition also holds for the spatially continuous FHN system, which has two pulse solutions (one stable and one unstable) for $\epsilon < \epsilon_c$. These solutions coalesce at ϵ_c and cease to exist for larger ϵ (see [26, 27]). A pulse consists of regions of smooth variation of u on the time scale t, separated by sharp interfaces in which u varies rapidly on the time scale $T = t/\epsilon$. In the regions where u varies smoothly, we can set $\epsilon = d = 0$, thereby obtaining the reduced problem

$$u_n(2 - u_n)(a - u_n) - v_n = 0, \tag{3.1}$$

$$\frac{dv_n}{dt} = u_n - B\,v_n. \tag{3.2}$$

These regions are separated by sharp interfaces (moving fronts), at which u_n varies rapidly as $u_n(t) = u(z)$, $v_n(t) = v(z)$, with $z = n - ct/\epsilon$. There, to leading order,

$$-c\frac{du}{dz} = d\left[u(z+1) - 2u(z) + u(z-1)\right] + u(z)\left(2 - u(z)\right)\left(a - u(z)\right) - v, \tag{3.3}$$

$$-c\frac{dv}{dz} = 0. \tag{3.4}$$

Thus v is a constant equal to the value $v_n(t)$ at the last point in the region of smooth variation before the front. Equation (3.3) has a wave front solution as discussed in the previous section. We can now discuss different regions in the asymptotic description of a pulse as follows:

1. The region of smooth variation of u in front of the pulse, described by (3.1) and (3.2). In this region, $u_n = U_1(v_n)$, so that
$$\frac{dv_n}{dt} = U_1(v_n) - B\,v_n,$$
and initial data evolve exponentially fast towards equilibrium, $u_n = v_n = 0$.
2. The pulse leading edge. Let $v(t)$ be the value of v_n at the last point of the region in front of the pulse. Eventually, $v \to 0$. At the leading edge, $u_n(t) = u(n - ct/\epsilon)$ is a wave front moving towards the right with speed $C = c(a, d, v)/\epsilon$ measured in points per unit time t. We have the boundary conditions $u(-\infty) = U_3(v)$ and $u(\infty) = U_1(v)$ for the monotone decreasing profile $u(z)$ which satisfies (3.3). It is convenient to call $c_-(v) = c(a, d, v)$. Eventually, $C \sim c_-(0)/\epsilon$, and u_n decreases from $u_n = 2$ to $u_n = 0$ across the leading edge of the pulse.
3. The region between fronts: $u_n = U_3(v_n)$ and
$$\frac{dv_n}{dt} = U_3(v_n) - B\,v_n.$$
There is a finite number of points in this region. On its far right, $v_n = v \to 0$. As we move towards the left, v_n increases until it reaches a certain value $V(t)$ corresponding to that in the trailing wave front.
4. The trailing wave front: $v_n(t) = v(z) = V$, and $u_n(t) = u(z)$ obeys (3.3) with boundary conditions $u(-\infty) = U_1(V)$ and $u(\infty) = U_3(V)$. This front increases monotonically with z, and it moves with speed $C = c(a, d, V)/\epsilon$ measured in points per unit time t. It is convenient to denote $c_+(V) = c(a, d, V)$. We shall indicate how to determine V below. Clearly, if the pulse is to move rigidly, we should have $c_+(V) = c_-(0)$ after a sufficiently long transient period.
5. Pulse tail. Again $u_n = U_1(v_n)$ and $dv_n/dt = U_1(v_n) - Bv_n$. Sufficiently far to the left, $v_n = u_n = 0$.

The number of points between wave fronts of the pulse is not arbitrary: it can be calculated following an argument due to Keener for the spatially continuous case [19]. Let τ be the delay between fronts, i.e., the time elapsed from the instant at which the leading front traverses the point $n = N$ to the instant when the trailing front is at $n = N$. Clearly,

$$\tau = \int_{v(t-\tau)}^{V(t)} \frac{dv}{U_3(v) - Bv}\,. \tag{3.5}$$

The number of points between fronts, $l(t)$, can be calculated as

$$l = \frac{1}{\epsilon} \int_{t-\tau}^{t} c_-(v(t))\, dt. \tag{3.6}$$

On the other hand, the separation between fronts satisfies the equation

$$\frac{dl}{dt} = \frac{c_-(v(t)) - c_+(V(t))}{\epsilon}\,. \tag{3.7}$$

The three equations (3.5), (3.6), and (3.7) can be solved to obtain the three unknowns τ, l, and $V(t)$. (The function $v(t)$ is determined by solving (3.2) with $u_n = U_1(v_n)$ in the region to the left of the leading front.)

After a transient period, $v(t) \to 0$ and $V(t) \to V$ (a constant value), so that we have the simpler expressions

$$\tau = \int_0^V \frac{dv}{U_3(v) - Bv}\,, \tag{3.8}$$

$$\frac{dl}{dt} = \frac{c_-(0) - c_+(V)}{\epsilon}\,, \tag{3.9}$$

instead of (3.5) and (3.7), respectively. The number of points at the pulse top is now

$$l = \frac{c_-(0)\tau}{\epsilon} = \frac{c_-(0)}{\epsilon} \int_0^V \frac{dv}{U_3(v) - Bv}\,. \tag{3.10}$$

This equation yields V as a function of l. Then (3.9) becomes an autonomous differential equation for l that has a stable constant solution at $l = l^*$ such that $c_-(0) = c_+(V(l))$: at $l = l^*$, the right-hand side of (3.9) has a slope $-[U_3(V) - BV]\, c_+'(V)/c_-(0) < 0$.

Recapitulating, for appropriate initial conditions, the leading and trailing fronts of a pulse evolve until l reaches its stable value at which $c_-(0) = c_+(V(l^*))$ and (3.10) holds. To compute l^*, we first determine $V^* = V(l^*)$ by using $c_-(0) = c_+(V(l^*))$. Then we calculate $\tau = \tau^*$ (which does not depend on ϵ!) from (3.8) and $l^* = c_-(0)\tau^*/\epsilon$. Our construction breaks down if the number of points between fronts falls below 1. This yields an upper bound for the critical value of ϵ above which pulse propagation fails: $\epsilon_c \sim c_-(0)\tau^*$.

The asymptotic length of the pulse tail is obtained by first calculating the time needed for v_n to go from 0 to $V(l^*)$ to the left of the trailing front: $T = \int_0^V dv/[U_1(v) - v]$. The tail length is then $L = c_-(0)T/\epsilon$.

4. Numerically calculated pulses. We shall compare numerical solutions for different representative values of d with the approximate pulses provided by our theory. As initial data, we have adopted our approximate pulses. We have also used hump-like profiles with compact support for $u_n(0)$ and $v_n(0)$. It is important that $v_n(0) = 0$ at the leading edge of $u_n(0)$ and to its right, and that $v_n(0) = V \approx w_{cr}(d)$ at the trailing edge, where $v_n(0)$ reaches its maximum. Had we chosen $v_n(0) = 0$ for all n, the u_n profile would have split into two pulses traveling in opposite directions as time elapsed; see Figure 4.1. The region between leading and trailing fronts of the pulse acquires its asymptotic shape quite quickly, but the pulse tail is usually rather long and evolves slowly towards its final form.

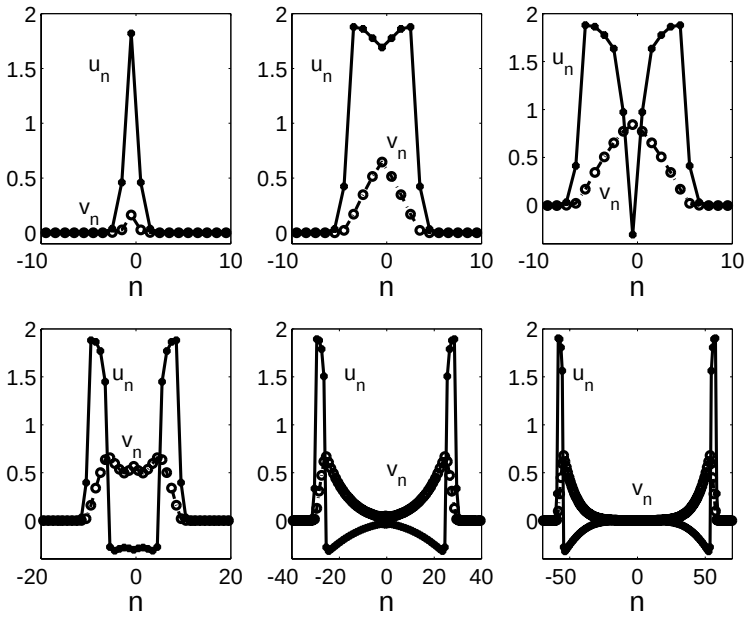


FIG. 4.1. *Splitting of an initial profile into two pulses propagating in opposite directions for* $d = 0.1$, $a = 0.5$, *and* $\epsilon = 0.006$.

- $d = 1$. In this case, $a_{cl} = 0.996$. We choose $a = 0.99 < a_{cl}$ and analyze front propagation for the rescaled Nagumo equation (2.1) first. The front propagation thresholds for these values of d and a are $w_{cl} = 0.0038$ and $w_{cr} = 0.0095$. Figure 4.2(a) shows the speeds of leading and trailing fronts as functions of w, as predicted by (2.2). For $w = 0$, the leading front should move at speed $c_-(0) = 0.0093$. The relation $c_+(V) = c_-(0)$ yields the asymptotic value $V^* = 0.0133$ at the trailing front joining $U_1(V^*) = -0.00665$ to $U_3(V^*) = 1.9933$. The time elapsed between fronts is $\tau^* = 0.00652$, as calculated from (3.8). Then our upper bound for the critical value of ϵ is $\epsilon_c = 0.000064$. Choosing a smaller value, $\epsilon = 0.000005$, we obtain a pulse speed of $C = c_-(0)/\epsilon = 1869$ points per unit time and a pulse width of $l^* = C\tau^* \sim 13$ points. Our numerical solution of the full FHN system (1.2) and (1.3) yields a pulse speed $C = 2000$ and a width of 13 points for $\epsilon = 0.000005$. The trailing front joins -0.006647 and 1.993 with $V^* = -0.0133$; see Figure 4.2(b). Note that the relative error in the predicted speed C is 0.0655. Obviously, rescaling the speed to $C = c_-(0)/\epsilon$ amplifies the error in our predictions. We have not been able to observe pulses for $\epsilon \geq 0.0000076$, which is smaller but not far from our estimation $\epsilon_c = 0.000064$. Let us now choose $a = 0.5$ which is far from a_{cl}. Then $w_{cl} = 0.3194$ and $w_{cr} = 0.3287$. Equation (2.2) predicts $c_-(0) = 0.09983$, whereas the trailing front joins -0.316 to 1.71 at $V^* = 0.6$. If $\epsilon = 0.01$, the speed and width of the pulse are $C = 9.983$ and $l^* = 0.351C \sim 4$, according to our theory. Numerically, we observe $C = 64.7$ and $l^* = 25$. The source of these large errors is the value $c_-(0) = 0.09983$ predicted with formula (2.2). If we replace this value by the numerical front speed calculated directly, $c_-(0) = 0.673$, we obtain $C = 67.3$ and $l^* \sim 24$ points, which better

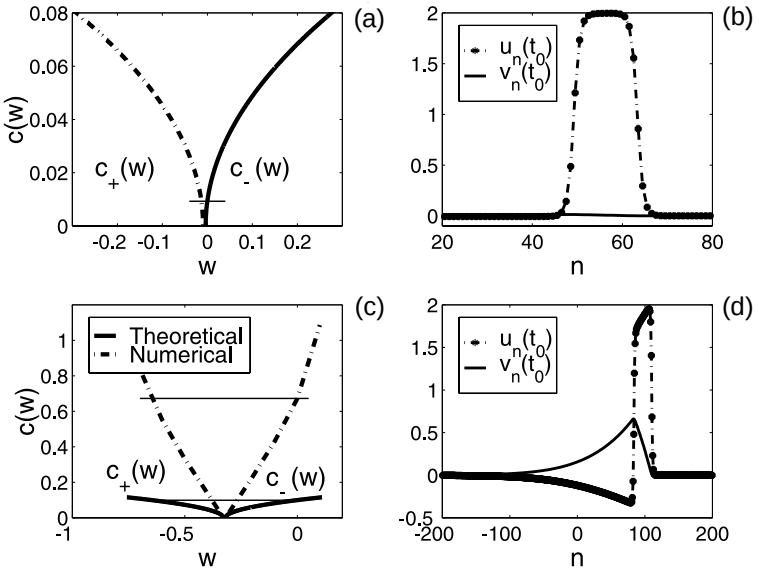


FIG. 4.2. (a) *Predicted speeds for wave fronts of the Nagumo equation* (2.1) *with* $d = 1$, $a = 0.99$. *The horizontal line marks the condition* $c_+(w) = c_-(0)$, *thereby graphically yielding* $w = V^*$. (b) *FHN pulse for* $\epsilon = 0.000005$. (c) *Predicted and numerical speeds for wave fronts of* (2.1) *with* $d = 1$, $a = 0.5$. *The horizontal lines mark* $c_+(w) = c_-(0)$. (d) *FHN pulse for* $\epsilon = 0.01$.

fit the numerically observed values.

- $d = 0.1$. In this case, $a_{cl} = 0.567$, and we shall choose $a = 0.5 < a_{cl}$. Let us first analyze front propagation for the rescaled Nagumo equation (2.1). For these values of d and a, we obtain $w_{cl} = 0.0307$ and $w_{cr} = 0.6175$. Figure 4.3(a) shows the predicted speeds of the leading and trailing fronts as functions of v, as given by formula (2.2). For $w = 0$, the leading front should move with speed $c_-(0) = 0.075662$. At the trailing front, $c_+(V) = c_-(0)$ yields $V^* = 0.648$, $U_1(V^*) = -0.33328$, and $U_3(V^*) = 1.666$. The time elapsed between fronts is $\tau^* = 0.39266$, which gives $l^* = 0.0297/\epsilon$. Our bound for the critical value of ϵ is $\epsilon_c = c_-(0)\tau^* = 0.029$. Selecting $\epsilon = 0.003$, we predict $C = 25.22$ and $l^* \sim 10$ points. Direct numerical calculations yield a pulse speed $C = 26.38$ and a pulse width of about 10 points. The trailing front joins -0.3269 to 1.675 with $V^* = 0.6578$; see Figure 4.3(b). We have not been able to obtain pulses for $\epsilon \geq 0.007$, which is four times smaller than our upper bound of 0.029.
- $d = 0.01$. In this case, $a_{cl} = 0.195$, and we shall choose $a = 0.1 < a_{cl}$. Let us first analyze front propagation for the rescaled Nagumo equation (2.1). The front propagation thresholds for these values of d and a are $w_{cl} = 0.0136$ and $w_{cr} = 1.0784$. Figure 4.4(a) shows the predicted speeds of leading and trailing fronts as functions of w according to (2.2). For $w = 0$, the leading front should move with speed $c_-(0) = 0.052$. Then the trailing front has $V^* = 1.092$ corresponding to $c_+(V) = c_-(0)$, and it joins $U_1(V^*) = -0.6$ to $U_3(V^*) = 1.4$. The time elapsed between fronts is $\tau^* = 0.748$, and the pulse width, $l^* = 0.297/\epsilon$. Our bound for the critical value of ϵ is $\epsilon_c = c_-(0)\tau^* = 0.058$. Selecting $\epsilon = 0.001$, we predict $C = 52$ and $l^* = 39$ points.

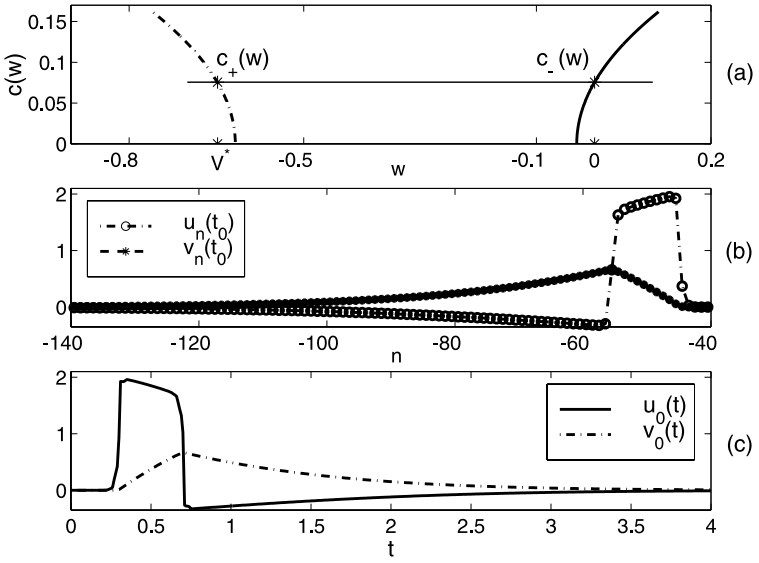


FIG. 4.3. (a) *Predicted speeds for the Nagumo equation* (2.1) *with* $d = 0.1$ *and* $a = 0.5$. *The horizontal line graphically yields* V^* *such that* $c_+(V^*) = c_-(0)$. (b) *Profiles of the FHN pulse for* $\epsilon = 0.003$ (c) *Trajectories of one point,* $u_0(t)$, $v_0(t)$, *as the FHN pulse propagates through it.*

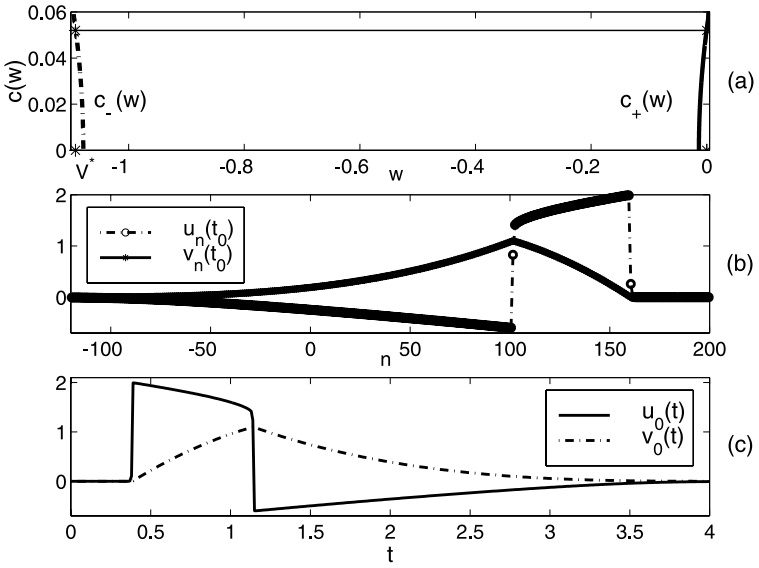


FIG. 4.4. (a) *Predicted speeds for the Nagumo equation* (2.1) *with* $d = 0.01$ *and* $a = 0.1$. (b) *FHN pulse for* $\epsilon = 0.001$. (c) *Trajectory of one point,* $u_0(t)$, *as the FHN pulse propagates through it.*

Numerical observations yield $C = 77.7$ (a relative error of 0.3) and a pulse width of 59 points. Furthermore, the trailing front joins -0.59 to 1.4 with $V^* = 1.095$; see Figure 4.4(b). Again the observed errors in the pulse speed and width are due to errors in the prediction of $c_-(0)$ given by formula (2.2). Replacing this value by the numerically computed front speed $c_+(0) = 0.078$, we obtain $C = 78$ and $l^* \sim 58$, better fit to the real values.

Let us now describe the situation for other values of d. Our asymptotic theory agrees with the numerical results, provided that ϵ is sufficiently small, but the velocity of the Nagumo wave fronts should be either approximated by (2.2) or calculated numerically depending on how close to zero w_{cl} happens to be. For $d < 0.01$, the length of the intervals in which fronts of the Nagumo equation propagate is very small. Then the front speeds are always very small and given by (2.2) with great accuracy. Our asymptotic description of the pulse agrees very well with numerical solutions of the FHN system. If $d > 1$, the spatially discrete FHN system can be approximated by its continuum limit. The length of the pinning intervals for the Nagumo equation is below 0.001, and the wave front velocities are essentially a correction of the wave front velocities for the spatially continuous Nagumo equation (see [21]):

$$c = \sqrt{d}\, c_0 \left(1 - \frac{k_1 c_0^2}{2d} + O\left(\frac{c_0^4}{d^2}\right)\right), \tag{4.1}$$

$$c_0 = \frac{2U_2(w) - U_1(w) - U_3(w)}{\sqrt{2}}, \tag{4.2}$$

$$k_1 = -\frac{\int_{-\infty}^{\infty} e^{-c_0^2 s} V_0'(s) V_0''''(s)\, ds}{12 c_0^4 \int_{-\infty}^{\infty} e^{-c_0^2 s} V_0'(s) V_0''(s)\, ds}. \tag{4.3}$$

Here V_0 is the appropriate wave front solution of the equation

$$c_0^{-2} V_0'' - V_0' + V_0 (2 - V_0)(V_0 - a) - w = 0. \tag{4.4}$$

5. Propagation failure. Two facts may lead to propagation failure: a value of ϵ that is too large or $a \in (a_{cl}(d), a_{cr}(d))$.

Let us consider the first cause of propagation failure now. If ϵ surpasses a certain critical value ϵ_c, recovery is too fast and a stable pulse cannot be sustained. This situation also occurs in spatially continuous FHN systems. In these systems, there exist two pulses (one pulse is stable, the other unstable) for $\epsilon < \epsilon_c$; they coalesce at ϵ_c and cease to exist for larger ϵ. In the discrete FHN system, the phenomenon of wave front propagation failure implies that pulses may propagate only if $a < a_{cl}(d)$ or $a > a_{cr}(d)$. As indicated by (3.10), the number of points between the two fronts of the stable pulse decreases as ϵ increases towards $\epsilon_c(a, d)$. Eventually the two fronts coalesce, and it is not possible to propagate a stable pulse for $\epsilon > \epsilon_c(a, d)$. If we start with an appropriate pulse-like initial condition, we find the scenario of propagation failure depicted in Figures 5.1 and 5.2. For small d ($d = 0.1$), the variable v_n ceases to be almost constant at the leading edge of the pulse, and the distance between the two fronts diminishes. While $v_n \sim 0$ at the rightmost point of the leading front, $v_n \sim w > 0$ at the leftmost point. Thus, u_n in this front decreases from $U_3(w)$ to zero as n increases. The value w increases with ϵ, and $U_3(w)$ decreases. At the same time, the leading front speed diminishes as w increases until w surpasses the propagation threshold and the leading front stops. Since the back front goes on moving, the pulse vanishes; see Figure 5.1. For large d ($d = 1$), a decremental pulse is formed. Its

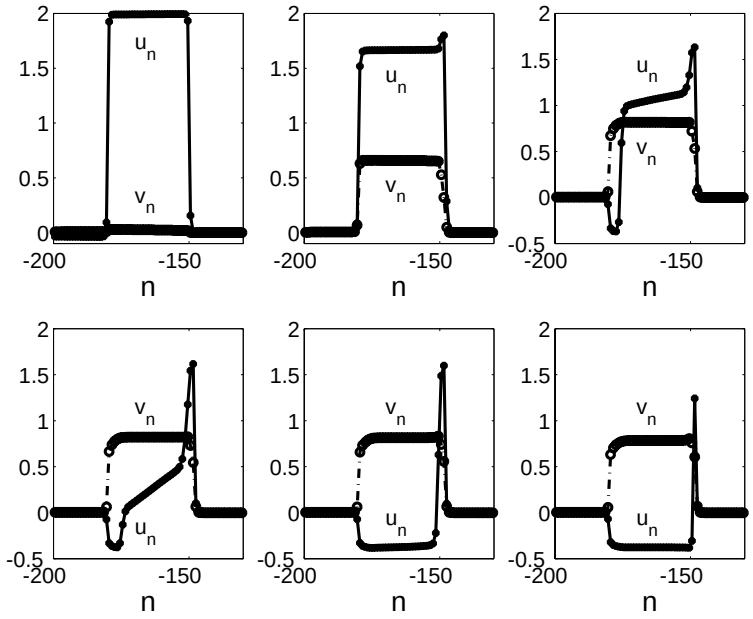


FIG. 5.1. *Snapshots of the excitation and recovery variables for* $d = 0.1$, $a = 0.5$, *and* $\epsilon = 0.007$, *illustrating propagation failure of the pulse for* $\epsilon > \epsilon_c$.

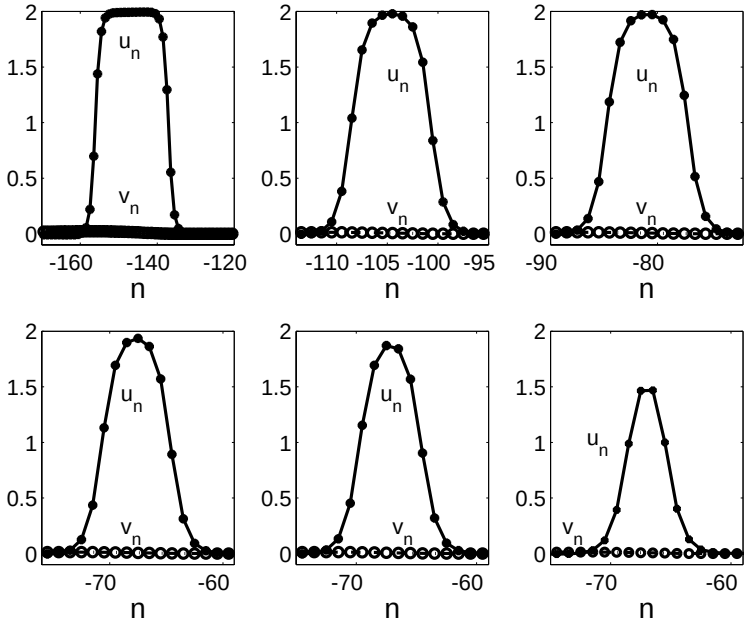


FIG. 5.2. *Same as Figure* 5.1 *for* $d = 1$, $a = 0.99$, *and* $\epsilon = 0.0000076$.

width and height decrease as it moves until it disappears; see Figure 5.2. Numerical simulations of the FHN system show that $\epsilon_c \to 0$ as a tends to either $a_{cl}(d)$ or $a_{cr}(d)$.

Let us assume now that $a \in (a_{cl}(d), a_{cr}(d))$. Then the leading front cannot propagate with $v_n = v = 0$. We need $v_n \sim v < w_{cl}(a,d) < 0$. However, in the region in front of the leading edge, v_n and u_n evolve towards 0, whereas we have $u_n > 0$ at the leading front. Thus $dv_n/dt = u_n - Bv_n \geq 0$ there, and v_n will increase until $v_n > 0$, which contradicts our previous assumption. Thus we cannot have stable propagating pulses. Furthermore, there are no stationary pulses of the type we have discussed for this range of a: if $v_n = u_n/B$, the source $u_n(2-u_n)(a-u_n) - v_n = u_n(2-u_n)(u_n-a) - u_n/B$ has only one zero, not three as in our construction. This does not preclude the existence of other pulses, such as those corresponding to the homoclinic orbit in the phase space of the spatially continuous FHN system. However, we have not observed stable stationary pulses of this type in the spatially discrete FHN system.

6. Pulse generation at a boundary. So far, we have considered the motion of a pulse (or its failure) in a sufficiently large myelinated nerve fiber. We have not discussed how such a pulse might be created in a more realistic situation. Clearly, nerve fibers have finitely many nodes of Ranvier, and pulses are typically generated at the fiber boundary. Thus we are led to consider how a pulse might be generated by an excitation at a boundary and how the pulse propagates or fails in a finite fiber. This problem was tackled by Booth and Erneux [5] using parameter values for which the FHN pulse fails to propagate. We shall now discuss different parameter ranges.

Nerve fibers may have either a few nodes of Ranvier (e.g., 20 for neurons of the central nervous system [25]) or several hundred nodes (in the peripheral nervous system [29]). Thus we shall consider a finite FHN system with N nodes and a Neumann boundary condition at the right end, $u_{N+1} = u_N$. At the left end, we impose $u_0(t) = 2$ for $0 \leq t \leq 0.05$, and $u_0(t) = 0$ for $t > 0.05$. The results corresponding to parameter values $d = 0.1$ and $a = 0.5$ are depicted in Figures 6.1 (for which $\epsilon = 0.006$) and 6.2 (for which $\epsilon = 0.003$). The asymptotic theory predicts that fully developed FHN pulses (corresponding to $N = \infty$) would have widths of $l^* \approx 5$ and $l^* \approx 10$, respectively. The left boundary condition ensures that the membrane potential u_n is excited during sufficient time, so that a wave is generated at the left end of the fiber.

The excitation at the left boundary induces a wave front that propagates with a velocity given approximately by $C = c_-(0)/\epsilon$ along the finite fiber for the parameter values we consider. For example, $C \approx 12.6$ for $\epsilon = 0.006$, which is close to the numerically observed value of 10 in Figure 6.1. Similarly, $C \approx 25.22$ for $\epsilon = 0.003$, which is close to the numerically observed value of 26 in Figure 6.2. If the fiber is long enough, a second wave front follows the first one, and their mutual distance rapidly approaches the asymptotic value l^*. (The number of nodes between fronts is 4 in Figure 6.1, while the asymptotic theory predicts $l^* \approx 5$; in Figure 6.2, numerical observation confirms the asymptotic value $l^* \approx 10$.) The numerical solution of the finite FHN system shows that an eventually truncated FHN pulse comprising the two wave fronts and the region between them is formed, provided that N is at least twice l^*. Otherwise, at best only the first wave front is shed at the boundary, as shown in Figure 6.2(a). Pulses fail to propagate in fibers whose parameters fall in the propagation failure region, as discussed in section 5.

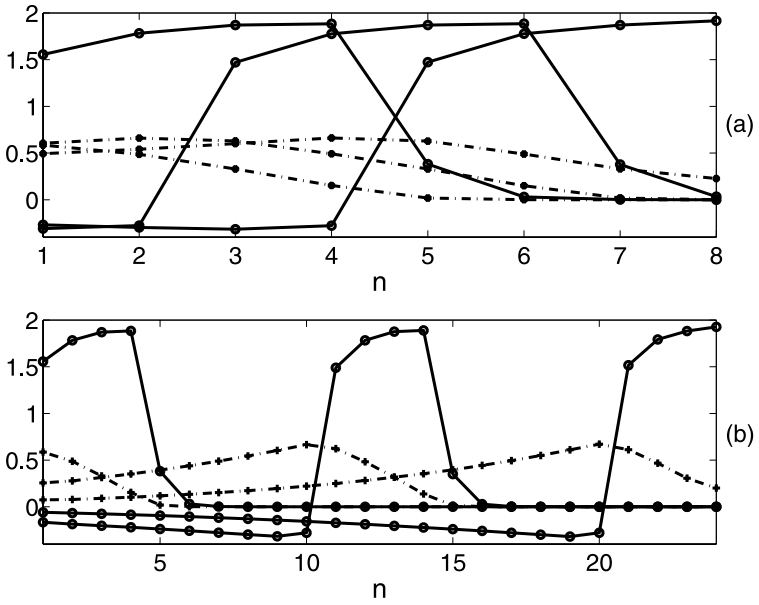


FIG. 6.1. *Snapshots of the excitation (solid line) and recovery (dotted line) variables for an FNH system with N nodes and* $d = 0.1$, $a = 0.5$, *and* $\epsilon = 0.006$. (a) *Profiles at times* 0.4, 0.6, *and* 0.8 *for* $N = 8$. (b) *Profiles at times* 0.4, 1.4, *and* 2.4 *for* $N = 24$.

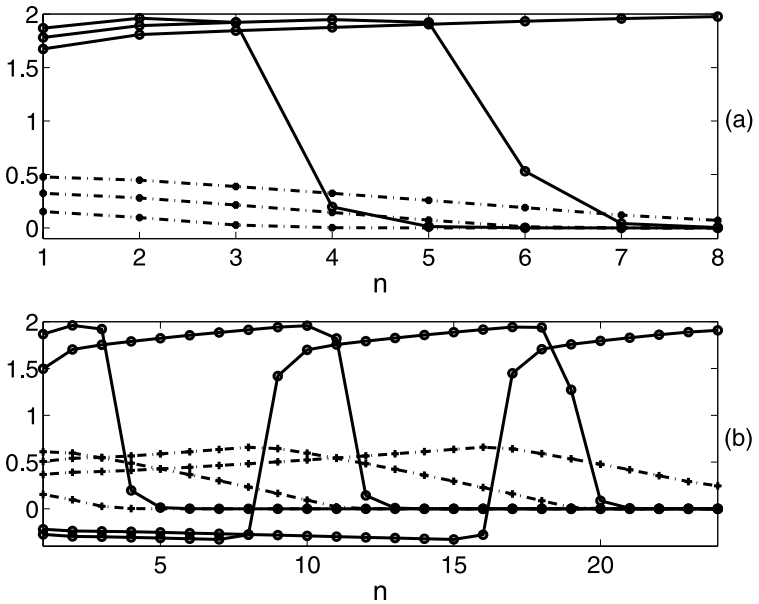


FIG. 6.2. *Snapshots of the excitation (solid line) and recovery (dotted line) variables for an FNH system with N nodes and* $d = 0.1$, $a = 0.5$, *and* $\epsilon = 0.003$. (a) *Profiles at times* 0.1, 0.2, *and* 0.3 *for* $N = 8$. (b) *Profiles at times* 0.1, 0.4, 0.7, *and* 1.0 *for* $N = 24$.

7. Conclusions. We have constructed stable pulses of the spatially discrete FHN system by asymptotic methods. In a pulse, there are regions where the excitation variable varies smoothly, separated by sharp fronts. These fronts are solutions of the discrete Nagumo equation with a constant value of the recovery variable. Their shape and speed can be approximately calculated near parameter values corresponding to front propagation failure or near the continuum limit. For long times, their width is given by the only stable solution of a one-dimensional autonomous system. We have compared the asymptotic results with numerical solutions of the FHN system and analyzed different scenarios for failure of pulse propagation. Besides the classical scenario of small separation between the time scales of excitation and recovery (large ϵ as in the spatially continuous FHN system), propagation failure of fronts for the spatially discrete Nagumo equation provides a different mechanism of propagation failure of pulses for the discrete FHN system. Wave fronts and pulses can be generated at a boundary and propagate or fail to propagate along a finite FHN system. If the number of nodes is sufficiently large, the two wave fronts comprising an FHN pulse can be shed at the boundary, and their separation rapidly reaches the value given by the asymptotic theory. This is true even if the fiber is too short to accommodate the slowly varying regions at the back of the second wave front of the pulse. In long fibers, a fully developed FNH pulse may be generated by an over-threshold stimulus applied during a short time at one end of the fiber.

REFERENCES

[1] A. Amann, A. Wacker, L. L. Bonilla, and E. Schöll, *Dynamic scenarios of multistable switching in semiconductor superlattices*, Phys. Rev. E (3), 63 (2001), pp. 1–8.

[2] A. R. A. Anderson and B. D. Sleeman, *Wave front propagation and its failure in coupled systems of discrete bistable cells modelled by FitzHugh-Nagumo dynamics*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 5 (1995), pp. 63–74.

[3] J. Bell and C. Costner, *Threshold behavior and propagation for nonlinear differential-difference systems motivated by modeling myelinated axons*, Quart. Appl. Math., 42 (1984), pp. 1–13.

[4] L. L. Bonilla, *Theory of nonlinear charge transport, wave propagation and self-oscillations in semiconductor superlattices*, J. Phys. Condensed Matter, 14 (2002), pp. R341–R381.

[5] V. Booth and T. Erneux, *Understanding propagation failure as a slow capture near a limit point*, SIAM J. Appl. Math., 55 (1995), pp. 1372–1389.

[6] J. W. Cahn, *Theory of crystal growth and interface motion in crystalline materials*, Acta Metallurgica, 8 (1960), pp. 554–562.

[7] A. Carpio and L. L. Bonilla, *Wave front depinning transitions in discrete one-dimensional reaction diffusion equations*, Phys. Rev. Lett., 86 (2001), pp. 6034–6037.

[8] A. Carpio, L. L. Bonilla, and G. Dell'Acqua, *Wave front motion in semiconductor superlattices*, Phys. Rev. E (3), 64 (2001), pp. 1–9.

[9] A. Carpio and L. L. Bonilla, *Depinning transitions in discrete reaction-diffusion equations*, SIAM J. Appl. Math., to appear.

[10] A. Carpio, S. J. Chapman, S. P. Hastings, and J. B. Mcleod, *Wave solutions for a discrete reaction-diffusion equation*, European J. Appl. Math., 11 (2000), pp. 399–412.

[11] G. Fáth, *Propagation failure of traveling waves in a discrete bistable medium*, Phys. D, 116 (1998), pp. 176–190.

[12] R. FitzHugh, *Impulse and physiological states in models of nerve membrane*, Biophys. J., 1 (1961), pp. 445–466.

[13] J. Frenkel and T. Kontorova, *On the theory of plastic deformation and twinning*, J. Phys. USSR, 13 (1938), pp. 1–10.

[14] G. Grüner, *The dynamics of charge-density waves*, Rev. Modern Phys., 60 (1988), pp. 1129–1181.

[15] S. P. Hastings, *The existence of homoclinic and periodic orbits for FitzHugh-Nagumo's equations*, Q. J. Math., 27 (1976), pp. 123–134.

[16] S. P. Hastings and X. Chen, *Pulse waves for a semi-discrete Morris-Lecar-type model*, J.

Math. Biol., 38 (1999), pp. 1–20.
[17] R. Hobart, *Peierls-barrier minima*, J. Appl. Phys., 36 (1965), pp. 1948–1952.
[18] J. Kastrup, R. Hey, K. Ploog, H. T. Grahn, L. L. Bonilla, M. Kindelan, M. Moscoso, A. Wacker, and J. Galán, *Electrically tunable GHz oscillations in doped GaAs-AlAs superlattices*, Phys. Rev. B (3), 55 (1997), pp. 2476–2488.
[19] J. P. Keener, *Waves in excitable media*, SIAM J. Appl. Math., 39 (1980), pp. 528–548.
[20] J. P. Keener, *Propagation and its failure in coupled systems of discrete excitable cells*, SIAM J. Appl. Math., 47 (1987), pp. 556–572.
[21] J. P. Keener and J. Sneyd, *Mathematical Physiology*, Springer, New York, 1998, Chapter 9.
[22] J. P. Keener, *Propagation of waves in an excitable medium with discrete release sites*, SIAM J. Appl. Math., 61 (2000), pp. 317–334.
[23] J. R. King and S. J. Chapman, *Asymptotics beyond all orders and Stokes lines in nonlinear differential-difference equations*, European J. Appl. Math., 12 (2001), pp. 433–463.
[24] J. Mallet-Paret, *The global structure of traveling waves in spatially discrete dynamical systems*, J. Dynam. Differential Equations, 11 (1999), pp. 49–127.
[25] C. C. McIntyre and W. M. Grill, *Excitation of central nervous system neurons by nonuniform electric fields*, Biophys. J., 76 (1999), pp. 878–888.
[26] J. Nagumo, S. Arimoto, and S. Yoshizawa, *An active pulse transmission line simulating nerve axon*, Proc. Inst. Radio Engineers, 50 (1962), pp. 2061–2070.
[27] J. Rinzel and J. B. Keller, *Traveling wave solutions of a nerve conduction equation*, Biophys. J., 13 (1973), pp. 1313–1337.
[28] A. C. Scott, *The electrophysics of a nerve fiber*, Rev. Modern Phys., 47 (1975), pp. 487–533.
[29] J. J. Struijk, *The extracellular potential of a myelinated nerve fiber in an unbounded medium and in nerve cuff models*, Biophys. J., 72 (1997), pp. 2457–2469.
[30] G. de Vries, A. Sherman, and H.-R. Zhu, *Diffusively coupled bursters: Effects of cell heterogeneity*, Bull. Math. Biol., 60 (1998), pp. 1167–1200.
[31] B. Zinner, *Existence of traveling wavefront solutions for the discrete Nagumo equation*, J. Differential Equations, 96 (1992), pp. 1–27.

SIAM J. APPL. MATH.
Vol. 63, No. 2, pp. 636–682

BIFURCATION ANALYSIS OF A PREDATOR-PREY SYSTEM WITH NONMONOTONIC FUNCTIONAL RESPONSE*

HUAIPING ZHU†, SUE ANN CAMPBELL‡, AND GAIL S. K. WOLKOWICZ†

Abstract. We consider a predator-prey system with nonmonotonic functional response: $p(x) = \frac{mx}{ax^2+bx+1}$. By allowing b to be negative ($b > -2\sqrt{a}$), $p(x)$ is concave up for small values of $x > 0$ as it is for the sigmoidal functional response. We show that in this case there exists a Bogdanov–Takens bifurcation point of codimension 3, which acts as an organizing center for the system. We study the Hopf and homoclinic bifurcations and saddle-node bifurcation of limit cycles of the system. We also describe the bifurcation sequences in each subregion of parameter space as the death rate of the predator is varied. In contrast with the case $b \geq 0$, we prove that when $-2\sqrt{a} < b < 0$, a limit cycle can coexist with a homoclinic loop. The bifurcation sequences involving Hopf bifurcations, homoclinic bifurcations, as well as the saddle-node bifurcations of limit cycles are determined using information from the complete study of the Bogdanov–Takens bifurcation point of codimension 3 and the geometry of the system. Examples of the predicted bifurcation curves are also generated numerically using XPPAUT. Our work extends the results in [F. Rothe and D. S. Shafer, *Proc. Roy. Soc. Edinburgh Sect. A*, 120 (1992), pp. 313–347] and [S. Ruan and D. Xiao, *SIAM J. Appl. Math.*, 61 (2001), pp. 1445–1472].

Key words. predator-prey system, Hopf bifurcation, homoclinic bifurcation, Bogdanov–Takens bifurcation, saddle-node bifurcation of limit cycles, limit cycle

AMS subject classifications. Primary, 34C25, 92D25; Secondary, 58F14

PII. S0036139901397285

1. Introduction.

The classical predator prey model with an inhibition response function was introduced in Freedman and Wolkowicz [12] to establish a veritable paradox of enrichment. In this paper we analyze the classical predator-prey model with a specific inhibition response function, the Holling type IV response function. In particular, we study the model

$$\begin{cases} \dot{x} &= rx\left(1-\dfrac{x}{K}\right) - yp(x) = p(x)(F(x)-y), \\ \dot{y} &= y(-d+cp(x)), \\ & x(0) \geq 0, y(0) \geq 0, \end{cases} \tag{1.1}$$

where x and y denote the density of the prey and predator populations, respectively, r, K, d, and c are positive constants, and $F(x) = rx(1-\frac{x}{K})/p(x)$.

The specific growth rate of the prey population in the absence of predator population is assumed to satisfy logistic growth, and so r denotes the intrinsic growth rate of the prey population, and K denotes the carrying capacity. The natural death rate of the predators is denoted by d, and the predator response function is denoted by $p(x)$. It is assumed that the rate of conversion of prey captured to predator is proportional

*Received by the editors October 31, 2001; accepted for publication (in revised form) May 23, 2002; published electronically December 19, 2002. This work was supported by the Natural Science and Engineering Research Council and the MITACS Network of Centres of Excellence.
http://www.siam.org/journals/siap/63-2/39728.html

†Department of Mathematics and Statistics, McMaster University, Hamilton, Ontario, Canada L8S 4K1 (hzhu@icarus.math.mcmaster.ca, wolkowic@icarus.math.mcmaster.ca).

‡Department of Applied Mathematics, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1 and Centre for Nonlinear Dynamics in Physiology and Medicine, McGill University, Montreal, Quebec, Canada H3G 1Y6 (sacampbe@duchatelet.math.uwaterloo.ca).

to the predator response function, where c is the constant of proportionality or yield constant.

The predator-prey system (1.1) has been extensively studied by many authors. (See Ruan and Xiao [28] and Wolkowicz [30] for an extended list of references.) In the literature, different functions have been used to model the predator response. (See Holling [15] and Freedman and Wolkowicz [12] for a description of the general conditions that this function should satisfy.) In Wolkowicz [30], assuming only these general conditions and a technical assumption, a complete one parameter bifurcation analysis of (1.1) was carried out using the carrying capacity K as the bifurcation parameter. It was proved that the model has rich dynamics, including parameters for which there is a homoclinic bifurcation. It was pointed out that both supercritical and subcritical Hopf bifurcations are possible, and that when there is a subcritical Hopf bifurcation, there must be parameters for which there is a saddle-node bifurcation of limit cycles and a range of parameters for which there are at least two limit cycles. To do a more detailed analysis, it is useful to specify the function $p(x)$.

The form most often used (see, for example, Holling [14]), is the Holling type II form, $p(x) = \frac{mx}{b+x}$, also associated with Monod and Michaelis–Menten. This is an increasing function that saturates, i.e., has a finite positive limit as x approaches infinity. For a very nice description of the biological interpretation of each of the parameters, see Rinaldi, Muratori, and Kuznetsov [24], where they also study the effect of periodically forcing each of the parameters individually, and propose a universal bifurcation diagram.

In this paper we consider the Holling type IV functional response, associated with Monod–Haldane (see Andrews [1]):

$$p(x) = \frac{mx}{ax^2 + bx + 1}, \tag{1.2}$$

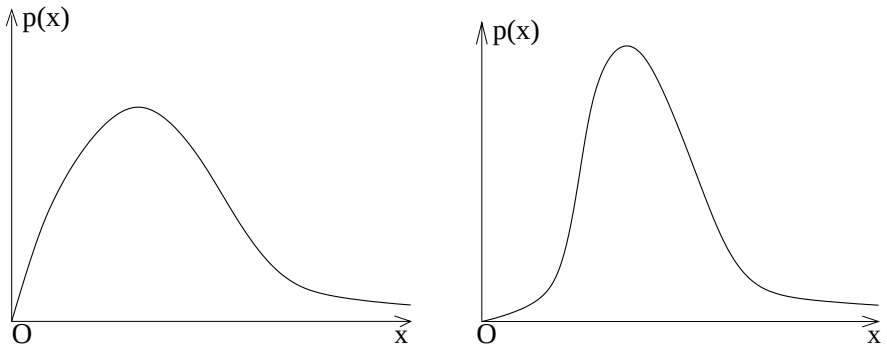
where a and m are positive constants, and $b > -2\sqrt{a}$ (so that $ax^2 + bx + 1 > 0$ for all $x \geq 0$ and hence $p(x) > 0$ for all $x > 0$).

When a is positive, this function increases to a maximum and then decreases, approaching zero as x approaches infinity. Thus, $p(x)$ models the situation where the prey can better defend or disguise themselves when their population becomes large enough, a phenomenon called group defense. See [12] and [28] for more information about this phenomenon and examples of populations that use this strategy.

The response function (1.2) has been primarily considered assuming m positive and a and b nonnegative. In this case, for x sufficiently small, $p(x)$ resembles the Holling type II model while for large x the effect of inhibition is seen (Figure 1.1(a)). If $-2\sqrt{a} < b < 0$ and a is nonnegative, $p(x)$ remains nonnegative, the inhibition effect is still observed for large x, however, for x small $p(x)$ resembles the Holling type III (sigmoidal) function (Figure 1.1(b)). We shall show in section 5 that for model (1.1) with the nonmonotonic response function (1.2), the organizing center of the bifurcation diagram is at

$$b = -\sqrt{a}, \qquad d = \frac{mc}{b + 2\sqrt{a}}, \qquad K = \frac{2}{\sqrt{a}},$$

where there is a Bogdanov–Takens bifurcation of codimension 3. Ignoring the negative, but physically relevant, values of b misses this important fact as well as some rich dynamics of the model.



(a) $b \geq 0$ (b) $-2\sqrt{a} < b < 0$

FIG. 1.1. *Response functions.*

Using the response function (1.2) in (1.1), the model to be considered is

$$\begin{cases} \dot{x} &= rx\left(1-\dfrac{x}{K}\right) - y\dfrac{mx}{ax^2+bx+1} = p(x)(F(x)-y), \\ \dot{y} &= y\left(-d+c\dfrac{mx}{ax^2+bx+1}\right) = y(-d+cp(x)), \\ & x(0)\geq 0, y(0)\geq 0, \end{cases} \tag{1.3}$$

where

$$F(x) = \frac{r}{m}\left(1-\frac{x}{K}\right)(ax^2+bx+1) \tag{1.4}$$

and

$$K,\ r,\ m,\ a,\ c,\ d > 0, \quad \text{and} \quad b > -2\sqrt{a}. \tag{1.5}$$

Rothe and Shafer [25, 26] considered the system

$$\begin{cases} \dfrac{dx}{d\tau} &= rx\left[\left(1-\dfrac{x}{K}\right)\left(\left(\dfrac{x}{\lambda}-1\right)\left(\dfrac{x}{\mu}-1\right)+x\right)-y\right], \\ \dfrac{dy}{d\tau} &= -y\Big(\dfrac{x}{\lambda}-1\Big)\Big(\dfrac{x}{\mu}-1\Big), \end{cases} \tag{1.6}$$

which may be obtained from (1.3) by rescaling time t via

$$\tau = \int_0^t \left[\frac{1}{ax^2(t)+bx(t)+1}\right] dt$$

and by assuming that the system always has two equilibria inside the positive quadrant. Using results for polynomial systems and taking $(\frac{1}{K}, \frac{1}{\lambda}, \frac{1}{\mu})$ as parameters, the authors studied the bifurcations of the model. Rothe and Shafer were the first to consider the case b negative ($-2\sqrt{a} < b$). However, after the transformation and reduction to (1.6), the effect of allowing $b < 0$ was hidden. They proved that there is a set of parameters for which there is a cusp of codimension 2, in a neighborhood of

which the system realizes every phase portrait possible under small smooth perturbation. They also indicated that there is a cusp of codimension at least 3, but did not prove that the codimension is exactly 3. They point out for parameters near this cusp there is a semi-stable limit cycle. The results of [26] are in terms of parameters which are composites of the model parameters and thus are more difficult to interpret for the biological system.

In [28], Ruan and Xiao restricted $b = 0$ in (1.3) and carried out a global analysis. In particular, they proved that a limit cycle cannot coexist with a homoclinic loop. They determined a set of parameters for which the system has a unique limit cycle which is stable and another for which no cycles exist. They also studied the Bogdanov–Takens bifurcation of codimension 2.

Model (1.3) involves 7 parameters (see (1.5)) that all have biological interpretations. By rescaling the state variables and time we could eliminate 3 of them. For example, we could eliminate a, m, and c by using

$$(t,\ x,\ y) \longrightarrow \left(\frac{\sqrt{a}}{mc}t,\ \frac{1}{\sqrt{a}}x,\ \frac{c}{\sqrt{a}}y\right) \tag{1.7}$$

and replacing r by $\frac{\sqrt{a}}{mc}r$, K by $\sqrt{a}K$, d by $\frac{\sqrt{a}}{mc}d$, and b by $\frac{1}{\sqrt{a}}b$. Similarly, using the change of variables

$$(t,\ x,\ y) \longrightarrow \left(\frac{t}{r},\ \frac{r}{mc}x,\ \frac{r}{m}y\right),$$

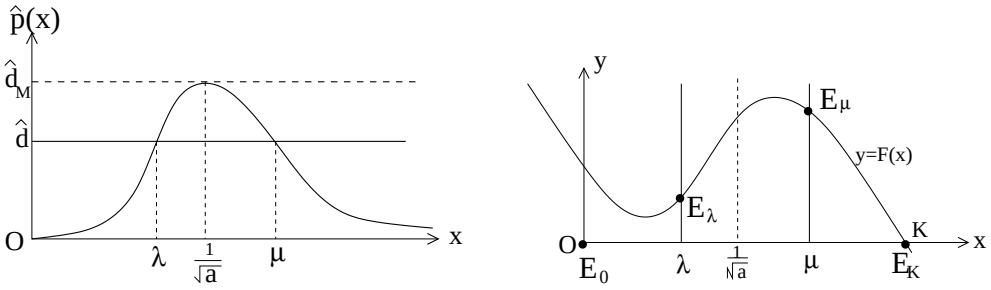
and replacing a by $(\frac{r}{mc})^2 a$, K by $\frac{mc}{r}K$, d by $\frac{1}{r}d$, and b by $\frac{r}{mc}b$ we could eliminate r, m, and c. We choose not to do this so that the effect of all the parameters may be easily seen in our results.

The x and y axes, the nonnegative cone and its interior are all invariant sets with respect to system (1.3). A standard phase plane argument can be used to show that all solutions initiating in the positive cone are bounded (see [30]).

This paper is organized as follows. Section 2 contains the linear analysis of the equilibria, where we emphasize how this depends on the slope of the prey isocline. In section 3 we give the geometric properties of the predator and prey isoclines, and we show how the parameters (1.5) influence the geometry of the isoclines. In section 4, we study the effect of this geometry on the existence, number, and criticality of Hopf bifurcations which occur as $\hat{d} = \frac{d}{mc}$ is varied. We determine that the parameter r plays no role here. In section 5, we continue our analysis, examining the degenerate equilibria, especially the cusp points of codimension 2 and 3. In section 6 we consider the global dynamics. We conclude with section 7 where we summarize our results, compare them with other closely related results, and indicate their biological significance.

2. Linear stability analysis. We consider equilibrium solutions to exist only if they lie in the nonnegative cone. System (1.3) has at most four equilibrium solutions. Two always lie on the boundary of the nonnegative cone: $E_0 = (0, 0)$, representing the extinction of both species, and $E_K = (K, 0)$, representing the extinction of the predator population and the density of the prey population equilibrating at the carrying capacity.

Let $\lambda \le \mu$ denote the two possible solutions of the quadratic equation $cp(x) = d$, and $E_\lambda = (\lambda, F(\lambda))$ and $E_\mu = (\mu, F(\mu))$ denote the corresponding equilibria. Whether zero, one, or both of these other equilibria exist and sit in the nonnegative cone



(a) Parameters λ and μ (b) Isoclines and equilibria

FIG. 2.1. *Typical geometry for system* (1.3).

depends on the relative positions of the prey isocline $y = F(x)$ and the predator isoclines $x = \lambda$ and $x = \mu$. Figure 2.1(b) illustrates one of the possible positions, where both E_λ and E_μ exist.

Define

$$\hat{d} = \frac{d}{mc}, \qquad \hat{p}(x) = \frac{p(x)}{m}. \tag{2.1}$$

Then $cp(x) = d$ or $\hat{p}(x) = \hat{d}$ is equivalent to

$$\hat{g}(x) = a\hat{d}x^2 - (1 - b\hat{d})x + \hat{d} = 0. \tag{2.2}$$

If we define

$$\hat{d}_M = \frac{1}{b + 2\sqrt{a}}, \qquad \Delta_0 = (1 - b\hat{d})^2 - 4a\hat{d}^2, \tag{2.3}$$

then from a simple calculation it follows that for $\hat{d} \in (0, \hat{d}_M]$, (2.2) has two positive roots $\lambda \le \mu$ where

$$\lambda := \frac{1 - b\hat{d} - \sqrt{\Delta_0}}{2a\hat{d}}, \qquad \mu := \frac{1 - b\hat{d} + \sqrt{\Delta_0}}{2a\hat{d}}. \tag{2.4}$$

If $\hat{d} \in (0, \hat{d}_M)$, then $\lambda < \frac{1}{\sqrt{a}} < \mu$. As $\hat{d}$ increases, λ increases and μ decreases. When $\hat{d} = \hat{d}_M$, $\lambda = \mu = \frac{1}{\sqrt{a}}$ (Figure 2.1(a)). When $\hat{d} > \hat{d}_M$, λ and μ are no longer real. Then E_0 and E_K are the only equilibria in the nonnegative cone, and E_K attracts all orbits initiating in the positive cone.

Next we investigate the stability of the equilibrium solutions E_0, E_K, E_λ, and E_μ by linearizing about each one. The variational matrix about any equilibrium $(\bar{x}, \bar{y})$ is

$$V(\bar{x}, \bar{y}) = \begin{bmatrix} p'(\bar{x})(F(\bar{x}) - \bar{y}) + p(\bar{x})F'(\bar{x}) & -p(\bar{x}) \\ cp'(\bar{x})\bar{y} & cp(\bar{x}) - d \end{bmatrix}. \tag{2.5}$$

TABLE 2.1
Linear analysis of the equilibrium solutions.

Fixed point	$K < \lambda < \mu$	$\lambda < K < \mu$	$\lambda < \mu < K$	
E_0	saddle	saddle	saddle	
E_K	attracting node	saddle	attracting node	
E_λ	does not exist	repeller	repeller	$F'(\lambda) > 0$
		attractor	attractor	$F'(\lambda) < 0$
E_μ	does not exist	does not exist	saddle	

An easy calculation indicates that E_0 is always a saddle point. For E_μ, note that the determinant of $V(\mu, F(\mu))$ is

$$\det\Big(V(\mu, F(\mu))\Big) = c\bar{y}p(\mu)p'(\mu) = \frac{c\hat{d}m^2F(\mu)(1 - a\mu^2)}{(a\mu^2 + b\mu + 1)^2}, \tag{2.6}$$

and that for $\hat{d} \in (0, \hat{d}_M)$, $\mu > \frac{1}{\sqrt{a}}$. Hence, if E_μ lies in the positive cone, i.e., when $\mu < K$, $\det(V(\mu, F(\mu))) < 0$, therefore E_μ is a saddle point.

Similarly, when E_λ lies in the positive cone, i.e., if $\lambda < K$, $\det(V(\lambda, F(\lambda))) > 0$. Further, the trace of $V(\lambda, F(\lambda))$ is

$$\text{tr } V(\lambda, F(\lambda)) = p(\lambda)F'(\lambda). \tag{2.7}$$

Hence, E_λ is an attractor (resp., repeller) if $F'(\lambda) < 0$ (resp., $F'(\lambda) > 0$).

From the discussion above, it is clear that there is a saddle-node bifurcation involving E_λ and E_μ when $\hat{d} = \hat{d}_M$.

The two eigenvalues of the variational matrix about E_K are $-r < 0$ and $-d + cp(K) = mc[\hat{p}(K) - \hat{d}]$. Thus E_K is an attracting node if $\hat{d} > \hat{p}(K)$. It is a saddle point if $\hat{d} < \hat{p}(K)$. If $\hat{d} = \hat{p}(K)$, E_K undergoes a transcritical bifurcation. As $\hat{d}$ increases from 0, the steady state bifurcations outlined below occur.

1. If $K < \frac{1}{\sqrt{a}}$, denote $\hat{d}_{\lambda K} = \hat{p}(K)$. A transcritical bifurcation involving E_λ and E_K occurs when $\hat{d} = \hat{d}_{\lambda K}$. E_K changes its stability from a saddle point to an attracting node. When $\hat{d} = \hat{d}_M$, a saddle-node bifurcation involving E_λ and E_μ occurs outside the nonnegative cone.
2. If $K > \frac{1}{\sqrt{a}}$, denote $\hat{d}_{\mu K} = \hat{p}(K)$. A transcritical bifurcation involving E_μ and E_K occurs when $\hat{d} = \hat{d}_{\mu K}$. E_K changes its stability from a saddle point to an attracting node. When $\hat{d} = \hat{d}_M$, a saddle-node bifurcation involving E_λ and E_μ occurs inside the positive cone.
3. If $K = \frac{1}{\sqrt{a}}$, then $\hat{p}(K) = \hat{d}_M$. When $\hat{d} = \hat{d}_M$, E_λ, E_μ, and E_K coalesce at E_K. Phase portrait analysis shows that E_K is an asymptotically stable degenerate node.

The linear analysis for system (1.1) is summarized in Table 2.1.

3. Geometry of the isoclines. The geometry of the prey and predator isoclines plays an important role in the analysis of both the local and global bifurcations. We begin with a useful observation about the intersections of the prey isocline $y = F(x)$ with the predator isoclines, the lines $x = \lambda$ and $x = \mu$.

LEMMA 3.1. *Consider $F(\lambda)$ and $F(\mu)$:*

1. *If* $0 < K < \frac{2}{\sqrt{a}}$ *and* $\hat{d} \in (0, \hat{d}_M)$, *then* $F(\lambda) > F(\mu)$.
2. *If* $K = \frac{2}{\sqrt{a}}$, *then* $F(\lambda) = F(\mu)$ *if and only if* $\hat{d} = \hat{d}_M$. *Otherwise, if* $\hat{d} \in (0, \hat{d}_M)$, *then* $F(\lambda) > F(\mu)$.
3. *If* $K > \frac{2}{\sqrt{a}}$, *then there exists* $\hat{d}_c$,

$$\hat{d}_c = \frac{1}{aK + b}, \tag{3.1}$$

satisfying $\hat{d}_c \in (0, \hat{d}_M)$ *such that*
- *if* $0 < \hat{d} < \hat{d}_c$, *then* $F(\lambda) > F(\mu)$;
- *if* $\hat{d} = \hat{d}_c$, *then* $F(\lambda) = F(\mu)$;
- *if* $\hat{d}_c < \hat{d} < \hat{d}_M$, *then* $F(\lambda) < F(\mu)$;
- *if* $\hat{d} = \hat{d}_M$, *then* $F(\lambda) = F(\mu)$.

Proof. If $\hat{d} \in (0, \hat{d}_M)$, two interior equilibria E_λ and E_μ exist and $\lambda < \mu$. By (2.2), we have

$$\lambda + \mu = \frac{1 - b\hat{d}}{a\hat{d}}, \qquad \lambda\mu = \frac{1}{a}. \tag{3.2}$$

Then

$$\begin{aligned} F(\lambda) - F(\mu) &= \frac{r}{mK}\left[-a(\lambda^3 - \mu^3) + (aK - b)(\lambda^2 - \mu^2) + (bK - 1)(\lambda - \mu)\right] \\ &= \frac{r(\lambda - \mu)}{mK}\left[-a((\lambda + \mu)^2 - \lambda\mu) + (aK - b)(\lambda + \mu) + (bK - 1)\right] \\ &= -\frac{r(\lambda - \mu)}{a\hat{d}^2 mK}\left[(1 - b\hat{d})^2 - a\hat{d}^2 - \hat{d}(aK - b)(1 - b\hat{d}) - a\hat{d}^2(bK - 1)\right] \\ &= -\frac{r(\lambda - \mu)}{a\hat{d}^2 mK}[1 - (aK + b)\hat{d}]. \end{aligned} \tag{3.3}$$

For $0 < K < \frac{2}{\sqrt{a}}$ and $0 < \hat{d} < \hat{d}_M$, $(aK + b)\hat{d} < (b + 2\sqrt{a})\hat{d}_M < 1$. Note that $\lambda < \mu$; therefore, $F(\lambda) > F(\mu)$.

For $K = \frac{2}{\sqrt{a}}$, if $\hat{d} \in (0, \hat{d}_M)$, then $(aK + b)\hat{d} < 1$, hence $F(\lambda) > F(\mu)$. It also follows from (3.3) that $F(\lambda) = F(\mu)$ if and only if $\hat{d} = \hat{d}_M$.

Assume that $K > \frac{2}{\sqrt{a}}$. By $F(\lambda) - F(\mu) = 0$ we obtain either $\lambda = \mu$ or

$$1 - (aK + b)\hat{d} = 0. \tag{3.4}$$

Hence, if $\hat{d} = \hat{d}_M$, $F(\lambda) = F(\mu)$. If $\hat{d} \in (0, \hat{d}_M)$, we can solve (3.4) to obtain $\hat{d} = \hat{d}_c$ such that the rest of the results follow. ∎

The prey isocline $y = F(x)$ is a cubic polynomial with $\lim_{x \to -\infty} F(x) = \infty$ and $\lim_{x \to \infty} F(x) = -\infty$. It is either decreasing or has two humps, a local minimum with x coordinate H_m and a local maximum with x coordinate H_M, where H_m and H_M are the solutions of the quadratic equation $F'(x) = 0$, or, equivalently,

$$3ax^2 - 2(aK - b)x + 1 - bK = 0. \tag{3.5}$$

Let

$$\Delta_1 = a^2K^2 + abK + b^2 - 3a. \tag{3.6}$$

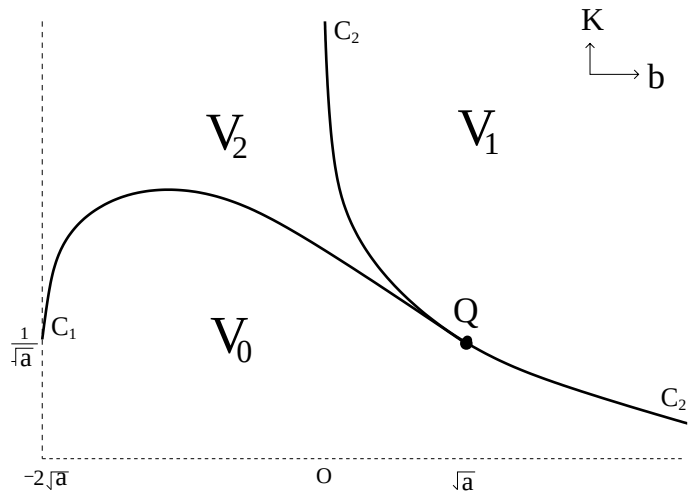


FIG. 3.1. *Three basic regions V_0, V_1, and V_2 in the bK plane, $Q = (\sqrt{a}, \frac{1}{\sqrt{a}})$ (See Proposition 3.2).*

Then when $\Delta_1 \geq 0$, we have

$$H_m := \frac{1}{3a}[aK - b - \sqrt{\Delta_1}\,], \qquad H_M := \frac{1}{3a}[aK - b + \sqrt{\Delta_1}\,]. \tag{3.7}$$

H_m is always to the left of H_M, i.e., $H_m \leq H_M$. The number and the position of the humps of the prey isocline inside the positive cone are determined by the signs of Δ_1 and $1 - bK$. For $K > 0$, the curve defined by $\Delta_1 = 0$ (part of an ellipse) is tangent to $1 - bK = 0$ at the point $Q(\sqrt{a}, \frac{1}{\sqrt{a}})$.

A straightforward analysis of the signs of the quantities Δ_1 and $1 - bK$ gives the following proposition.

PROPOSITION 3.2. *The two curves*

$$\begin{aligned} C_1 &: K = \frac{1}{2a}[\sqrt{3(4a - b^2)} - b], \qquad -2\sqrt{a} \leq b \leq \sqrt{a}, \\ C_2 &: K = \frac{1}{b}, \qquad b > 0, \end{aligned} \tag{3.8}$$

divide the region $K > 0$, $b > -2\sqrt{a}$ into 3 subregions V_0, V_1, and V_2 (see Figure 3.1):

$$\begin{aligned} V_0 &= \left\{ \begin{array}{l} -2\sqrt{a} < b < \sqrt{a} \\ 0 < K < \frac{1}{2a}[\sqrt{3(4a - b^2)} - b] \end{array} \right\} \cup \left\{ \begin{array}{l} \sqrt{a} \leq b \\ 0 < K < \frac{1}{b} \end{array} \right\}, \\ V_1 &= \left\{ 0 < b, \frac{1}{b} \leq K \right\}, \\ V_2 &= \left\{ \begin{array}{l} -2\sqrt{a} < b \leq 0 \\ \frac{1}{2a}[\sqrt{3(4a - b^2)} - b] < K \end{array} \right\} \cup \left\{ \begin{array}{l} 0 < b < \sqrt{a} \\ \frac{1}{2a}[\sqrt{3(4a - b^2)} - b] < K < \frac{1}{b} \end{array} \right\}. \end{aligned} \tag{3.9}$$

In regions V_0, V_1, and V_2, the prey isocline has 0, 1, and 2 humps in the first quadrant, respectively (see Figure 3.2).

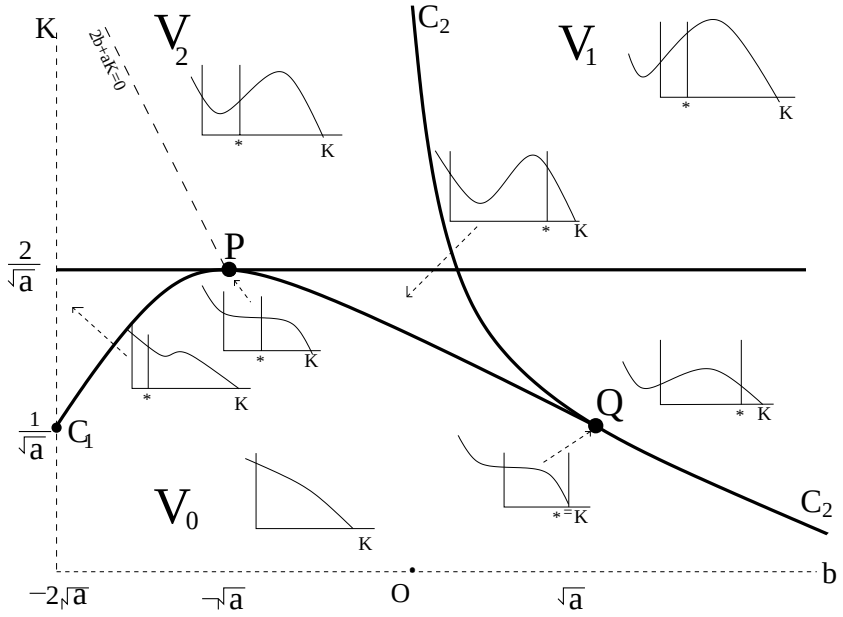


FIG. 3.2. *The positions of the two humps of the prey isocline as a function of K and b. A $*$ indicates the position of the line $x = \frac{1}{\sqrt{a}}$.*

1. *Along C_1, the two humps of the prey isocline coalesce at an inflection point with x coordinate*

$$H_I = \frac{aK - b}{3a} \quad (H_I = H_m = H_M).$$

2. *Along C_2,*
   - *if $K > \frac{1}{\sqrt{a}}$, then $H_m = 0$, i.e., the left hump sits on the y-axis;*
   - *if $0 < K < \frac{1}{\sqrt{a}}$, then $H_M = 0$, i.e., the right hump sits on the y-axis;*
   - *if $K = \frac{1}{\sqrt{a}}$, then $H_m = H_M = H_I = 0$, i.e., the inflection point sits on the y-axis.*
3. *In region V_0, the prey isocline is decreasing and hence there are no humps inside the first quadrant.*
4. *In region V_1, only the right hump H_M sits inside the positive cone and*
   - *if $K = \frac{2}{\sqrt{a}}$, then $H_M = \frac{1}{\sqrt{a}}$,*
   - *if $K > \frac{2}{\sqrt{a}}$, then $H_M > \frac{1}{\sqrt{a}}$,*
   - *if $K < \frac{2}{\sqrt{a}}$, then $H_M < \frac{1}{\sqrt{a}}$.*
5. *In region V_2, both humps sit inside the first quadrant. For $(b, K) \in V_2$,*
   - *if $K = \frac{2}{\sqrt{a}}$ and $b < -\sqrt{a}$, then $H_m = \frac{1}{\sqrt{a}}$,*
   - *if $K = \frac{2}{\sqrt{a}}$ and $b > -\sqrt{a}$, then $H_M = \frac{1}{\sqrt{a}}$,*
   - *if $K = \frac{2}{\sqrt{a}}$ and $b = -\sqrt{a}$, i.e., at the point $P = (-\sqrt{a}, \frac{2}{\sqrt{a}})$, $H_m = H_M = \frac{1}{\sqrt{a}}$,*
   - *if $K > \frac{2}{\sqrt{a}}$, then $H_m < \frac{1}{\sqrt{a}} < H_M$,*
   - *if $K < \frac{2}{\sqrt{a}}$ and $-2\sqrt{a} < b < -\sqrt{a}$, then $\frac{1}{\sqrt{a}} < H_m < H_M$*

- *if* $K < \frac{2}{\sqrt{a}}$ *and* $-\sqrt{a} < b < \sqrt{a}$, *then* $H_m < H_M < \frac{1}{\sqrt{a}}$.

Next we consider how the left and right humps of the prey isocline move as either b or K is varied.

PROPOSITION 3.3.

1. *For any fixed* b *satisfying* $b > -2\sqrt{a}$, *as* K *increases,* H_M *moves to the right and* H_m *moves to the left. Also*

$$\lim_{K\to\infty} H_M = \infty,$$
$$\lim_{K\to\infty} H_m = -\frac{b}{2a}. \quad \textit{(Note that this is positive if and only if } b < 0.\textit{)}$$

2. *For any fixed* $K > 0$,
   - *if* $2b + aK \le 0$ *(and* $-2\sqrt{a} < b < -\sqrt{a}$*), as* b *increases,*
     - H_M *moves left,*
     - *if* $0 < K < \frac{2}{\sqrt{a}}$, *then* H_m *moves right,*
     - *if* $K > \frac{2}{\sqrt{a}}$, *then* H_m *moves left,*
     - *if* $K = \frac{2}{\sqrt{a}}$, *then* $H_m = \frac{1}{\sqrt{a}}$;
3. *if* $2b + aK > 0$, *as* b *increases,*
   - H_m *moves left,*
   - *if* $0 < K < \frac{2}{\sqrt{a}}$, *then* H_M *moves right,*
   - *if* $K > \frac{2}{\sqrt{a}}$, *then* H_M *moves left,*
   - *if* $K = \frac{2}{\sqrt{a}}$, *then* $H_M = \frac{1}{\sqrt{a}}$.

Also

$$\lim_{b\to -2\sqrt{a}} H_m = \frac{2 + \sqrt{a} - |\sqrt{a}K - 1|}{3\sqrt{a}},$$
$$\lim_{b\to -2\sqrt{a}} H_M = \frac{2 + \sqrt{a} + |\sqrt{a}K - 1|}{3\sqrt{a}},$$
$$\lim_{b\to \infty} H_m = -\infty,$$
$$\lim_{b\to \infty} H_M = \frac{K}{2}.$$

Proof. The limits follow from straightforward calculations.

To study the movement of the humps, we need the following derivatives from (3.7):

$$\frac{\partial H_M}{\partial K} = \frac{b + 2aK + 2\sqrt{\Delta_1}}{6a\sqrt{\Delta_1}},$$
$$\frac{\partial H_m}{\partial K} = \frac{-(b + 2aK) + 2\sqrt{\Delta_1}}{6a\sqrt{\Delta_1}}; \tag{3.10}$$

$$\frac{\partial H_M}{\partial b} = \frac{2b + aK - 2\sqrt{\Delta_1}}{6a\sqrt{\Delta_1}},$$
$$\frac{\partial H_m}{\partial b} = -\frac{2b + aK + 2\sqrt{\Delta_1}}{6a\sqrt{\Delta_1}}. \tag{3.11}$$

1. Inside the region $V_1 \cup V_2$, for any fixed $b > -2\sqrt{a}$, we have $b + 2aK > 0$. By (3.10), $\frac{\partial H_M}{\partial K} > 0$, i.e., as we increase K, the right hump moves to the right. For the

left hump, by (3.10) we have

$$\frac{\partial H_m}{\partial K} = \frac{b^2 - 4a}{2a\sqrt{\Delta_1}(b + 2aK + 2\sqrt{\Delta_1})}.$$

For $b \geq 2\sqrt{a}$, the left hump does not lie inside the positive cone. Hence, if it exists, $\frac{\partial H_m}{\partial K} < 0$; i.e., as we increase K, the left hump moves left.

2. Fix K inside the region $V_1 \cup V_2$.

First note that the distance between the two humps has a minimum $\frac{2}{\sqrt{3}a}\sqrt{b^2 - a}$ along the line segment $2b + aK = 0$ when $-2\sqrt{a} \leq b \leq -\sqrt{a}$.

In the subregion where $2b + aK < 0$, by (3.11), $\frac{\partial H_M}{\partial b} < 0$, so as we increase b, the right hump moves left. The sign of

$$\frac{\partial H_m}{\partial b} = \frac{a(\frac{4}{a} - K^2)}{2\sqrt{\Delta_1}(2\sqrt{\Delta_1} - 2b - aK)}$$

indicates the direction that the left hump moves in.

Similarly, in the subregion where $2b + aK > 0$, the results follow from $\frac{\partial H_m}{\partial b} < 0$ and

$$\frac{\partial H_M}{\partial b} = \frac{a(\frac{4}{a} - K^2)}{2\sqrt{\Delta_1}(2b + aK + 2\sqrt{\Delta_1})}. \qquad \square$$

4. Hopf bifurcations. From the analysis in section 2, E_λ is the only candidate for a Hopf bifurcation. It follows from (2.7) that if a Hopf bifurcation occurs, it occurs when λ coincides with a hump of the prey isocline $y = F(x)$, i.e., when λ is such that $F'(\lambda) = 0$ or, equivalently,

$$3a\lambda^2 - 2(Ka - b)\lambda + 1 - Kb = 0. \tag{4.1}$$

Eliminating λ from $\hat{g}(\lambda) = 0$ and $F'(\lambda) = 0$, we obtain

$$(4a - b^2)(aK^2 + bK + 1)\hat{d}^2 + 2(abK^2 + 2(b^2 - 2a)K + b)\hat{d} + 3(1 - bK) = 0. \tag{4.2}$$

For each fixed $a > 0$, (4.2) determines the Hopf bifurcation surface $\hat{d}(b, K)$. This is illustrated in Figure 4.1. The significance of P, Q, and R will be discussed in Proposition 4.3 and Theorem 5.2.

Note that, depending on the values of a and b, there may be one or two values of $\hat{d}$ at which a Hopf bifurcation occurs. These may be found explicitly by solving (4.2) for $\hat{d}$ to obtain

$$\hat{d}_\pm = \frac{-(abK^2 + 2(b^2 - 2a)K + b) \pm (2 + bK)\sqrt{\Delta_1}}{(4a - b^2)(aK^2 + bK + 1)}. \tag{4.3}$$

4.1. Existence of Hopf bifurcations. Our analysis is based on the positions of the humps of the prey isocline relative to the vertical line $x = \frac{1}{\sqrt{a}}$. We study the Hopf bifurcation on the surface (4.2) by fixing (b, K) in each region V_i $(i = 0, 1, 2)$ and using $\hat{d}$ as a bifurcation parameter.

THEOREM 4.1. *Fix all parameters except $\hat{d} > 0$. Provided that $H_m \neq H_M$, a generic Hopf bifurcation occurs*

1. *at $E_\lambda = (H_m, F(H_m))$, when $\hat{d} = \hat{d}_-$, if $0 < H_m < \frac{1}{\sqrt{a}}$ and*

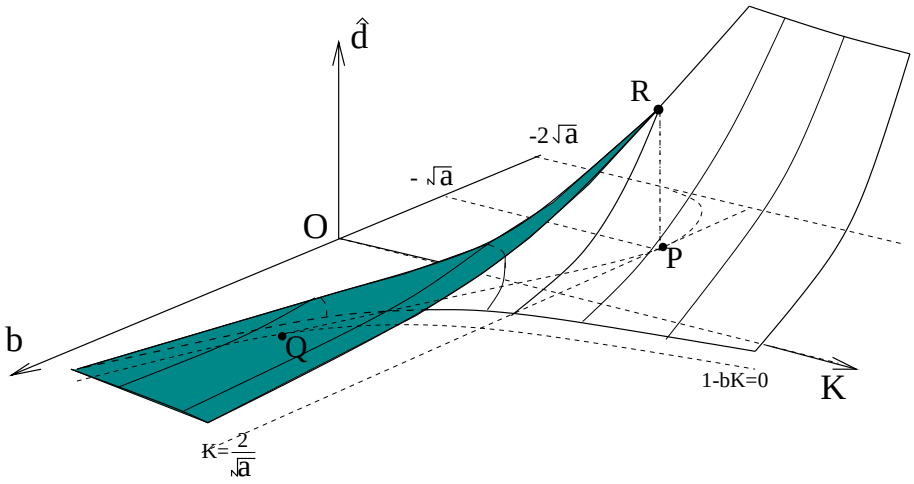


FIG. 4.1. *Hopf bifurcation surface* $\hat{d} = \hat{d}(b, K)$.

2. *at* $E_\lambda = (H_M, F(H_M))$, *when* $\hat{d} = \hat{d}_+$, *if* $0 < H_M < \frac{1}{\sqrt{a}}$.

No other nondegenerate Hopf bifurcations occur in the interior of the positive cone.

Proof. Consider the variational matrix about $E_\lambda = (\lambda, F(\lambda))$ (see (2.5)). It is clear from (2.2), (2.5), and (2.7) that $V(\lambda, F(\lambda))$ has pure imaginary eigenvalues if and only if $0 < \lambda < \frac{1}{\sqrt{a}}$ and $F'(\lambda) = 0$, and hence $\lambda = H_m$ or $\lambda = H_M$.

If $\lambda = H_m$, then $\hat{d} = \hat{p}(H_m)$. Substituting (3.7) in $\hat{p}(H_m)$ yields

$$\hat{d} = \hat{p}(H_m) = \frac{3[aK - b - \sqrt{\Delta_1}]}{2a^2K^2 + 2abK - b^2 + 6a - (2aK + b)\sqrt{\Delta_1}} = \hat{d}_-.$$

Similarly, if $\lambda = H_M$, then $\hat{d} = \hat{p}(H_m)$, and we can show that $\hat{p}(H_M) = \hat{d}_+$.

Next we verify the transversality condition. At E_λ with $\lambda = H_m$ or H_M, let γ be the real part of the eigenvalue. Then a straightforward calculation from (2.5) and (2.7) gives

$$\gamma = \frac{1}{2}p(\lambda)F'(\lambda) = -\left(\frac{rm}{2K}\right)\frac{\lambda(3a\lambda^2 - 2(aK - b)\lambda + 1 - bK)}{a\lambda^2 + b\lambda + 1}. \tag{4.4}$$

Using Maple [29], we obtain

$$\begin{aligned}\frac{\partial\gamma}{\partial\hat{d}} = -\frac{rm\frac{\partial\lambda}{\partial\hat{d}}}{27a^2K}\Big[&3a(bK-1)(2a^2K^2 + 2abK - b^2 + 6a)\\ &-(4a^3K^3 + 3a^2bK^2 + 2b^3 - 9ab + 2)\Big(Ka - b - \sqrt{\Delta_0}\Big)\Big].\end{aligned} \tag{4.5}$$

Note that $\frac{\partial\lambda}{\partial\hat{d}} = -\frac{\lambda}{\hat{d}\sqrt{\Delta_0}} \neq 0$ as long as $\lambda > 0$ and so $\frac{\partial\gamma}{\partial\hat{d}} = 0$ if and only if the term in square brackets in (4.5) equals 0. This occurs only when $H_m = 0$ or $H_M = 0$ or $H_m = H_M = H_I$. Thus the transversality condition is satisfied.

The only other equilibria of (1.3) are E_0, E_μ, and E_K. No nondegenerate Hopf bifurcation can occur at any of these equilibria, since the corresponding variational matrix always has real eigenvalues in each case. ∎

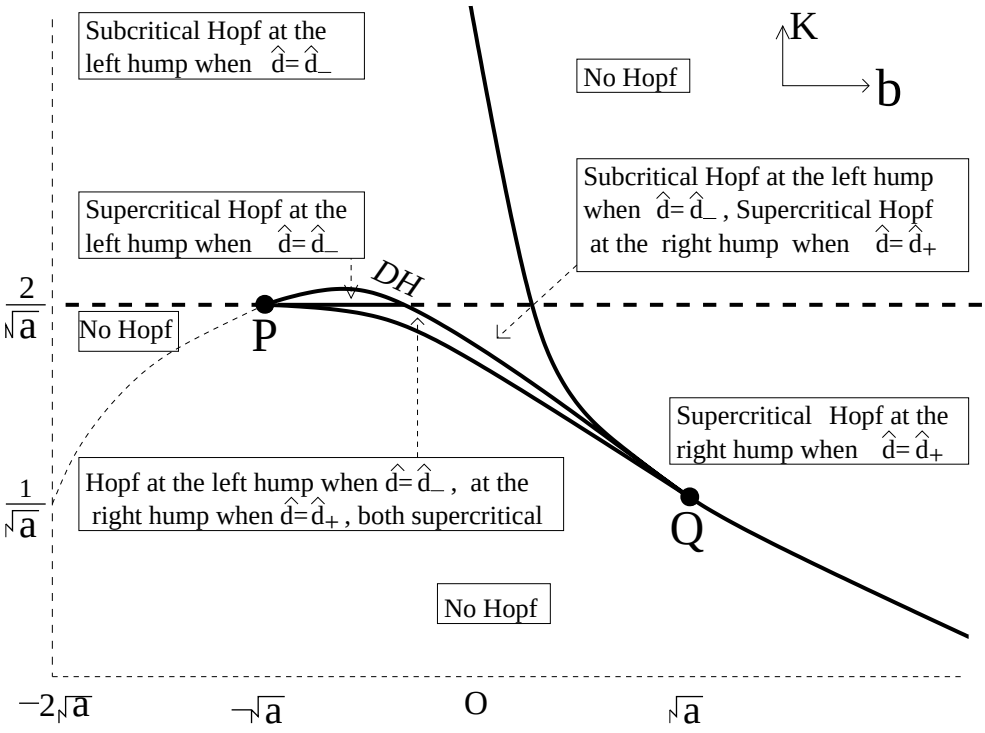


FIG. 4.2. *The existence and criticality of Hopf bifurcations in the bK plane as $\hat{d}$ is varied.*

By the above theorem, if the predator isocline has a hump inside the positive cone, and the hump is to the left of the vertical line $x = \frac{1}{\sqrt{a}}$, there exists a $\hat{d}$ defined by (4.3) such that system (1.3) undergoes a Hopf bifurcation. Hence, if we define

$$V_1^1 = \left\{ (b,K) \in V_1 \middle| K > \tfrac{2}{\sqrt{a}} \right\},$$

$$V_2^0 = \left\{ (b,K) \in V_2 \middle| b < -\sqrt{a},\ K < \tfrac{2}{\sqrt{a}} \right\}$$

from Theorem 4.1 and Proposition 3.2 (as shown in Figure 3.2), we obtain the following corollary as illustrated in Figure 4.1 and Figure 4.2.

COROLLARY 4.2. *Fix all parameters except $\hat{d} > 0$ and allow $\hat{d}$ to vary (see Figure 4.2).*

1. *No Hopf bifurcation occurs if*
   (a) $(b,K) \in V_0 \cup V_1^1 \cup V_2^0$,
   (b) $(b,K) \in C_1$,
   (c) $(b,K) \in C_2$ *and* $K < \frac{1}{\sqrt{a}}$ *or* $K > \frac{2}{\sqrt{a}}$.
2. *There is exactly one Hopf bifurcation and it occurs at $(H_M, F(H_M))$ when $\hat{d} = \hat{d}_+$ if*
   (a) $(b,K) \in C_2$ *and* $\frac{1}{\sqrt{a}} < K < \frac{2}{\sqrt{a}}$ *or*
   (b) $(b,K) \in V_1$, $K < \frac{2}{\sqrt{a}}$.
3. *There is exactly one Hopf bifurcation and it occurs at $(H_m, F(H_m))$ when $\hat{d} = \hat{d}_-$ if*

- $(b, K) \in V_2$, $K > \frac{2}{\sqrt{a}}$.

4. *There are exactly two Hopf bifurcations: one occurs at* $(H_m, F(H_m))$ *when* $\hat{d} = \hat{d}_-$, *and the other occurs at* $(H_M, F(H_M))$ *when* $\hat{d} = \hat{d}_+$ *if*
   - $(b, K) \in V_2$, $K < \frac{2}{\sqrt{a}}$, *and* $b > -\sqrt{a}$.

Proof. Recall that a Hopf bifurcation occurs when $E_\lambda = (\lambda, F(\lambda))$ coincides with a hump, i.e., when either $H_m = \lambda$ or $H_M = \lambda$ with $0 < \lambda < \frac{1}{\sqrt{a}}$. The results 2(b), 3, and 4 are the consequences of Theorem 4.1 and Proposition 3.2. It remains to prove 1 and 2(a).

Now for $(b, K) \in V_0$, $y = F(x)$ has no humps. For $(b, K) \in V_1^1$, $y = F(x)$ has only one hump, with $H_M > \frac{1}{\sqrt{a}}$. For $(b, K) \in V_2^0$, $y = F(x)$ has two humps, with $\frac{1}{\sqrt{a}} < H_m < H_M$. For $(b, K) \in C_1$, $H_m = H_M = H_I$. For $(b, K) \in C_2$, if $K < \frac{1}{\sqrt{a}}$, $H_M = 0$; if $K > \frac{2}{\sqrt{a}}$, $H_m = 0$ and $H_M < \frac{1}{\sqrt{a}}$. Thus a Hopf bifurcation is precluded in all cases. □

4.2. Criticality of the Hopf bifurcations. In a study [30] of Hopf bifurcation in systems of the form (1.1), the following formula for the Liapunov coefficient, σ, was obtained:

$$\sigma(x) = -\frac{p(x)F''(x)p''(x)}{p'(x)} + p(x)F'''(x) + 2p'(x)F''(x). \tag{4.6}$$

We will use this formula to give a complete description of the criticality of the Hopf bifurcation at $\hat{d} = \hat{d}_+$ and $\hat{d} = \hat{d}_-$.

PROPOSITION 4.3.

1. *When the Hopf bifurcation occurs at* E_λ *with* $\lambda = H_M$, *it is always supercritical.*
2. *Define the curve DH (Figures* 4.1 *and* 4.2*) connecting the two points* $P(-\sqrt{a}, \frac{2}{\sqrt{a}})$ *and* $Q(\sqrt{a}, \frac{1}{\sqrt{a}})$*:*

$$DH: \begin{aligned} &16a^4K^4 + a^2b(8a - 3b^2)K^3 - a^2(144a - 15b^2)K^2 \\ &\qquad - 8ab(9a - b^2)K + 16b^4 - 144ab^2 + 300a^2 = 0. \end{aligned} \tag{4.7}$$

   When the Hopf bifurcation occurs at E_λ *with* $\lambda = H_m$, *it is supercritical if* $(b, K) \in V_2$ *below DH; it is subcritical if* $(b, K) \in V_2$ *and* $-2\sqrt{a} < b < -\sqrt{a}$ *or above DH.*
3. *For* $(b, K) \in DH$, *a degenerate Hopf bifurcation occurs at* E_λ *with* $\lambda = H_m$ *when* $\hat{d} = \hat{d}_-$.

Proof. If the Hopf bifurcation occurs at E_λ with $\lambda = H_M$, it follows from (4.6) that we have

$$\begin{aligned} \sigma(H_M) &= \frac{F''(H_M)}{p'(H_M)}[2p'^2(H_M) - p(H_M)p''(H_M)] + p(H_M)F'''(H_M) \\ &= \frac{2m^2}{(aH_M^2 + bH_M + 1)^3}\frac{F''(H_M)}{p'(H_M)} - \frac{6ra}{mK}p(H_M). \end{aligned} \tag{4.8}$$

Note that $p'(H_M) > 0$ and $F''(H_M) < 0$. It then follows from (4.8) that $\sigma < 0$, i.e., when the Hopf bifurcation occurs at $E_\lambda = (H_M, F(H_M))$ with $\hat{d} = \hat{d}_+$, it is supercritical.

Assume that when $\hat{d} = \hat{d}_-$, a Hopf bifurcation occurs at E_λ with $\lambda = H_m$. By (4.6),

$$\sigma(\lambda) = \frac{2r[3a^2\lambda^3 - 9a\lambda - 2(b - aK)]}{K(1 - a\lambda^2)(a\lambda^2 + b\lambda + 1)}. \tag{4.9}$$

Note that when $\lambda \in (0, \frac{1}{\sqrt{a}})$, the denominator of (4.9) is positive. Consider σ defined by (4.9) as a function of b, K, and λ. Using λ as the parameter with $\lambda \in (0, \frac{1}{\sqrt{a}})$, then $\sigma = 0$ defines a simple curve connecting the two points $P(-\sqrt{a}, \frac{2}{\sqrt{a}})$ and $Q(\sqrt{a}, \frac{1}{\sqrt{a}})$. Now we develop an expression for this curve segment called DH.

If the Hopf bifurcation occurs at $x = \lambda$, then λ satisfies both (2.2) and (4.1). For the Hopf bifurcation to occur at $E_\lambda = (H_m, F(H_m))$ when $\hat{d} = \hat{d}_-$, it follows that

$$\lambda = \frac{(2 + bK)\hat{d}_-}{3 - (b + 2aK)\hat{d}_-}. \tag{4.10}$$

Substituting (4.10) into (4.9) and using (4.3), we obtain an implicit equation $\sigma = 0$ which defines a simple curve in the bK plane, along which the Liapunov coefficient of the Hopf bifurcation vanishes. Using Maple [29], we can solve the equation and obtain $K = \frac{-b \pm \sqrt{b^2 - 4a}}{2a}$, which is not real for $-\sqrt{a} \le b \le \sqrt{a}$, and an implicit equation of b and K, which can be simplified to

$$\begin{aligned} 16a^4K^4 + a^2b(8a - 3b^2)K^3 - a^2(144a - 15b^2)K^2 \\ - 8ab(9a - b^2)K + 16b^4 - 144ab^2 + 300a^2 = 0. \end{aligned} \tag{4.11}$$

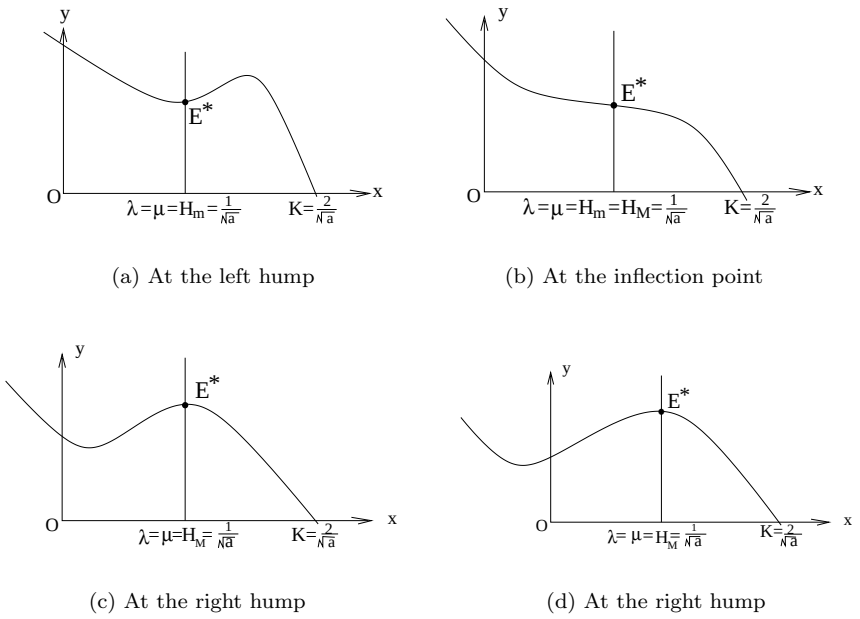
For $K > 0$, (4.11) defines a cusp curve with the cusp point located at P. The upper branch of the cusp curve is above the line $K = \frac{2}{\sqrt{a}}$ and is not relevant because it is an artifact of simplification. The lower branch passes through the point Q where it is tangent to C_2. Only the curve segment DH is relevant since there are no humps for $(b, K) \in V_0$. Along the curve segment DH, the Hopf bifurcation is degenerate. One can verify that for (b, K) above the curve DH, $\sigma > 0$, and so a subcritical Hopf bifurcation occurs at the left hump. Below the curve segment DH, $\sigma < 0$, and so a supercritical Hopf bifurcation occurs at the left hump.

At the two points $P(-\sqrt{a}, \frac{2}{\sqrt{a}})$ and $Q(\sqrt{a}, \frac{1}{\sqrt{a}})$, if the Hopf bifurcation occurs, it occurs at $\lambda = \frac{1}{\sqrt{a}}$, and the associated $\sigma = 0$. The Hopf bifurcation is therefore degenerate. □

The next theorem follows from Theorem 4.1, Corollary 4.2, and Propositions 4.3.

THEOREM 4.4. *Fix all parameters except $\hat{d} > 0$ and allow $\hat{d}$ to vary (see Figures* 3.2*,* 4.1*, and* 4.2*).*

1. *In region $V_1 \setminus V_1^1$, a supercritical Hopf bifurcation occurs at $(H_M, F(H_M))$ when $\hat{d} = \hat{d}_+$.*
2. *In region V_2, for $K > \frac{2}{\sqrt{a}}$, above the curve DH, a subcritical Hopf bifurcation occurs at $(H_m, F(H_m))$ when $\hat{d} = \hat{d}_-$.*
3. *In region V_2, for $K > \frac{2}{\sqrt{a}}$, below the curve DH, a supercritical Hopf bifurcation occurs at $(H_m, F(H_m))$ when $\hat{d} = \hat{d}_-$.*
4. *In region V_2, for $b > -\sqrt{a}$, $K < \frac{2}{\sqrt{a}}$, below the curve DH, two Hopf bifurcations occur. When $\hat{d} = \hat{d}_-$, one occurs at $(H_m, F(H_m))$. When $\hat{d} = \hat{d}_+$, one occurs at $(H_M, F(H_M))$. They are both supercritical.*
5. *In region V_2, for $K < \frac{2}{\sqrt{a}}$, above the curve DH, two Hopf bifurcations occur. When $\hat{d} = \hat{d}_-$, one occurs at $(H_m, F(H_m))$ and is subcritical. When $\hat{d} = \hat{d}_+$, one occurs at $(H_M, F(H_M))$ and is supercritical.*



(a) At the left hump (b) At the inflection point

(c) At the right hump (d) At the right hump

FIG. 5.1. *Positions of the isoclines and equilibria at the cusp bifurcation of codimension* 2, (a), (c), (d), *and* 3, (b).

5. The cusp points of codimension 2 and 3. From the analysis in sections 3 and 4, when

$$K = \frac{2}{\sqrt{a}}, \qquad d = \frac{mc}{b+2\sqrt{a}} = mc\hat{d}_M, \tag{5.1}$$

two equilibria E_λ and E_μ coalesce on the vertical line $x = \frac{1}{\sqrt{a}}$. That is,

$$E_\lambda = E_\mu = \left(\frac{1}{\sqrt{a}}, \frac{r(b+2\sqrt{a})}{2m\sqrt{a}}\right) := E^*.$$

Using (2.6) in (2.5) it follows that the equilibrium E^* has two zero eigenvalues.

From Proposition 3.3 and Figure 5.1, the position of E^* is described below. If $b \in (-2\sqrt{a}, -\sqrt{a})$, then E^* is at the left hump (Figure 5.1(a)). As b increases in this range, the right hump moves to the left until $b = -\sqrt{a}$, when the two humps coalesce and E^* is at the inflection point (Figure 5.1(b)). For $b \in (-\sqrt{a}, \infty)$, E^* is at the right hump (Figure 5.1(c)). As b increases, the left hump moves to the left until $b = \frac{\sqrt{a}}{2}$, when the left hump reaches the y-axis and leaves the first quadrant (Figure 5.1(d)).

In this section, we prove that E^* is a cusp singularity of codimension 2 for all $b \in (-2\sqrt{a}, \infty)$ except at $b = -\sqrt{a}$ where it is a cusp singularity of codimension 3. This generalizes the results in [28] and [26]. In [28] they only consider $b = 0$, and hence the cusp singularity of codimension 2. In [26], they proved that there is a set of parameters for which there is a cusp of codimension 2 for (1.6). They also indicated that there is a cusp of codimension at least 3, but did not prove that the codimension is exactly 3.

For any $b \in (-2\sqrt{a}, \infty)$ and K, d satisfying condition (5.1), system (1.3) becomes

$$\begin{cases} \dot{x} = rx\left(1 - \dfrac{\sqrt{a}}{2}x\right) - \dfrac{mxy}{ax^2 + bx + 1}, \\ \dot{y} = y\left[-\dfrac{mc}{b + 2\sqrt{a}} + \dfrac{mcx}{ax^2 + bx + 1}\right], \end{cases} \tag{5.2}$$

which has a unique equilibrium E^* in the positive cone. Using a series of transformations, we shall reduce system (5.2) to normal form.

The translation

$$X = x - \frac{1}{\sqrt{a}}, \qquad Y = y - \frac{r(b + 2\sqrt{a})}{2m\sqrt{a}}, \tag{5.3}$$

brings E^* to the origin. Expanding the right-hand side of the resulting system in a Taylor series about the origin, we obtain

$$\begin{cases} \dot{X} = -\dfrac{m}{b + 2\sqrt{a}}Y - \dfrac{r\sqrt{a}(b + \sqrt{a})}{2(b + 2\sqrt{a})}X^2 + R_{10}(X, Y), \\ \dot{Y} = -\dfrac{acr}{2(b + 2\sqrt{a})}X^2 + R_{20}(X, Y), \end{cases} \tag{5.4}$$

where R_{i0} $(i = 1, 2)$ is C^∞ in (X, Y) and $R_{i0}(X, Y) = O(|(X, Y)|^3)$.

Reversing time and making the transformation

$$X = X, \qquad Z = \frac{m}{b + 2\sqrt{a}}Y,$$

system (5.4) becomes

$$\begin{cases} \dot{X} = Z + \dfrac{r\sqrt{a}(b + \sqrt{a})}{2(b + 2\sqrt{a})}X^2 + R_{11}(X, Z), \\ \dot{Z} = \dfrac{acmr}{2(b + 2\sqrt{a})^2}X^2 + R_{21}(X, Z), \end{cases} \tag{5.5}$$

where R_{i1} $(i = 1, 2)$ is C^∞ in (X, Z) and $R_{i1}(X, Z) = O(|(X, Z)|^3)$.

Making the near-identity transformation

$$u = X, \qquad v = Z + \frac{r\sqrt{a}(b + \sqrt{a})}{2(b + 2\sqrt{a})}X^2 + R_{11}(X, Z), \tag{5.6}$$

we obtain

$$\begin{cases} \dot{u} = v, \\ \dot{v} = \delta_1 u^2 + \delta_2 uv + R_{22}(u, v), \end{cases} \tag{5.7}$$

where R_{22} is C^∞ in (X, Z), $R_{22}(u, v) = O(|(u, v)|^3)$, and

$$\delta_1 = \frac{acmr}{2(b + 2\sqrt{a})^2}, \quad \delta_2 = \frac{r\sqrt{a}(b + \sqrt{a})}{(b + 2\sqrt{a})}. \tag{5.8}$$

Thus we have the following theorem.

THEOREM 5.1. *For any $b > -2\sqrt{a}$, if d and K satisfy (5.1) and $b \neq -\sqrt{a}$, then the equilibrium E^* is a cusp point of codimension 2 (a Bogdanov–Takens bifurcation point).*

If $b = -\sqrt{a}$, the cusp point is at the inflection point of $F(x)$ (Figure 5.1(b)), and by (5.8), $\delta_2 = 0$ in the normal form (5.7). Thus the Bogdanov–Takens bifurcation is degenerate and the codimension of the cusp singularity is at least 3. To show that the codimension is exactly 3, one needs to show that system (5.7) is C^∞ equivalent to the generic normal form of the cusp point of codimension 3. This is the approach taken in the following theorem.

THEOREM 5.2. *If $(b, \hat{d}, K)$ is at the point R (Figure 4.1), i.e.,*

$$b = -\sqrt{a}, \qquad K = \frac{2}{\sqrt{a}}, \qquad d = \frac{mc}{\sqrt{a}}, \tag{5.9}$$

then the equilibrium $E^* = (\frac{1}{\sqrt{a}}, \frac{r}{2m})$ is a cusp point of codimension 3 (a degenerate Bogdanov–Takens bifurcation point).

Proof. It has been shown [8, 18, 21] that any system which has a cusp point of codimension 3 is C^∞ equivalent to the following:

$$\begin{cases} \dot{x} = y, \\ \dot{y} = x^2 + y[\beta x^3 + O(x^4)] + y^2 Q_4(x, y), \end{cases} \tag{5.10}$$

where $\beta \neq 0$. Thus, we will prove this theorem by showing that there exist smooth coordinate changes which take system (1.3) with the parameter values (5.9) into (5.10).

Under condition (5.9), $E^* = (\frac{1}{\sqrt{a}}, \frac{r}{2m})$. As in the case $b \neq -\sqrt{a}$, after translating the equilibrium to the origin and performing a Taylor expansion, we obtain

$$\begin{cases} \dot{X} = -\dfrac{m}{\sqrt{a}} Y + \sqrt{a} m X^2 Y - \dfrac{1}{2} a r X^3 + Q_{10}(X, Y), \\ \dot{Y} = -\dfrac{1}{2}\sqrt{a} c r X^2 - mc\sqrt{a} X^2 Y + \dfrac{1}{2} a r c X^3 + Q_{20}(X, Y), \end{cases} \tag{5.11}$$

where Q_{i0} $(i = 1, 2)$ is C^∞ in (X, Y) and $Q_{i0}(X, Y) = O(|(X, Y)|^4)$.

Reversing time and rescaling

$$X = X, \qquad Z = \frac{m}{\sqrt{a}} Y,$$

system (5.11) becomes

$$\begin{cases} \dot{X} = Z - a X^2 Z + \dfrac{1}{2} a r X^3 + Q_{11}(X, Z), \\ \dot{Z} = \dfrac{crm}{2} X^2 + mc\sqrt{a} X^2 Z - \dfrac{crm\sqrt{a}}{2} X^3 + Q_{21}(X, Z), \end{cases} \tag{5.12}$$

where Q_{i1} $(i = 1, 2)$ is C^∞ in (X, Z) and $Q_{i1}(X, Z) = O(|(X, Z)|^4)$.

Using the near-identity transformation

$$u = X, \qquad v = Z - a X^2 Z + \frac{1}{2} a r X^3 + Q_{11}(X, Z), \tag{5.13}$$

and changing u, v into x, y, we obtain

$$\begin{cases} \dot{x} = y, \\ \dot{y} = g_1(x) + y g_2(x) + y^2 Q_2(x, y), \end{cases} \tag{5.14}$$

where

$$g_1(x) = \frac{cmr}{2}x^2 + O(x^3),$$
$$g_2(x) = \left(mc\sqrt{a} + \frac{3}{2}ra\right)x^2 - mcax^3 + O(x^4),$$

$Q_2(x,y)$ $(i=1,2)$ is C^∞ in (x,y) and $Q_2(x,y) = O(|(X,Z)|^4)$.

Let

$$\omega_0 = ydy - [g_1(x) + yg_2(x) + y^2Q_2(x,y)]dx. \tag{5.15}$$

By using the 1-form (5.15) for system (5.14), we develop the normal form for the cusp singularity.

(1) Reduction of $g_1(x)$ to x^2. Since $g_1''(0) = crm \neq 0$, there exists a local diffeomorphism in x near the origin,

$$X = L(x) = \sqrt[3]{\frac{crm}{2}}x + O(x^2),$$

such that

$$X^2dX = g_1(x)dx. \tag{5.16}$$

By this diffeomorphism, the term $g_1(x)$ is reduced to x^2. Writing x instead of X, ω_0 becomes

$$\omega_0 = ydy - [x^2 + yg_3(x) + y^2Q_3(x,y)]dx, \tag{5.17}$$

where

$$g_3(x) = \alpha x^2 + \beta x^3 + O(x^4)$$

with $\alpha = (mc\sqrt{a} + \frac{3}{2}ar)\frac{\sqrt[3]{4acmr}}{acmr}$ and $\beta = -\frac{2}{r}$.

(2) Elimination of x^2 term from $g_3(x)$. Let $S(x,y) = \frac{1}{2}y^2 - \frac{1}{3}x^3$. Then $dS(x,y) = ydy - x^2dx$. Thus

$$yx^2dx = y^2dy - ydS. \tag{5.18}$$

Substituting (5.18) into (5.17), we obtain

$$\omega_0 = (1+\alpha y)dS(x,y) - \alpha y^2dy - y[\beta x^3 + O(x^4) + yQ_3(x,y)]dx. \tag{5.19}$$

It follows that

$$\begin{aligned}\frac{\omega_0}{1+\alpha y} &= dS(x,y) - \frac{\alpha y^2}{1+\alpha y}dy - \frac{y[\beta x^3 + O(x^4)] + yQ_4(x,y)}{1+\alpha y}dx \\ &= dS(x,y) - \frac{\alpha y^2}{1+\alpha y}dy - y[\beta x^3 + O(x^4) + yQ_4(x,y)]dx,\end{aligned} \tag{5.20}$$

where for $i = 3, 4$, $Q_i(x,y)$ is C^∞ in (x,y) and $Q_i(x,y) = O(|(x,y)|^4)$. Now a near-identity transformation

$$X = x, \quad Y = y + \cdots \tag{5.21}$$

transforms the exact 1-form $dS(x,y) - \frac{\alpha y^2}{1+\alpha y}dy$ into $dS(X,Y)$, where the term YX^3dX remains unchanged. Writing x and y instead of X and Y, system (1.3) in a neighborhood of E^* is thus equivalent to (5.10). Since $\beta = -\frac{2}{r} \neq 0$, E^* is a cusp point of codimension 3. □

By Theorem 5.1, if $K = \frac{2}{\sqrt{a}}$ and $d = \frac{mc}{b+2\sqrt{a}}$ but $b \neq -\sqrt{a}$, the cusp point is of codimension 2. One can find standard analysis for this codimension 2 bifurcation in Dumortier [6], Dumortier and Roussarie [7], Kuznetsov [18], and Mardešić [21]. In [28], the authors study this cusp point in the case $b = 0$ and develop a versal unfolding using K and d as distinguished parameters. The analysis of the cusp point of codimension 2 in the case $b \neq 0$ is similar; thus we will present only the codimension 3 versal unfolding of the cusp singularity. By the analysis in section 3, we have several different choices for the parameters to unfold the codimension 3 singularity: (b,d,K), (b,m,K), (b,c,K), $(a,b,d),\ldots$. In this paper, we take (b,d,K) as the bifurcation parameters and develop a versal unfolding for the codimension 3 cusp singularity when these three parameters are perturbed near the point $(b_0,d_0,K_0) = (-\sqrt{a}, \frac{mc}{\sqrt{a}}, \frac{2}{\sqrt{a}})$. We study the bifurcations of this unfolding by using the results in [8] and [21].

We wish to study system (1.3) for parameters (b,d,K) in a neighborhood of $(-\sqrt{a}, \frac{mc}{\sqrt{a}}, \frac{2}{\sqrt{a}})$. Thus we let

$$\begin{cases} b = -\sqrt{a} + \varepsilon_1, \\ d = \frac{mc}{\sqrt{a}} + \varepsilon_2, \\ K = \frac{2}{\sqrt{a}} + \varepsilon_3 \end{cases} \tag{5.22}$$

in (1.3) and we study the bifurcations of the resulting system

$$\begin{cases} \dot{x} = rx\left[1 - \dfrac{x}{\frac{2}{\sqrt{a}} + \varepsilon_3}\right] - \dfrac{mxy}{ax^2 + (-\sqrt{a} + \varepsilon_1)x + 1}, \\ \dot{y} = y\left[-\left(\dfrac{mc}{\sqrt{a}} + \varepsilon_2\right) + \dfrac{mcxy}{ax^2 + (-\sqrt{a} + \varepsilon_1)x + 1}\right], \end{cases} \tag{5.23}$$

for $\varepsilon = (\varepsilon_1, \varepsilon_2, \varepsilon_3)$ sufficiently small.

THEOREM 5.3. *For parameters $\varepsilon = (\varepsilon_1, \varepsilon_2, \varepsilon_3)$ sufficiently small, system* (5.23) *is a generic unfolding of the cusp singularity of codimension* 3.

Proof. It has been shown in [8] that a generic unfolding, with the parameters (ν_1, ν_2, ν_3), of the codimension 3 cusp singularity is C^∞ equivalent to

$$\begin{cases} \dot{x} &= y, \\ \dot{y} &= \nu_1 + x^2 + y[\nu_2 + \nu_3 x + x^3 + O(x^4)] + y^2 Q(x,y). \end{cases} \tag{5.24}$$

Using the method and results of [8, 10], we will show that system (5.23), with parameters $(\epsilon_1, \epsilon_2, \epsilon_3)$, is also a generic unfolding of the codimension 3 singularity by showing that there exist smooth coordinate changes which take (5.23) into (5.24) with

$$\frac{D(\nu_1, \nu_2, \nu_3)}{D(\varepsilon_1, \varepsilon_2, \varepsilon_3)}\Big|_{\varepsilon=(0,0,0)} \neq 0.$$

System (5.23) has a cusp point at $(\frac{1}{\sqrt{a}}, \frac{r}{2m})$ if $\varepsilon = (0,0,0)$. Applying the translation

$$\bar{x} = x - \frac{1}{\sqrt{a}}, \qquad \bar{y} = y - \frac{r}{2m},$$

and expanding system (5.23) in the power series about the origin, we have

$$\begin{cases} \dot{\bar{x}} = \overline{L}_{11}(\bar{x}) + \bar{y}\overline{L}_{12}(\bar{x}) + \bar{y}^2 Q_{10}(\bar{x},\bar{y}), \\ \dot{\bar{y}} = \overline{L}_{21}(\bar{x}) + \bar{y}\overline{L}_{22}(\bar{x}) + \bar{y}^2 Q_{20}(\bar{x},\bar{y}), \end{cases} \tag{5.25}$$

where $Q_{i0}(0,0) = 0$ $(i = 1, 2)$ and

$$\begin{aligned}
\overline{L}_{11}(\bar{x}) &= \frac{r(a\varepsilon_3 + 2\sqrt{a}\varepsilon_1\varepsilon_3 + 2\varepsilon_1)}{2\sqrt{a}(\sqrt{a}+\varepsilon_1)(2+\sqrt{a}\varepsilon_3)} + \frac{\sqrt{a}r\varepsilon_3}{2+\sqrt{a}\varepsilon_3}\bar{x} \\
&\quad + \frac{\sqrt{a}r(a^{3/2}\varepsilon_3 - 4\sqrt{a}\varepsilon_1 - 2\varepsilon_1^2)}{2(\sqrt{a}+\varepsilon_1)^2(2+\sqrt{a}\varepsilon_3)}\bar{x}^2 - \frac{a^2 r}{2(\sqrt{a}+\varepsilon_1)^2}\bar{x}^3 + O(\bar{x}^4), \\
\overline{L}_{12}(\bar{x}) &= -\frac{m}{\sqrt{a}+\varepsilon_1} + \frac{a^{3/2}m}{(\sqrt{a}+\varepsilon_1)^2}\bar{x}^2 - \frac{a^2 m}{(\sqrt{a}+\varepsilon_1)^2}\bar{x}^3 + O(\bar{x}^4), \\
\overline{L}_{21}(\bar{x}) &= \frac{r}{2m}\left(-\varepsilon_2 - \frac{mc\varepsilon_1}{\sqrt{a}(\sqrt{a}+\varepsilon_1)}\right) - \frac{a^{3/2}cr}{2(\sqrt{a}+\varepsilon_1)^2}\bar{x}^2 + \frac{a^2 cr}{2(\sqrt{a}+\varepsilon_1)^2}\bar{x}^3 + O(\bar{x}^4), \\
\overline{L}_{22}(\bar{x}) &= -\varepsilon_2 - \frac{mc\varepsilon_1}{\sqrt{a}(\sqrt{a}+\varepsilon_1)} - \frac{a^{3/2}cm}{(\sqrt{a}+\varepsilon_1)^2}\bar{x}^2 + \frac{a^2 mc}{(\sqrt{a}+\varepsilon_1)^2}\bar{x}^3 + O(\bar{x}^4).
\end{aligned}$$

By the transformation

$$\tilde{x} = \bar{x}, \qquad \tilde{y} = \overline{L}_{11}(\bar{x}) + \bar{y}\overline{L}_{12}(\bar{x}) + \bar{y}^2 Q_{10}(\bar{x},\bar{y}), \tag{5.26}$$

system (5.25) is C^∞ equivalent to

$$\begin{cases} \dot{\tilde{x}} = \tilde{y}, \\ \dot{\tilde{y}} = L_{21}(\tilde{x}) + \tilde{y}L_{22}(\tilde{x}) + \tilde{y}^2 Q_2(\tilde{x},\tilde{y}), \end{cases} \tag{5.27}$$

where

$$\begin{aligned}
L_{21}(\tilde{x}) &= \frac{r}{2a\sqrt{a}}\left((cm\varepsilon_1 - a\varepsilon_2) + O(|\varepsilon|^2)\right) + \frac{r}{2\sqrt{a}}\left(\varepsilon_3(cm\varepsilon_1 - a\varepsilon_2) + O(|\varepsilon|^3)\right)\tilde{x} \\
&\quad + \left[\frac{1}{2}mcr + O(|\varepsilon|)\right]\tilde{x}^2 - \left[\frac{1}{2}\sqrt{a}cmr + O(|\varepsilon|)\right]\tilde{x}^3 + O(\tilde{x}^4), \\
L_{22}(\tilde{x}) &= \frac{1}{2a}\left(-2mc\varepsilon_1 + 2a\varepsilon_2 + a\sqrt{a}r\varepsilon_3 + O(|\varepsilon|^2)\right) + r\left(-\varepsilon_1 + a\varepsilon_3 + O(|\varepsilon|^2)\right)\tilde{x} \\
&\quad - \frac{\sqrt{a}}{2}\left(3\sqrt{a}r + 2cm + O(|\varepsilon|)\right)\tilde{x}^2 + \left(acm + O(|\varepsilon|)\right)\tilde{x}^3 + O(\tilde{x}^4).
\end{aligned}$$

Note that for ε sufficiently small, $c_{20}(\varepsilon) := \frac{\partial^2 L_{21}}{\partial \tilde{x}^2}(0) = \frac{1}{2}mcr + O(|\varepsilon|) \neq 0$. Thus, in the proof of Theorem 5.3, we can reduce $L_{21}(\tilde{x})$ to a quadratic polynomial without linear terms. First, by rescaling $\tilde{y}$ and time t using

$$\hat{x} = \tilde{x}, \quad \hat{y} = \tilde{y}\sqrt{c_{20}(\varepsilon)}, \quad t = \frac{1}{\sqrt{c_{20}(\varepsilon)}}\tilde{t},$$

the coefficient of $\tilde{x}^2$ in $L_{21}(\tilde{x})$ becomes $1 + O(|\varepsilon|)$, and the coefficient of $\tilde{x}$ becomes

$$c_{10}(\varepsilon) = \frac{1}{\sqrt{a}mc}\left[\varepsilon_3(mc\varepsilon_1 - a\varepsilon_2) + O(|\varepsilon|^3)\right].$$

Then the translation

$$\hat{x} = c_{10}(\varepsilon) + O(|\varepsilon|^3) + \hat{u}, \quad \hat{y} = \hat{v}, \tag{5.28}$$

brings system (5.27) to

$$\begin{cases} \dot{\hat{u}} = \hat{v}, \\ \dot{\hat{v}} = \hat{L}_{21}(\hat{u},\varepsilon) + \hat{v}\hat{L}_{22}(\hat{u},\varepsilon) + \hat{v}^2\hat{Q}_2(\hat{u},\hat{v}), \end{cases} \tag{5.29}$$

where

$$\begin{aligned} &\hat{L}_{21}(\hat{u},\varepsilon) = \hat{\nu}_1(\varepsilon) + \hat{u}^2 + O(\hat{u}^3), \\ &\hat{L}_{22}(\hat{u},\varepsilon) = \hat{\xi}_0(\varepsilon) + \hat{\xi}_1(\varepsilon)\hat{u} + \hat{\xi}_2(\varepsilon)\hat{u}^2 - \hat{\xi}_3(\varepsilon)\hat{u}^3 + O(\hat{u}^4), \end{aligned}$$

and

$$\begin{aligned} \hat{\nu}_1 &= \frac{1}{a\sqrt{a}mc}\left(mc\varepsilon_1 - a\varepsilon_2 + O(|\varepsilon|^2)\right), \\ \hat{\xi}_0 &= \frac{1}{a\sqrt{2rmc}}\left[-2cm\varepsilon_1 + 2a\varepsilon_2 + a\sqrt{a}r\varepsilon_3 + O(|\varepsilon|^2)\right], \\ \hat{\xi}_1 &= \sqrt{\frac{2r}{mc}}\left(-\varepsilon_1 + a\varepsilon_3 + O(|\varepsilon|^2)\right), \\ \hat{\xi}_2 &= \sqrt{\frac{a}{2mcr}}\left(3\sqrt{a}r + 2mc + O(|\varepsilon|)\right), \\ \hat{\xi}_3 &= 3a\sqrt{\frac{ar}{2mc}} + O(|\varepsilon|). \end{aligned}$$

Consider the corresponding 1-form of (5.29):

$$\hat{v}d\hat{v} - \left[\hat{L}_{21}(\hat{u},\varepsilon) + \hat{v}\hat{L}_{22}(\hat{u},\varepsilon) + \hat{v}^2\hat{Q}_2(\hat{u},\hat{v})\right]d\hat{u} = 0. \tag{5.30}$$

Now we reduce $\hat{L}_{21}(\hat{u},\varepsilon)$ to $\tilde{\nu}_1 + \hat{u}^2$. Denote

$$\hat{L}(\hat{u},\varepsilon) = \hat{\nu}_1(\varepsilon)\hat{u} + \frac{1}{3}\hat{u}^3 + O(\hat{u}^4).$$

Using the Malgrange preparation theorem [5], we find a coordinate change of the form

$$\hat{u} = \Phi(\tilde{u},\varepsilon) = \phi(\varepsilon)\tilde{u} + O(\tilde{u}^2), \tag{5.31}$$

where $\phi(0) = 1$ such that

$$\hat{L}(\Phi(\tilde{u},\varepsilon),\varepsilon) = \tilde{\nu}_1(\varepsilon)\tilde{u} + \frac{1}{3}\tilde{u}^3,$$

and $\tilde{\nu}_1(\varepsilon) = \hat{\nu}_1(\varepsilon) + O(|\varepsilon|^2)$. Performing this coordinate change to family (5.30) and writing $\hat{v} = \tilde{v}$, we obtain

$$\tilde{v}d\tilde{v} - \left[\tilde{L}_{21}(\tilde{u},\varepsilon) + \tilde{v}\tilde{L}_{22}(\tilde{u},\varepsilon) + \tilde{v}^2\tilde{Q}_2(\tilde{u},\tilde{v})\right]d\tilde{u} = 0, \tag{5.32}$$

where

$$\tilde{L}_{21}(\tilde{u},\varepsilon) = \hat{\nu}_1(\varepsilon) + \tilde{u}^2,$$
$$\tilde{L}_{22}(\tilde{u},\varepsilon) = \tilde{\xi}_0(\varepsilon) + \tilde{\xi}_1(\varepsilon)\tilde{u} + \tilde{\xi}_2(\varepsilon)\tilde{u}^2 - \tilde{\xi}_3(\varepsilon)\tilde{u}^3 + O(\tilde{u}^4),$$

and $\tilde{\xi}_i(\varepsilon) = \xi_i(\varepsilon) + O(|\varepsilon|^2)$ $(i = 0, 1, 2, 3)$.

Then similar to step (2) in the proof of Theorem 5.2, using $S(\tilde{u},\tilde{v}) = \frac{1}{2}\tilde{v}^2 - \frac{1}{3}\tilde{u}^3$, and a near-identity transformation of the form

$$\begin{aligned} \tilde{u} &= u, \\ \tilde{v} &= v + \tfrac{1}{3}\xi_2(\varepsilon)v^2 + O(v^3), \end{aligned} \tag{5.33}$$

we eliminate the term $\tilde{v}\tilde{u}^2$ in (5.32), and system (5.23) is C^∞ equivalent to

$$\begin{cases} \dot{u} = v, \\ \dot{v} = \tilde{\nu}_1 + u^2 + v[\tilde{\xi}_0 + \tilde{\xi}_2 u - \tilde{\xi}_3 u^3 + O(u^4)] + v^2 Q(u,v). \end{cases} \tag{5.34}$$

For ε sufficiently small, the dominant terms in $\tilde{\nu}_1$ and $\tilde{\xi}_i$ $(i = 0, 2, 3)$ remain unchanged. Hence, we keep the previous notation.

For system (5.34), $\tilde{\xi}_3(0) = 3a\sqrt{\frac{ar}{2mc}} > 0$. Thus the rescaling

$$u = \frac{1}{\tilde{\xi}_3^{\frac{2}{5}}} U, \qquad v = \frac{1}{\tilde{\xi}_3^{\frac{3}{5}}} V, \qquad \tilde{t} = \xi_3^{\frac{1}{5}}\tau$$

shows that system (5.23) is equivalent to

$$\begin{cases} \dot{x} = y, \\ \dot{y} = \tilde{\tilde{\nu}}_1 + x^2 + y[\tilde{\tilde{\xi}}_0 + \tilde{\tilde{\xi}}_1 x - x^3 + O(x^4)] + y^2 Q(x,y), \end{cases} \tag{5.35}$$

where we have replaced (U, V, τ) with (x, y, t).

To apply the results from [8], we change the sign of the term $x^3 y$ in the second equation of (5.36) to positive by the transformation

$$\left(x,\ y,\ t,\ \tilde{\tilde{\nu}}_1,\ \tilde{\tilde{\xi}}_0,\ \tilde{\tilde{\xi}}_1\right) \longrightarrow \left(x,\ -y,\ -t,\ \tilde{\tilde{\nu}}_1,\ -\tilde{\tilde{\xi}}_0,\ -\tilde{\tilde{\xi}}_1\right);$$

then system (5.35) becomes

$$\begin{cases} \dot{x} = y, \\ \dot{y} = \tilde{\tilde{\nu}}_1 + x^2 + y[\tilde{\tilde{\xi}}_0 + \tilde{\tilde{\xi}}_1 x + x^3 + O(x^4)] + y^2 Q(x,y). \end{cases} \tag{5.36}$$

To simplify the expressions for the parameters, we make the rescaling

$$\begin{cases} \nu_1 = \sqrt[5]{\dfrac{81r^2}{4a^4(mc)^{12}}}\,\tilde{\tilde{\nu}}_1, \\ \nu_2 = \sqrt[10]{\dfrac{9}{64a^7r^4(mc)^6}}\,\tilde{\tilde{\xi}}_0, \\ \nu_3 = \sqrt[10]{\dfrac{64r^4}{9a^3(mc)^2}}\,\tilde{\tilde{\xi}}_1. \end{cases} \tag{5.37}$$

This yields (5.24), where using Maple, we obtain

$$
\begin{cases}
\nu_1 = mc\sqrt{a}(mc\varepsilon_1 + a\varepsilon_2) \\
\qquad + 2(mc)^2\varepsilon_1^2 + 4amc\varepsilon_1\varepsilon_2 + a^2\varepsilon_2^2 + O(|(\varepsilon_1,\varepsilon_2)|^3), \\
\nu_2 = -2mc\varepsilon_1 - 2a\varepsilon_2 + a\sqrt{a}r\varepsilon_3 \\
\qquad + \frac{1}{4}\left[-4(mc)^2\varepsilon_1^2 - 3a^2\sqrt{a}mcr\varepsilon_3^2 - 4a^2\varepsilon_2^2 - 16amc\varepsilon_1\varepsilon_2\right. \\
\qquad \left. + 2amc(3\sqrt{a}r + mc)\varepsilon_1\varepsilon_3 + 2a^2(\sqrt{a}r + mc)\varepsilon_2\varepsilon_3\right] + O(|\varepsilon|^3), \\
\nu_3 = -\varepsilon_1 + a\varepsilon_3 + \frac{1}{4}\left[-2rmc\varepsilon_1^2 - 3a^2mcr\varepsilon_3^2 - 2ar\varepsilon_1\varepsilon_2\right. \\
\qquad \left. + \sqrt{a}mc(11\sqrt{a}r + 4mc)\varepsilon_1\varepsilon_3 + 4\sqrt{a}(2\sqrt{a}r + mc)\varepsilon_2\varepsilon_3\right] + O(|\varepsilon|^3),
\end{cases}
\tag{5.38}
$$

and

$$
\frac{D(\nu_1,\nu_2,\nu_3)}{D(\varepsilon_1,\varepsilon_2,\varepsilon_3)}\Big|_{\varepsilon=(0,0,0)} = -a^3rmc \neq 0.
$$

So system (5.23) with parameters $\varepsilon = (\varepsilon_1, \varepsilon_2, \varepsilon_3)$ is a generic family unfolding the codimension 3 cusp singularity. □

By Theorem 5.3, system (5.23) is a generic family unfolding the cusp singularity of codimension 3. So by the main theorem in [8], system (5.23) has the same bifurcation set with respect to ε_3 as (5.24) has with respect to ν, at least up to a homeomorphism in the parameter space. This bifurcation set is a cone with vertex at the origin of the parameter space.

If $\nu_1 > 0$, system (5.24) obviously has no equilibria. In a neighborhood of the origin, $\nu_1 = 0$ is a saddle-node bifurcation plane. Crossing the plane in the direction of decreasing ν_1, two equilibria are created: a saddle and an antisaddle (node or focus). Correspondingly, from (5.38), there is a surface in the parameter space $(\varepsilon_1, \varepsilon_2, \varepsilon_3)$ defined by $\nu_1(\varepsilon_1, \varepsilon_2) = 0$:

$$
\varepsilon_2 = -\frac{mc}{a}\varepsilon_1 + \frac{mc}{a\sqrt{a}}\varepsilon_1^2 + O(\varepsilon_1^3). \tag{5.39}
$$

Along this surface, system (5.23) has a saddle-node bifurcation. Substituting (5.22) into (2.2), it follows that the exact saddle-node bifurcation surface is given by

$$
\Sigma_{SN}: \quad \varepsilon_2 = -\frac{mc}{\sqrt{a}}\frac{\varepsilon_1}{\sqrt{a}+\varepsilon_1}, \tag{5.40}
$$

which is consistent with (5.39). To the right of the surface Σ_{SN}, system (5.23) has no equilibria, thus all the bifurcation surfaces are located to the left of Σ_{SN}. Since up to a homeomorphism in the parameter space, each bifurcation surface is a cone with vertex at the origin; they can best be visualized by drawing their trace on the sphere

$$
S = \left\{(\nu_1,\nu_2,\nu_3)\Big|\nu_1 < 0, \nu_1^2 + \nu_2^2 + \nu_3^2 = \varepsilon_0, \varepsilon_0 > 0 \quad \text{sufficiently small}\right\}
$$

to the left of the surface Σ_{SN}.

As in Figure 5.2, let $\Gamma = S \cap \Sigma_{SN}$ be the intersection of the "half" sphere S with Σ_{SN}. Then along Γ, except for the two points b_1 and b_2, there is a saddle-node bifurcation. The next result follows from the bifurcation diagram given in Figure 3 of [8].

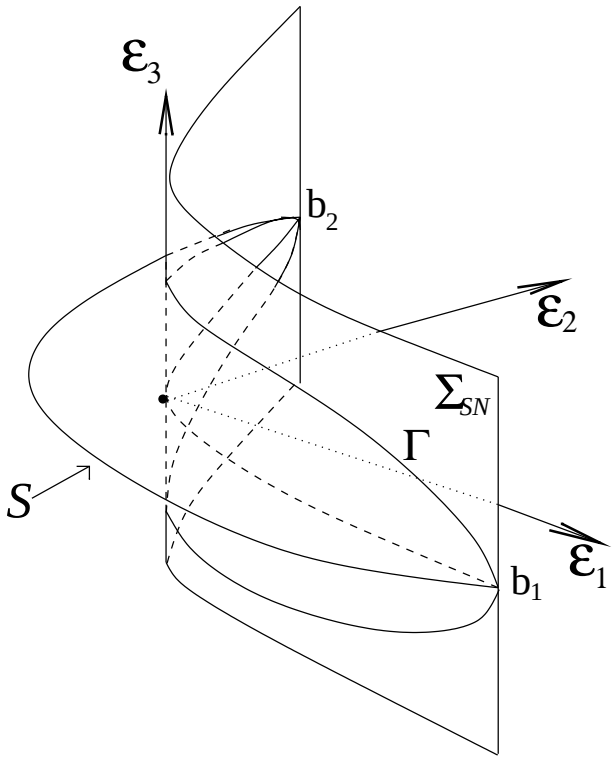


FIG. 5.2. Γ, *the intersection curve of the saddle-node bifurcation surface* Σ_{SN} *with the half sphere* S.

THEOREM 5.4. *For system* (5.24)*, using* $\varepsilon = (\varepsilon_1, \varepsilon_2, \varepsilon_3)$ *as parameters, the bifurcation diagram on* S *is given in Figure* 5.3.

On S, there are three bifurcation curves as shown in Figure 5.3:

- a curve H of Hopf bifurcations,
- a curve H_{om} of homoclinic bifurcations, and
- a curve SN_{lc} of saddle-node bifurcations of limit cycles.

As shown in Figure 5.3, the curve SN_{lc} joins a point h_2 on H to a point c_2 on H_{om}, and SN_{lc} is tangent to H at h_2 and tangent to H_{om} at c_2. The curves H and H_{om} have first order contact with the boundary of S at the points b_1 and b_2. In the neighborhood of b_1 and b_2, system (5.24) is an unfolding of the cusp singularity of codimension 2. This corresponds to the bifurcations along $K = \frac{2}{\sqrt{a}}$ with (d, b) in the neighborhood of $(\frac{mc}{\sqrt{a}}, -\sqrt{a})$. If $b > -\sqrt{a}$, the cusp singularity of codimension 2 is at the right hump, while if $b < -\sqrt{a}$, it is at the left hump.

Along the arc b_1h_2 of the curve H, a supercritical Hopf bifurcation occurs with a stable limit cycle appearing when the arc b_1h_2 is crossed from right to left. Along the arc h_2b_2 of the curve H, a subcritical Hopf bifurcation occurs with an unstable limit cycle appearing when the arc h_2b_2 is crossed from left to right. The point h_2 is a degenerate Hopf bifurcation point, i.e., a Hopf bifurcation point of codimension 2. The point h_2 in Figure 5.3 corresponds to the degenerate Hopf curve DH in Figures 4.2 and 6.1, which represents a three dimensional curve of codimension 2 Hopf bifurcations, projected onto the bK plane.

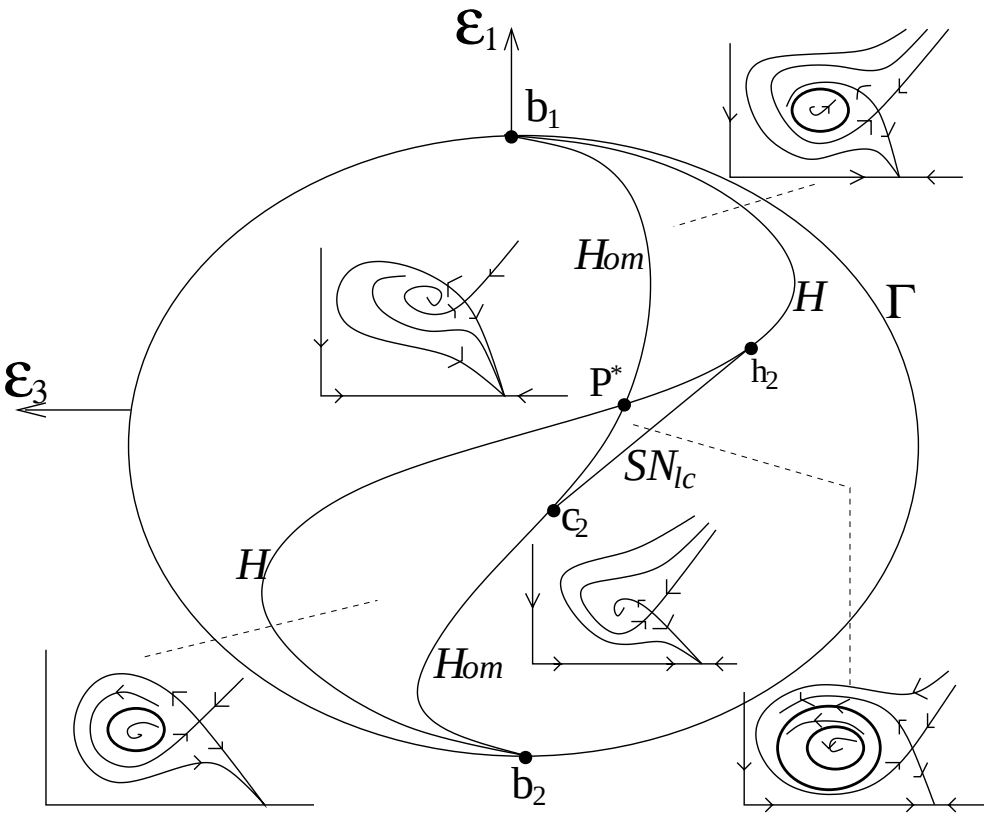


FIG. 5.3. *Bifurcation diagram for system* (5.24) *on* S.

Along the curve H_{om}, except at the point c_2, a homoclinic bifurcation of codimension 1 occurs. When the arc b_1c_2 of H_{om} is crossed from left to right, the two separatrices of the saddle point coincide and a stable limit cycle appears. The same phenomenon gives rise to an unstable limit cycle when the arc c_2b_2 of H_{om} is crossed from right to left. The point c_2 corresponds to a homoclinic bifurcation of codimension 2 (see [2, 23, 27] for references).

The curves H and H_{om} intersect transversally at a unique point P^* representing a parameter value of simultaneous Hopf and homoclinic bifurcation. The point P^* in Figure 5.3 corresponds to the curve HH in Figure 6.1, a projection to the bK plane of the three dimensional curve along which Hopf and homoclinic bifurcations occur simultaneously.

For parameter values in the curved triangle $P^*h_2c_2$, there exist exactly two limit cycles; the inner one is unstable and the outer one is stable. These two limit cycles coalesce in a generic way in a saddle-node bifurcation of limit cycles when the curve SN_{lc} is crossed from left to right. On the arc SN_{lc} itself, there exists a unique semistable limit cycle.

6. Global dynamics. In the previous section we proved that a degenerate Bogdanov–Takens bifurcation of codimension 3 occurs when (5.9) is satisfied. Therefore, when the parameters (1.5) are varied in a neighborhood of (5.9), a degenerate homoclinic bifurcation, degenerate Hopf bifurcation, and saddle-node bifurcation of

limit cycles must occur. In this section, using the information obtained from the analysis of the codimension 3 Bogdanov–Takens bifurcation and the geometry of system (1.3), we study the role of each of the parameters (1.5) in the global bifurcations of system (1.3). In order to do this, we determine when certain phenomena occur simultaneously. We then combine this information with the local dynamics studied in section 4 to determine the sequence of bifurcations in the different regions of parameter space as $\hat{d} = \frac{d}{mc}$ is varied.

6.1. Periodic orbits and homoclinic loops. It was proved in [30] that if system (1.3) has a limit cycle in the positive cone, it has to surround a hump of the prey isocline. In this subsection we prove several theorems which help determine whether system (1.3) has periodic orbits or homoclinic loops. Throughout this section, by periodic orbits we mean nontrivial periodic orbits.

THEOREM 6.1. *For system* (1.3)*, the horizontal line* $y = F(\lambda)$ *can intersect the prey isocline* $y = F(x)$ *at most three points in the first quadrant. If there is a periodic orbit, it lies entirely to the left of* E_K *and* E_μ *(if* E_μ *exists). Furthermore,*

- *if* $\lambda \neq H_m$, H_M*, the periodic orbit must surround* E_λ *and another intersection point* $(x^*, F(x^*))$*, where* $F(x^*) = F(\lambda)$*;*
- *if* $\lambda = H_m$ *or* H_M*, the periodic orbit must surround* E_λ*, the tangent point of* $y = F(\lambda)$ *with* $y = F(x)$*.*

Proof. From (1.4) it is clear that $y = F(x)$ is a cubic polynomial and hence $y = F(\lambda)$ can intersect $y = F(x)$ at at most three points.

If $\hat{d} \geq \hat{d}_M$, then $\dot{y} \leq 0$ along all orbits and hence system (1.3) has no periodic orbits. The only case remaining is $0 < \hat{d} < \hat{d}_M$.

Using standard phase plane arguments, it is clear that any periodic orbit must lie entirely to the left of E_K and E_μ (if E_μ exists). By a consequence of the Poincaré–Bendixson theorem, a periodic orbit in the plane must surround an equilibrium. By phase plane analysis, E_λ is the only candidate.

Consider an auxiliary function of the form $L(x, y) = M(x) + N(y)$, where $M(x)$ and $N(y)$ are continuous and differentiable and satisfy the following equations, respectively:

$$\begin{aligned} &p(x)M'(x) = d - cp(x), \qquad M(\lambda) = 0,\ x > 0, \\ &yN'(y) = F(\lambda) - y, \qquad N(F(\lambda)) = 0,\ y > 0. \end{aligned} \tag{6.1}$$

Solving these equations, we obtain a function $L(x, y)$ defined in the first quadrant. Along the trajectories of system (1.3) we have

$$\frac{d}{dt}L(x, y) = (d - cp(x))[F(x) - F(\lambda)]. \tag{6.2}$$

For $\hat{d} \in (0, \hat{d}_M)$, (6.2) can be rewritten as

$$\dot{L}(t) = \frac{d}{dt}L(x(t), y(t)) = \frac{mca\hat{d}}{ax^2 + bx + 1}(x - \lambda)(x - \mu)[F(x) - F(\lambda)]. \tag{6.3}$$

Denote

$$\bar{\mu} = \min\{\mu, K\}. \tag{6.4}$$

If there is a closed orbit, $\dot{L}$ must undergo a change of sign along this orbit. If $\lambda = H_m$ or H_M, then $\dot{L}$ changes sign when $x = \lambda$, and if $\lambda \neq H_m$ or H_M, $\dot{L}$ can change

sign only when $x = \mu$ or $x = x^* \neq \lambda$, where $F(x^*) = F(\lambda)$. Therefore, $\lambda = H_m$ or H_M, and the periodic orbit surrounds E_λ, or there exists a $x^* \in (0, \lambda) \cup (\lambda, \bar{\mu})$ such that $F(x^*) = F(\lambda)$, and $(x^*, F(x^*))$ sits inside the closed orbit. □

COROLLARY 6.2. *For system* (1.3)*, if E_λ is the only intersection point of the horizontal line $y = F(\lambda)$ with the prey isocline $y = F(x)$, and $\lambda \neq H_M$, then there are neither periodic orbits nor homoclinic loops.*

THEOREM 6.3. *Assume $\hat{d} \in (0, \hat{d}_M)$. Neither periodic orbits nor homoclinic loops exist if either*

1. *$F'(x) \leq 0$ for all $x \in (\lambda, \bar{\mu})$, or*
2. *$F'(\lambda) \geq 0$ and $F'(\mu) \geq 0$.*

Proof. To prove the theorem, we make a change of variables and rescale time by setting

$$u = \ln x, \quad v = \ln y, \quad \tau = \int_0^t \frac{1}{ax^2 + bx + 1} dt \tag{6.5}$$

to obtain

$$\begin{cases} \dot{u} = m[F(e^u) - e^v], \\ \dot{v} = -mc\hat{g}(e^u), \end{cases} \tag{6.6}$$

where $\hat{g}(x)$ is defined in (2.2).

1. We proceed by using the Dulac criterion with the positive auxiliary function $B(v) = e^{m\beta v}$, where β is a nonnegative constant to be determined.

The divergence

$$\begin{aligned} \operatorname{div}(B(v)[\dot{u}, \dot{v}]) &= -me^{m\beta v}[-e^u F'(e^u) + \beta mc\hat{g}(e^u)] \\ &= -me^{m\beta v} R(e^u, \beta), \end{aligned} \tag{6.7}$$

where

$$\begin{aligned} R(x, \beta) &= -xF'(x) + \beta mc\hat{g}(x) \\ &= \frac{3ar}{mK}x^3 + \left(a\hat{d}mc\beta - \frac{2r(aK - b)}{mK}\right)x^2 \\ &\quad + \left[(b\hat{d} - 1)mc\beta - \frac{r(bK - 1)}{mK}\right]x + mc\beta\hat{d}. \end{aligned} \tag{6.8}$$

It follows from Theorem 6.1 that a periodic solution or a homoclinic loop must lie entirely inside the strip

$$\left\{(x, y) \middle| 0 < x < \bar{\mu}, y > 0\right\}.$$

Thus it is enough to show that there exists a $\beta_1 \geq 0$ such that

$$R(x, \beta_1) \geq 0, \quad x \in (0, \bar{\mu}). \tag{6.9}$$

Consider the cubic, $R(x, 0)$. By the hypothesis, $R(x, 0) > 0$ for $\lambda < x < \bar{\mu}$. Therefore (6.9) is satisfied with $\beta_1 = 0$, unless there is either one or two simple roots of $R(x, 0)$ inside the interval $(0, \lambda)$. We now consider this case.

Note that $\lim_{x \to \pm\infty} R(x, \beta) = \pm\infty$ for any $\beta \geq 0$, and $\hat{g}(\lambda) = \hat{g}(\mu) = 0$, $\hat{g}(x) < 0$ if $x \in (\lambda, \mu)$, and $\hat{g}(x) > 0$, otherwise. Therefore for $\beta > 0$, there is always one negative root of $R(x, \beta)$ and for $\beta > 0$ sufficiently small there are two positive roots in $(0, \lambda]$.

Since $R(\lambda,\beta) = -\lambda F'(\lambda) \geq 0$ for all $\beta \geq 0$, and since $\hat{g}(x) > 0$ for $x \in [0,\lambda)$, there exists $\beta_1 > 0$ such that $R(x,\beta_1)$ has a double root in $(0,\lambda]$, and thus $R(x,\beta_1) \geq 0$ for $x \in (0,\bar{\mu})$.

2. The argument is similar. Using the auxiliary function $B(v) = e^{-m\beta v}$, the divergence

$$\operatorname{div}(B(v)[\dot{u},\dot{v}]) = me^{-m\beta v}R(e^u,\beta), \tag{6.10}$$

where $R(x,\beta) = xF'(x) + \beta mc\hat{g}(x)$. Note that $\lim_{x\to\pm\infty} R(x,\beta) = \mp\infty$ for all $\beta \geq 0$, the hypothesis implies that $F'(x) > 0$ for $\lambda < x < \mu$, and there is always a root of $R(x,\beta)$ such that $x \geq \beta$ rather than a negative root. ∎

COROLLARY 6.4.

1. *Assume $K \neq \frac{2}{\sqrt{a}}$. There exists a $\tilde{d} \in (0,\hat{d}_M)$ such that for all $\hat{d} > \tilde{d}$, system* (1.3) *has neither periodic orbits nor homoclinic loops.*
2. *Assume $K > \frac{2}{\sqrt{a}}$. Then for all $\hat{d} > \hat{d}_c$ ($\hat{d}_c$ was defined in* (3.1)*), system* (1.3) *has neither periodic orbits nor homoclinic loops if*
   - *$(b,K) \in V_1^1$, or*
   - *$(b,K) \in V_2$ and $F(0) < F(\lambda)$.*

Proof. 1. As $\hat{d} \in (0,\hat{d}_M)$ increases, λ and μ tend to $\frac{1}{\sqrt{a}}$ monotonically from the left and right side, respectively. Hence if $K \neq \frac{2}{\sqrt{a}}$, neither the left nor the right hump is at $\frac{1}{\sqrt{a}}$, and so there exists a $\tilde{d} \in (0,\hat{d}_M)$ such that for $\hat{d} \in (\tilde{d},\hat{d}_M)$, there are no interior equilibria or both of the interior equilibria satisfy either $F'(\lambda) > 0$ and $F'(\mu) > 0$ or $F'(x) < 0$ for all $x \in (\lambda,\bar{\mu})$. By Theorem 6.3, in either case, system (1.3) has no periodic orbits nor homoclinic loops.

2. This is a direct consequence of part 3 of Lemma 3.1 and Theorem 6.1. ∎

COROLLARY 6.5. *For system* (1.3) *with parameters* (1.5)*, if $(b,K) \in V_0 \cup V_2^0$, then for any $\hat{d} > 0$, system* (1.3) *has neither periodic orbits nor homoclinic loops.*

Proof. For any $(b,K) \in V_0$, $F'(x) < 0$ for all $x > 0$, so the result follows from Theorem 6.3.

Assume $(b,K) \in V_2^0$, and $\hat{d} \in (0,\hat{d}_M)$. In this region, $0 < \frac{1}{\sqrt{a}} < H_m < H_M < K$. By Lemma 3.1, since $K < \frac{2}{\sqrt{a}}$, $F(\lambda) > F(\mu)$. If $\mu \geq H_M$, then $F(\lambda) > F(H_M)$, and there is a unique intersection of $y = F(\lambda)$ and $y = F(x)$. The result follows from Corollary 6.2. If $\frac{1}{\sqrt{a}} < \mu < H_M$, besides $(\lambda, F(\lambda))$, any other intersection of $y = F(\lambda)$ and $y = F(x)$ must have x coordinate great than μ. The result follows from Theorem 6.1. ∎

THEOREM 6.6. *Fix all parameters except $\hat{d} > 0$.*

1. *No homoclinic bifurcation can occur for $\hat{d} \in (0,\hat{d}_{\mu K})$ (where $\hat{d}_{\mu K}$ is the value of $\hat{d}$ at the transcritical bifurcation involving E_μ and E_K).*
2. *When a homoclinic bifurcation occurs,*
   - *if $F'(\mu) < 0$, then it is supercritical,*
   - *if $F'(\mu) > 0$, then it is subcritical.*

Proof. Part 1 is obvious since in this case, if E_μ does not exist, E_λ is the only equilibrium inside the positive cone and it is never a saddle. E_K cannot form a saddle loop as the x-axis is invariant. Part 2 follows from a standard result [2], since $\operatorname{tr}(V(\mu,F(\mu))) = p(\mu)F'(\mu)$ (see (2.5)). ∎

THEOREM 6.7. *Fix all parameters except $\hat{d} > 0$. For $(b,K) \in V_2$ and $K > \frac{2}{\sqrt{a}}$, there exists a $\hat{d}_l \in (\hat{d}_{\mu K},\hat{d}_M)$ such that a homoclinic loop bifurcation involving E_μ occurs when $\hat{d} = \hat{d}_l$.*

Proof. If a homoclinic loop bifurcation occurs, it involves E_μ; hence it occurs for $\hat{d} \in (\hat{d}_{\mu K}, \hat{d}_M)$. For $\hat{d} < \hat{d}_{\mu K} (< \hat{d}_-)$, E_λ is the only equilibrium in the interior of the first quadrant and it is asymptotically stable. Since solutions are bounded, there must either be no limit cycles or an even number, excluding semistable periodic orbits. For $\hat{d} > \hat{d}_M$, system (1.3) has no limit cycles. By Corollary 4.2, there is exactly one Hopf bifurcation which occurs at $(H_m, F(H_m))$ when $\hat{d} = \hat{d}_-$ and changes the parity of the limit cycles. Therefore, there must exist a $\hat{d}_l \in (\hat{d}_{\mu K}, \hat{d}_M)$ such that a homoclinic loop bifurcation occurs to compensate for this change in the number of limit cycles. ∎

6.2. Simultaneous phenomena. We will subdivide V_0, V_1, and V_2 using the curves defined below along which simultaneous phenomena occur for some $\hat{d} > 0$:

NS:	a Hopf bifurcation at E_λ with $\lambda = H_m$ and a neutral saddle at E_μ when $\mu = H_M$;
$E_{\mu K}$:	a Hopf bifurcation at E_λ with $\lambda = H_m$ or $\lambda = H_M$ and a transcritical bifurcation involving E_μ and E_K;
HH:	a Hopf bifurcation at E_λ with $\lambda = H_m$ and a homoclinic loop involving E_μ;
$Dhom$:	a homoclinic bifurcation involving E_μ when E_μ is a neutral saddle, i.e., $\mu = H_M$ (degenerate homoclinic bifurcation);
ST:	a saddle-node bifurcation of limit cycles and a transcritical bifurcation involving E_μ and E_K.

In the rest of this subsection, we will find analytic expressions for curves $E_{\mu K}$ and NS and prove that curves HH and $Dhom$ must exist in certain regions in parameter space. We will prove the existence of the curve ST in the next subsection. Information about these curves is summarized in Table 6.1 and illustrated in Figure 6.1 and Figure 6.4.

PROPOSITION 6.8. *In the bK plane (Figure* 6.1*), along the curve*

$$NS : K = -\frac{2}{b}, \qquad -\sqrt{a} < b < 0, \tag{6.11}$$

there exists a unique $\hat{d} \in (0, \hat{d}_M)$ such that a Hopf bifurcation at $(H_m, F(H_m))$ and a neutral saddle at $(H_M, F(H_M))$ occur simultaneously. To the left of NS the neutral saddle at $(H_M, F(H_M))$ occurs before the Hopf bifurcation at $(H_m, F(H_m))$. To the right of NS this ordering is reversed.

Proof. Recall from Theorem 4.1 that a Hopf bifurcation occurs at $(H_m, F(H_m))$ where $\hat{d} = \hat{d}_-$ as given in (4.3). In a similar manner to the proof of Theorem 4.1, it can be shown that for $\frac{2}{\sqrt{a}} < K < \frac{1}{b}$, a neutral saddle occurs at $(H_M, F(H_M))$ when $\hat{d} = \hat{d}_+$, as given in (4.3). We thus obtain

$$\hat{d}_+ - \hat{d}_- = \frac{(2 + bK)\sqrt{\Delta_1}}{(4a - b^2)(aK^2 + bK + 1)}. \tag{6.12}$$

Clearly, when $K = -\frac{2}{b}$, $\hat{d}_+ = \hat{d}_- = \frac{3K}{2(aK^2-1)}$, thus the Hopf bifurcation and the neutral saddle occur simultaneously at this value of $\hat{d}$. Note that $b^2 - 4a < 0$. Hence when $-2\sqrt{a} < b < -\frac{2}{K}$, $\hat{d}_+ < \hat{d}_-$, the neutral saddle occurs before the Hopf bifurcation, and when $-\frac{2}{\sqrt{a}} < b < \frac{1}{K}$, the order is reversed. To complete the proof we note that $\Delta_1 = 0$ only when $\lambda = H_m = H_M = H_I = \mu$, which corresponds to the point P. ∎

Recall that a transcritical bifurcation involving E_μ and E_K occurs when $K > \frac{1}{\sqrt{a}}$ and $\hat{d} = \hat{d}_{\mu K} = \hat{p}(K)$ (see (2.1)).

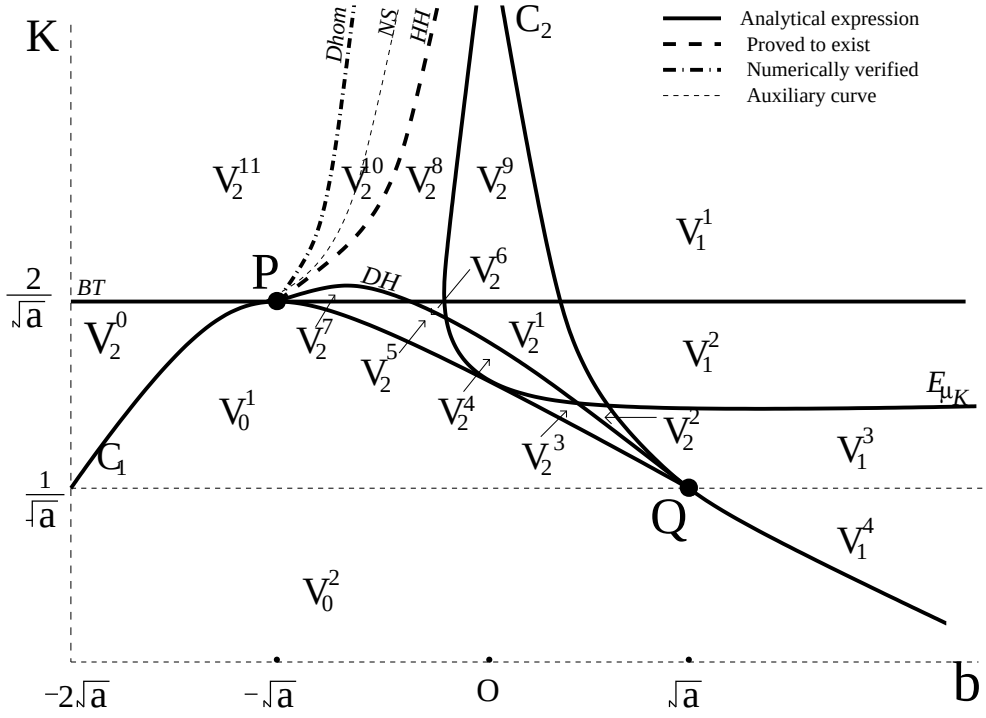


FIG. 6.1. *The subregions of the bK plane determined by the degenerate phenomena of Table* 6.1.

PROPOSITION 6.9. *In the bK plane (Figure* 6.1*), along the curve*

$$E_{\mu K} : b = \frac{3 - aK^2}{K(aK^2 - 2)}, \qquad K > \sqrt{\frac{2}{a}}, \tag{6.13}$$

there exists a unique $\hat{d} \in (0, \hat{d}_M)$ *such that the Hopf bifurcation and the* E_μ *and* E_K *transcritical bifurcation occur simultaneously. The curve* $E_{\mu K}$ *is tangent to* C_1 *at* $(0, \sqrt{3/a})$*. Along this curve, if* $b < 0$*, the Hopf bifurcation occurs at the left hump, and if* $b > 0$*, the Hopf bifurcation occurs at the right hump.*

1. *For* $(b, K) \in V_1 \setminus V_1^1$*,*
   (a) *if* (b, K) *is above* $E_{\mu K}$*, the transcritical bifurcation occurs before the Hopf bifurcation at* E_λ *with* $\lambda = H_M$*;*
   (b) *if* (b, K) *is below* $E_{\mu K}$*, the Hopf bifurcation at* E_λ *with* $\lambda = H_M$ *occurs before the transcritical bifurcation.*
2. *For* $(b, K) \in V_2 \setminus V_2^0$*,*
   (a) *if* (b, K) *is to the left of* $E_{\mu K}$*, the transcritical bifurcation occurs before the Hopf bifurcation at* E_λ *with* $\lambda = H_m$*;*
   (b) *if* (b, K) *is to the right of* $E_{\mu K}$*, the transcritical bifurcation occurs after the Hopf bifurcation at* E_λ *with* $\lambda = H_m$ *and, if* $K < \frac{2}{\sqrt{a}}$*, it occurs before the Hopf bifurcation at* E_λ *with* $\lambda = H_M$*;*
   (c) *if* (b, K) *is below* $E_{\mu K}$*, the Hopf bifurcations at* E_λ *with* $\lambda = H_m$ *and* $\lambda = H_M$ *occur before the transcritical bifurcation.*

Proof. Let $K = \mu$ and set $H_m = \lambda$ (see (2.4) and (3.7)),

$$\begin{cases} K = \dfrac{1 - b\hat{d} + \sqrt{\Delta_0}}{2a\hat{d}}, \\ \dfrac{aK - b - \sqrt{\Delta_1}}{3a} = \dfrac{1 - b\hat{d} - \sqrt{\Delta_0}}{2a\hat{d}}. \end{cases} \tag{6.14}$$

Eliminating $\hat{d}$ from (6.14), we obtain (6.13).

Eliminating $\hat{d}$ from $\mu = K$ and $H_M = \lambda$ yields (6.13).

From Theorem 4.4 it follows that the branch of $E_{\mu K}$ with $b < 0$ corresponds to the Hopf bifurcation at $(H_m, F(H_m))$, whereas the branch with $b > 0$ corresponds to the Hopf bifurcation at $(H_M, F(H_M))$. The ordering of the Hopf bifurcation at $(H_M, F(H_M))$ and the transcritical bifurcation follows from part 1 of Proposition 3.3. As K increases across $E_{\mu K}$ with $b > 0$, the right hump moves to the right. Thus in the region above $E_{\mu K}$, the transcritical bifurcation must occur before the Hopf bifurcation at $(H_M, F(H_M))$, and in the region below, the ordering must be reversed.

To the left of $E_{\mu K}$, it follows from Proposition 6.8 that a Hopf bifurcation can occur at the same time as a neutral saddle. Thus in this region the transcritical bifurcation must occur before the Hopf bifurcation at $(H_m, F(H_m))$. To the right of $E_{\mu K}$, notice that on C_2 above Q, $H_m = 0$. When $\hat{d} = \hat{d}_{\mu K}$, $\lambda > H_m$. Fix $\hat{d}$. For K slightly below C_2, $\lambda > H_m$, but $K < \mu$. Therefore the Hopf bifurcation at $(H_m, F(H_m))$ occurs before the transcritical bifurcation in this region. ∎

For system (1.3), if $b = 0$, it was proved in [28] that the Hopf bifurcation and homoclinic bifurcation cannot occur simultaneously. However, for $b < 0$ they can happen simultaneously along the curve HH. The bifurcation analysis of the codimension 3 cusp singularity in section 5 (Theorem 5.3) indicates that in a neighborhood of the point P, there exist curves DH, HH, and $Dhom$ emanating from P:

$$\begin{aligned} DH &: K = K_{DH}(b), \\ HH &: K = K_{HH}(b), \\ Dhom &: K = K_{Dhom}(b), \end{aligned} \tag{6.15}$$

HH and $Dhom$ are tangent to DH at P, and $Dhom$ is to the left of HH, which is to the left of DH. Recall that an analytic expression for DH was derived (see (4.11)). For any (b, K) along HH in a neighborhood of P, there exists a $\hat{d}_l$ such that when $\hat{d} = \hat{d}_l$, the system undergoes both a Hopf bifurcation and a homoclinic loop bifurcation. For any (b, K) along $Dhom$ in a neighborhood of P, there exists a $\hat{d}_{Dhom} \in (0, \hat{d}_M)$ such that when $\hat{d} = \hat{d}_{Dhom}$, system (1.3) undergoes a degenerate (codimension 2) homoclinic loop bifurcation. Also in a neighborhood of P, for any (b, K) in the region between DH and $Dhom$, there exists a $\hat{d}_{sn} \in (0, \hat{d}_M)$ such that when $\hat{d} = \hat{d}_{sn}$, system (1.3) undergoes a saddle-node bifurcation of limit cycles. In Figure 6.1 and Figure 6.2, we use a dot-dash line to illustrate $Dhom$ and a dashed line to illustrate HH. The global extension of the curve $Dhom$ has been observed numerically using XPPAUT. In the following we prove the position of NS with respect to $Dhom$ and HH.

LEMMA 6.10. *Fix all parameters as in* (1.5) *except* $\hat{d} > 0$. *For* $K < \frac{2}{\sqrt{a}}$ *and* $K > \frac{2}{\sqrt{a}}$ *to the right of* NS, *if a homoclinic bifurcation occurs, it is supercritical.*

Proof. By Proposition 3.2, if $K < \frac{2}{\sqrt{a}}$, $H_M < \frac{1}{\sqrt{a}} \le \mu$. If $K > \frac{2}{\sqrt{a}}$ to the right of NS, then $H_M < \mu$. Hence in both cases, $F'(\mu) < 0$. It follows from part 2 of Theorem 6.6 that if a homoclinic bifurcation occurs in either of these cases, it is supercritical. ∎

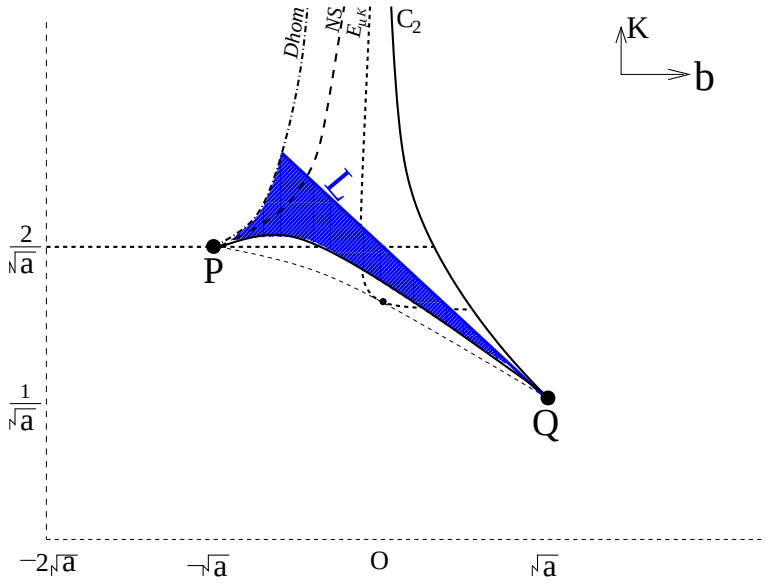


FIG. 6.2. *V^*, subregion of V_2, bounded by Dhom, DH, and C_2. For $(b, K) \in V_{sn}$, the shaded region, there exists a $\hat{d}_{sn} > 0$ such that when $\hat{d} = \hat{d}_{sn}$, system (1.3) undergoes a saddle-node bifurcation of limit cycles.*

COROLLARY 6.11. *Dhom lies to the left of NS.*

Before we establish the relative positions of NS and HH, we need the following results.

LEMMA 6.12.

1. *For $(b, K) \in V_2 \setminus V_2^0$ and to the left of NS, any homoclinic bifurcation that occurs must occur before the Hopf bifurcation.*
2. *For $(b, K) \in V_2$, $K \geq \frac{2}{\sqrt{a}}$ to the right of $E_{\mu K}$, any homoclinic bifurcation that occurs must occur after the Hopf bifurcation.*
3. *For $(b, K) \in V_2 \setminus V_2^0$, $K < \frac{2}{\sqrt{a}}$ to the right of $E_{\mu K}$, any homoclinic bifurcation that occurs must occur after the Hopf bifurcation at $(H_m, F(H_m))$ and before the Hopf bifurcation at $(H_M, F(H_M))$.*
4. *For $(b, K) \in V_2 \setminus V_2^0$, below $E_{\mu K}$ with $b > 0$, both Hopf bifurcations occur before any homoclinic bifurcation.*
5. *For $(b, K) \in V_1 \setminus V_1^1$, any homoclinic bifurcation that occurs must occur before the Hopf bifurcation at $(H_M, f(H_M))$ when $\hat{d} = \hat{d}_+$.*

Proof. 1. For $(b, K) \in V_2 \setminus V_2^0$ and to the left of NS, from Theorem 4.4 and Proposition 6.8, the Hopf bifurcation at $(H_m, F(H_m))$ with $\hat{d} = \hat{d}_-$ occurs after the neutral saddle at $(H_M, F(H_M))$ with $\hat{d} = \hat{d}_+$. This implies that for $\hat{d} \geq \hat{d}_+$ $H_m \leq \lambda \leq \mu < H_M$, and hence $F'(\lambda) \geq 0$ and $F'(\mu) > 0$. By part 2 of Theorem 6.3, there are no periodic orbits or homoclinic loops for $\hat{d} \geq \hat{d}_+$.

2. For $(b, K) \in V_2$ and $K \geq \frac{2}{\sqrt{a}}$, it follows from Theorem 4.4 that the only Hopf bifurcation that occurs is at the $(H_m, F(H_m))$ when $\hat{d} = \hat{d}_-$, and it is subcritical. By part 2(b) of Proposition 6.9, the Hopf bifurcation occurs before the E_μ and E_K transcritical bifurcation. Hence by part 1 of Theorem 6.6, no homoclinic bifurcation occurs for $\hat{d} \in (0, \hat{d}_-]$.

3. For $(b,K) \in V_2$ and $K < \frac{2}{\sqrt{a}}$, by part 2(b) of Proposition 6.9, the transcritical bifurcation involving E_μ and E_K occurs after the Hopf bifurcation at $(H_m, F(H_m))$ when $\hat{d} = \hat{d}_-$ and before the Hopf bifurcation at $(H_M, F(H_M))$ when $\hat{d} = \hat{d}_+$. A similar argument to that for part 2 shows that no homoclinic bifurcation occurs in $(0, \hat{d}_-]$, and a similar argument to that for part 1 shows that no homoclinic bifurcation occurs for $\hat{d} \geq \hat{d}_+$.

4. For $(b,K) \in V_2 \setminus V_2^0$, below $E_{\mu K}$ ($b > 0$), by Theorem 4.4, Hopf bifurcations occur at E_λ when $\lambda = H_m$ and $\lambda = H_M$. But by part 2(c) of Proposition 6.9, both Hopf bifurcations occur before the transcritical bifurcation involving E_μ and E_K, which implies any homoclinic bifurcation must occur after both Hopf bifurcations.

5. For $(b,K) \in V_1 \setminus V_1^1$, by Theorem 4.4, there is one Hopf bifurcation at $(H_M, F(H_M))$ when $\hat{d} = \hat{d}_+$. If $\hat{d} \geq \hat{d}_+$, then $H_M \leq \lambda < \bar{\mu}$ and $F'(x) \leq 0$ for $x \in (\lambda, \bar{\mu})$. Thus by part 1 of Theorem 6.3, for $\hat{d} \geq \hat{d}_+$ no homoclinic loops can exist. Hence any homoclinic bifurcation that occurs must occur before the Hopf bifurcation. □

PROPOSITION 6.13. *Fix all the parameters in* (1.5) *except* $\hat{d} > 0$. *If* (b,K) *is outside the region bounded by* NS, C_1, *and* $E_{\mu K}$ *(Figure* 6.1*), for all* $\hat{d} \in (0, \hat{d}_M)$, *a Hopf bifurcation and a homoclinic bifurcation cannot occur simultaneously.*

Proof. It follows from Theorem 4.4 that no Hopf bifurcations can occur for (b,K) in the regions V_0, V_2^0, and V_1^1. The proofs for $V_1 \setminus V_1^1$ and $V_2 \setminus V_2^0$ follow from Lemma 6.12. □

Recall that in the neighborhood of P, there exists a curve HH along which the Hopf bifurcation at the left hump and a homoclinic bifurcation occur simultaneously. Let V_{HH} be the region bounded by curves NS, $K = \frac{2}{\sqrt{a}}$, and $E_{\mu K}$ ($b < 0$). Now we prove the following theorem regarding the extension of the curve HH in V_{HH}.

THEOREM 6.14. *Fix all parameters except* $\hat{d} > 0$. *In* V_{HH}, *there exists a curve* HH: $K = K_{HH}(b)$ *with finite end point at* $P(-\sqrt{a}, \frac{2}{\sqrt{a}})$ *(Figure* 6.1*). For any* $(b,K) \in HH$, *there exists a unique* $\hat{d}_{hh} \in (\hat{d}_{\mu K}, \hat{d}_M)$, *such that when* $\hat{d} = \hat{d}_{hh}$, *the subcritical Hopf bifurcation at* $(H_m, F(H_m))$ *and a homoclinic loop bifurcation occur simultaneously.*

Proof. By Theorem 6.7, for $(b,K) \in V_{HH}$ there exists $\hat{d}_l \in (\hat{d}_{\mu K}, \hat{d}_M)$ such that a homoclinic bifurcation occurs at $\hat{d} = \hat{d}_l$. By Proposition 6.13, if the curve $K = K_{HH}(b)$ exists, it lies inside the region V_{HH}.

Now we prove that in V_{HH}, for fixed $K > \frac{2}{\sqrt{a}}$, there exists a $b \in (-\frac{2}{K}, \frac{3-aK^2}{K(aK^2-2)})$ such that the subcritical Hopf bifurcation at the left hump and a homoclinic bifurcation involving E_μ happen simultaneously when $\hat{d} = \hat{d}_- (< \hat{d}_+)$.

(a) First we show that along NS, any homoclinic bifurcation at $\hat{d} = \hat{d}_l$ occurs before the Hopf bifurcation.

From Proposition 6.8, if $b = -\frac{2}{K}$ (i.e., $(b,K) \in NS$), a Hopf bifurcation at $(H_m, F(H_m))$ and a neutral saddle at $(H_M, F(H_M))$ occur simultaneously when $\hat{d} = \frac{3K}{2(ak^2-1)}$. Then for $\hat{d} \in [\frac{3K}{2(aK^2-1)}, \hat{d}_M)$, we have $F'(\lambda) \geq 0$ and $F'(\mu) \geq 0$. By Theorem 6.3 neither periodic orbits nor homoclinic loops can exist. Hence along NS, any homoclinic bifurcation occurs before the Hopf bifurcation.

(b) Next we show that along $E_{\mu K}$, the Hopf bifurcation occurs before any homoclinic bifurcation.

This is obvious since the Hopf bifurcation occurs at $\hat{d} = \hat{d}_{\mu K}$, and any homoclinic bifurcation occurs when $\hat{d} > \hat{d}_{\mu K}$.

TABLE 6.1
Curves in the bK plane corresponding to degenerate phenomena. For the last 7 curves, there exists a $\hat{d}$ such that the indicated bifurcation occur.

Name	Phenomenon	Expression	
C_1	$H_m = H_M$	$K = \frac{\sqrt{3(4a-b^2)}-b}{2a}$, $-2\sqrt{a} < b \leq \sqrt{a}$	(3.8)
C_2	$H_m = 0$ or $H_M = 0$	$K = \frac{1}{b}$, $b > 0$	(3.8)
$E_{\mu K}$	Hopf & $E_\mu - E_K$ transcritical	$b = \frac{3 - aK^2}{K(aK^2 - 2)}$	(6.13)
NS	Hopf & neutral saddle	$K = -\frac{2}{b}$	(6.11)
BT	Bogdanov–Takens	$K = \frac{2}{\sqrt{a}}$	(5.1)
DH	Degenerate Hopf	$\sigma(K, b) = 0$	(4.11)
HH	Hopf & homoclinic	$K = K_{HH}(b)$	Theorem 6.14
$Dhom$	Degenerate homoclinic	$K = K_{Dhom}(b)$	(6.15)

By (a) and (b), for any K with (b, K) in V_{HH} there exists at least one $b^* \in (-\frac{2}{K}, \frac{3-aK^2}{K(aK^2-2)})$ such that at (b^*, K), there exists a unique $\hat{d} \in (\hat{d}_{\mu K}, \hat{d}_M)$ such that a Hopf bifurcation and a homoclinic loop bifurcation occur simultaneously. □

Remark 6.15. Theorem 5.4 proves the local existence of the curve HH emanating from P and lying between the curves NS and DH. Theorem 6.14 shows that this curve may be globally extended into the region V_{HH}. Since both NS and $E_{\mu K}$ tend to $b = 0$ as $K \to \infty$, the global extension of HH should do the same. In Figure 6.1 we have drawn the curve HH as a single branch emanating from P and lying above DH in V_{HH}. This representation of HH is supported by numerical simulations using XPPAUT [11]; however, we have not been able analytically to preclude that HH has multiple branches or crosses the curve DH.

We summarize the relevant curves in the bK plane in Table 6.1. As shown in Figure 6.1, we use these curves to divide V_0, V_1, and V_2 into the following subregions:

$$V_0 = V_0^1 \cup V_0^2,$$
$$V_1 = \bigcup_{k=1}^{4} V_1^k,$$
$$V_2 = \bigcup_{k=0}^{11} V_2^k.$$

6.3. Saddle-node bifurcation of limit cycles. Let V^* be the subregion of V_2 (Figure 6.1) bounded by $Dhom$, DH, and C_2; i.e.,

$$V^* = V_2^1 \cup V_2^2 \cup V_2^6 \cup V_2^8 \cup V_2^9 \cup V_2^{10}. \tag{6.16}$$

To study the saddle-node bifurcation of limit cycles for parameters away from P, we introduce a line segment L in the following proposition (see Figure 6.2).

PROPOSITION 6.16. *In the bK plane, the line*

$$L: \quad K = -\frac{b}{a} + \frac{2}{\sqrt{a}} \tag{6.17}$$

lies between the curves C_2 and DH and is tangent to C_1, DH, and C_2 at Q. In region V_2, below this line, $F(0) > F(H_M)$; above the line, $F(0) < F(H_M)$; and on the line, $F(0) = F(H_M)$.

Proof. The proof follows from direct calculations. □

As shown in Figure 6.2, the line L subdivides the region V^* into two subregions. Denote the shaded subregion below L by

$$V_{sn} = \left\{ (b,K) \in V^* \,\middle|\, K < -\frac{b}{a} + \frac{2}{\sqrt{a}} \right\}.$$

PROPOSITION 6.17. *Fix all the parameters except $\hat{d} > 0$. If $(b,K) \in V_{sn}$, there exists a $\hat{d}_{sn} \in (0, \hat{d}_-)$ such that when $\hat{d} = \hat{d}_{sn}$, system (1.3) undergoes a saddle-node bifurcation of limit cycles.*

Proof. For $(b,K) \in V_{sn}$, it follows from Proposition 6.16 that there exists a $\hat{d}_0 > 0$ such that for $\hat{d} \in (0, \hat{d}_0)$, $F(\lambda) > F(H_M)$. By Corollary 6.2, system (1.3) has neither periodic orbits nor homoclinic loops for $(b,K) \in V_{sn}$ and $\hat{d} \in (0, \hat{d}_0)$. Further, it follows from Theorem 4.4 that there exists a $\hat{d}_- \in [\hat{d}_0, \hat{d}_M)$ such that when $\hat{d} = \hat{d}_-$, system (1.3) undergoes a subcritical Hopf bifurcation at $(H_m, F(H_m))$. Recall that E_λ is asymptotically stable for $0 < \hat{d} < \hat{d}_-$ (see Table 2.1) and so an unstable periodic orbit must be destroyed as $\hat{d}$ increases through $\hat{d}_-$.

(1) In V_{sn} for $K \geq \frac{2}{\sqrt{a}}$ and to the right of the curve NS, from Theorem 4.4, there are no other Hopf bifurcations. From Lemma 6.10 any homoclinic bifurcation that occurs in supercritical. Thus, the only way to create the unstable limit cycle that must be destroyed in the subcritical Hopf bifurcation is to first have a saddle-node bifurcation of limit cycles.

(2) In V_{sn} for $K < \frac{2}{\sqrt{a}}$, by part 5 of Theorem 4.4, in addition to the subcritical Hopf bifurcation at $\hat{d} = \hat{d}_-$, and there is a supercritical Hopf bifurcation at $\hat{d} = \hat{d}_+$. From Lemma 6.10, any homoclinic bifurcation that occurs is supercritical. By a similar argument as for case (1), there must be a saddle-node bifurcation of limit cycles before the two Hopf bifurcations.

(3) In V_{sn} for $K \geq \frac{2}{\sqrt{a}}$, by Theorem 6.7 there exists a $\hat{d}_l$ such that when $\hat{d} = \hat{d}_l$ system (1.3) undergoes a homoclinic bifurcation. By Lemma 6.12, this homoclinic bifurcation must occur before the Hopf bifurcation. By definition, to the right of *Dhom* there must be a supercritical homoclinic bifurcation. Thus, an unstable limit cycle must already surround the asymptotically stable equilibrium. This limit cycle must have been created by a saddle-node bifurcation of limit cycles. □

PROPOSITION 6.18.

1. *In the bK plane (Figure 6.3), the curve*

$$C_{sn}: \quad b = \frac{3 - aK^2}{K}, \quad K \geq \sqrt{\frac{3}{a}} \tag{6.18}$$

lies between the curves C_1 and $E_{\mu K}$ and is tangent to them at $(0, \sqrt{3/a})$. For $(b,K) \in V_{sn}$, below this curve, if $\hat{d} \in (0, \hat{d}_{\mu K})$, system (1.3) has no closed orbits. Therefore the saddle-node bifurcation of limit cycles must occur after the transcritical bifurcation involving E_μ and E_K.

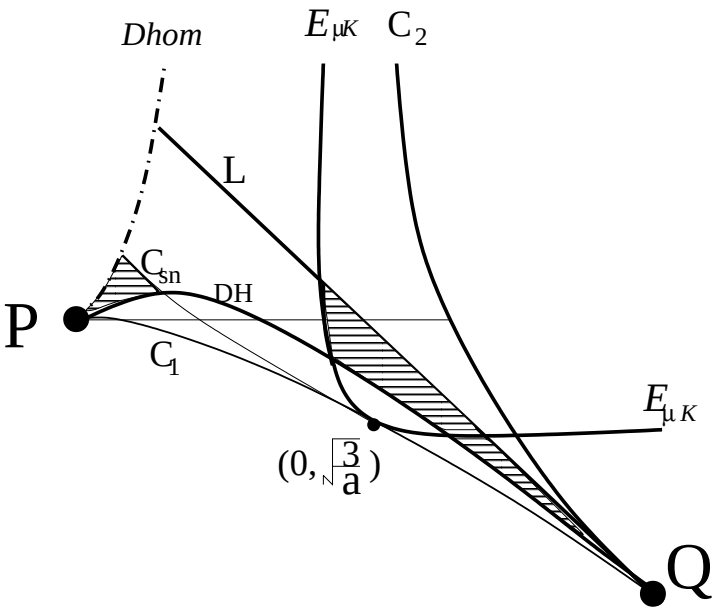


FIG. 6.3. *In the shaded subregion of V_{sn} below C_{sn}, the saddle-node bifurcation of limit cycles occurs after the transcritical bifurcation involving E_μ and E_K. In the shaded subregion of V_{sn} to the right of $E_{\mu K}$, the saddle-node bifurcation of limit cycles occurs before the transcritical bifurcation.*

2. *For $(b, K) \in V_{sn}$ to the right of the branch of $E_{\mu K}$ with $b < 0$, a saddle-node bifurcation of limit cycles occurs before the transcritical bifurcation involving E_μ and E_K.*

Proof. Note that in V_{sn}, $F(0) > F(H_M)$. As $\hat{d}$ increases from 0, there exists a $\hat{d}_0$ such that for $\hat{d} = \hat{d}_0$, $\lambda = \lambda_0$ where $F(\lambda_0) = F(H_M)$ and $\lambda_0 < H_m$. By Corollary 6.2, for $0 < \hat{d} < \hat{d}_0$, there are no periodic orbits or homoclinic loops. Therefore, the saddle-node bifurcation of limit cycles must occur for some $\hat{d} > \hat{d}_0$.

Setting $\mu = K$ and $F(\lambda) = F(H_M)$ in (1.4), (2.4), and (3.7) and eliminating $\hat{d}$, we obtain (6.13) and (6.18). By Proposition 6.9, if (b, K) satisfies (6.13), then when $\hat{d} = \hat{d}_{\mu K}$, we have $\lambda = H_m$. Hence (6.13) is not relevant and (6.18) corresponds to $\hat{d}_{\mu K} = \hat{d}_0$. Straightforward calculations show that curve (6.18) lies between C_1 and $E_{\mu K}$, is tangent to both C_1 and $E_{\mu K}$ at $(0, \sqrt{3/a})$, and lies below L in V_{sn}. Consideration of the sign of $F(\lambda) - F(H_M)$ when $\hat{d} = \hat{d}_{\mu K}$ shows that for $(b, K) \in V_{sn}$ below curve (6.18), $\hat{d}_0 > \hat{d}_{\mu K}$, and above curve (6.18), $\hat{d}_0 < \hat{d}_{\mu K}$. Therefore, the result follows.

On the other hand, for $(b, K) \in V_{sn}$ to the right of the branch of the curve $E_{\mu K}$ with $b < 0$, the Hopf bifurcation at $(H_m, F(H_m))$ occurs at the same time as or before the transcritical bifurcation involving E_μ and E_K. By Proposition 6.17, for $(b, K) \in V_{sn}$, the Hopf bifurcation at $(H_m, F(H_m))$ occurs after the saddle-node bifurcation of limit cycles. Therefore in this region a saddle-node bifurcation of limit cycles must occur before the transcritical bifurcation. □

As shown in Figure 6.4, the segment of C_{sn} between *Dhom* and *DH* and the segment of L between *Dhom* and $E_{\mu K}$ divide the regions V_2^8 and V_2^{10} into subregions V_2^{ia}, V_2^{ib}, and V_2^{ic} $(i = 8, 10)$. For the saddle-node bifurcation of limit cycles and its

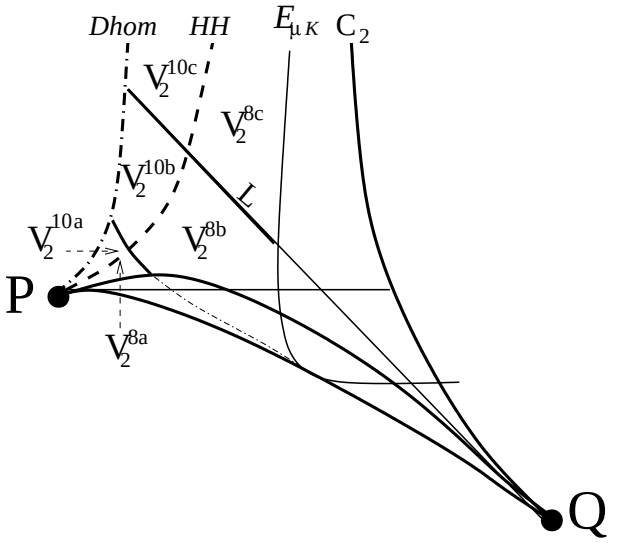


FIG. 6.4. *In V_{sn}, the curve C_{sn} and line L divide the regions V_2^8 and V_2^{10} into subregions V_2^{ia}, V_2^{ib}, and V_2^{ic} $(i = 8, 10)$.*

relative order with respect to the transcritical bifurcation involving E_μ and E_K, we make the following remark.

Remark 6.19.

1. By Proposition 6.18, for (b, K) in V_2^{8a} and V_2^{10a}, the saddle-node bifurcation of limit cycles occurs after the transcritical bifurcation involving E_μ and E_K.
2. When $\hat{d} = 0$, system (1.3) has a degenerate graphic with a line segment of equilibria. The bifurcation analysis of this type of graphic is very complicated (see [9] for reference), and will appear elsewhere.
3. For (b, K) in V_2^{ic} $(i = 8, 10)$ and V_2^9, above the line L, a saddle-node bifurcation of limit cycles may occur or the required limit cycles may come from a bifurcation of the degenerate singularity when $\hat{d} = 0$. Numerical simulations show that for (b, K) in V_2^{ic} $(i = 8, 10)$ and V_2^9, above the line L, system (1.3) has two limit cycles for $\hat{d} > 0$ very small.
4. For (b, K) in V_2^{ib} $(i = 8, 10)$ and V_2^6, it follows from Proposition 6.18 that there exists a curve that lies between C_{sn} and $E_{\mu K}$. For (b, K) on this curve, there exists a $\hat{d}_{sn} > 0$ such that a saddle-node bifurcation of limit cycles and the transcritical bifurcation involving E_μ and E_K occur simultaneously. This curve ST may not be unique.

6.4. Sequences of bifurcations. Although a three dimensional Hopf bifurcation surface is shown in Figure 4.1, it is not convenient to visualize the entire bifurcation diagram for system (1.3) in $(b, \hat{d}, K)$ space. Instead we describe the bifurcation diagram as $\hat{d}$ varies for fixed b and K inside each subregion (see Figure 6.1) of the bK plane. The sequences of bifurcations that occur in the interior of each subregion are given using the "dictionary of phase portraits" in Table 6.2. The sequences on the boundaries of each subregion are not included. These can easily be deduced by reading Tables 6.3–6.5 vertically. On the boundaries various simultaneous or degenerate bifurcations occur.

TABLE 6.2
Dictionary of phase portraits.

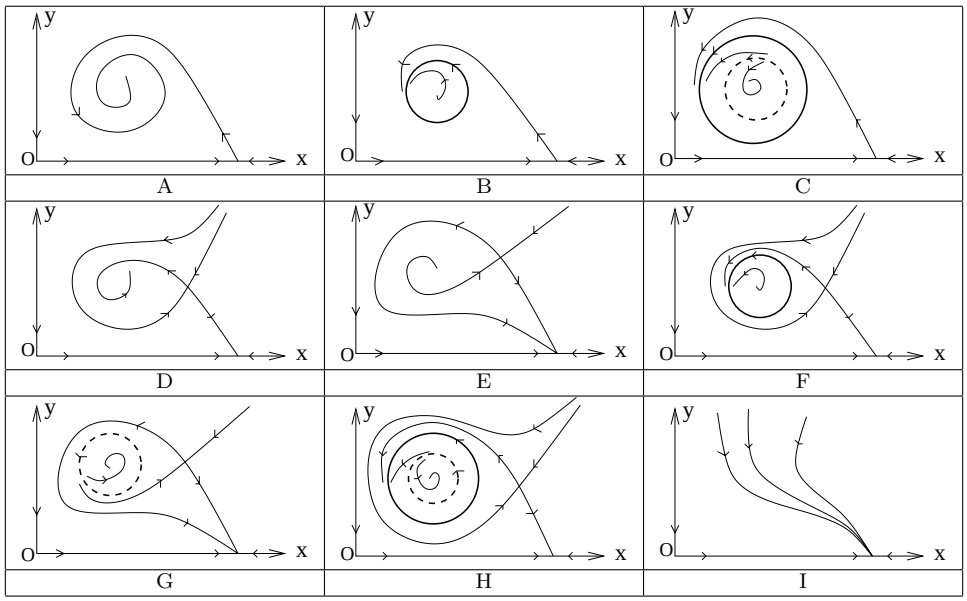


TABLE 6.3
Sequences of phase portraits for $(b, K) \in V_0$.

V_0^1	A	D	I
	$(0, \hat{d}_{\mu K})$	$(\hat{d}_{\mu K}, \hat{d}_M)$	$(\hat{d}_M, \infty)$
V_0^2	A	I	
	$(0, \hat{d}_{\lambda K})$	$(\hat{d}_{\lambda K}, \infty)$	

THEOREM 6.20. *Fix all parameters except* $\hat{d} > 0$. *For* $(b, K) \in V_0$, *when* $\hat{d} > 0$ *increases, the sequence of phase portraits occurring in the interior of each subregion is given in Table* 6.3. *In the table, moving from left to right as* $\hat{d}$ *increases, the phase portrait changes as a result of one of the following bifurcations:*

- *the transcritical bifurcation involving* E_λ *and* E_K *that occurs at* $\hat{d} = \hat{d}_{\lambda K}$;
- *the transcritical bifurcation involving* E_μ *and* E_K *that occurs at* $\hat{d} = \hat{d}_{\mu K}$;
- *the saddle-node bifurcation involving* E_λ *and* E_μ *that occurs at* $\hat{d} = \hat{d}_M$.

Proof. By Corollary 6.5, system (1.3) has neither periodic orbits nor homoclinic loops for $(b, K) \in V_0$. As $\hat{d} \in (0, \hat{d}_M)$ is varied, if $K > \frac{1}{\sqrt{a}}$, the only bifurcation that can occur is the transcritical bifurcation involving E_μ and E_K that occurs when $\hat{d} = \hat{d}_{\mu K}$ and the saddle-node bifurcation that occurs when $\hat{d} = \hat{d}_M$. If $0 < K < \frac{1}{\sqrt{a}}$, the only bifurcation that can occur is the transcritical bifurcation involving E_λ and E_K that occurs when $\hat{d} = \hat{d}_{\lambda K}$. □

THEOREM 6.21. *Fix all parameters except* $\hat{d} > 0$. *For* $(b, K) \in V_1$, *when* $\hat{d} > 0$ *increases, the sequence of phase portraits occurring in the interior of each subregion is given in Table* 6.4. *In the table, moving from left to right as* $\hat{d}$ *increases, the phase*

TABLE 6.4
Sequences of phase portraits for $(b, K) \in V_1$.

V_1^1	B	F	E	I
	$(0, \hat{d}_{\mu K})$	$(\hat{d}_{\mu K}, \hat{d}_l)$	$(\hat{d}_l, \hat{d}_M)$	$(\hat{d}_M, \infty)$
V_1^2	B	F	D	I
	$(0, \hat{d}_{\mu K})$	$(\hat{d}_{\mu K}, \hat{d}_+)$	$(\hat{d}_+, \hat{d}_M)$	$(\hat{d}_M, \infty)$
V_1^3	B	A	D	I
	$(0, \hat{d}_+)$	$(\hat{d}_+, \hat{d}_{\mu K})$	$(\hat{d}_{\mu K}, \hat{d}_M)$	$(\hat{d}_M, \infty)$
V_1^4	B	A	I	
	$(0, \hat{d}_+)$	$(\hat{d}_+, \hat{d}_{\lambda K})$	$(\hat{d}_{\lambda K}, \infty)$	

portrait changes due to one of the following bifurcations:

- *the transcritical bifurcation involving* E_λ *and* E_K *that occurs at* $\hat{d} = \hat{d}_{\lambda K}$;
- *the transcritical bifurcation involving* E_μ *and* E_K *that occurs at* $\hat{d} = \hat{d}_{\mu K}$;
- *the saddle-node bifurcation involving* E_λ *and* E_μ *that occurs at* $\hat{d} = \hat{d}_M$;
- *the supercritical Hopf bifurcation that occurs when* $\hat{d} = \hat{d}_+$;
- *a supercritical homoclinic bifurcation that occurs when* $\hat{d} = \hat{d}_l$.

For $(b, K) \in V_1^3 \cup V_1^4$, *the sequence is complete up to an even number of saddle-node bifurcations of limit cycles. For* $(b, K) \in V_1^1 \cup V_1^2$, *the sequences are complete up to an even number of saddle-node bifurcations of limit cycles and an even number of extra supercritical homoclinic bifurcations.*

Proof. In V_1 for $\hat{d} > 0$ sufficiently small, E_λ is an unstable node. Since solutions are bounded, a simple phase plane argument shows that there must be a stable limit cycle surrounding E_λ.

If $(b, K) \in V_1^1$, by Theorem 4.4, system (1.3) does not undergo Hopf bifurcations. Thus there must exist $\hat{d}_l$ such that when $\hat{d} = \hat{d}_l$ there is a supercritical homoclinic bifurcation destroying the limit cycle described above. Further, a transcritical bifurcation involving E_μ and E_K occurs when $\hat{d} = \hat{d}_{\mu K}$. By Theorem 6.6, $\hat{d}_{\mu K} < \hat{d}_l$. Thus, for $\hat{d} \in (0, \hat{d}_{\mu K})$, the system has a stable periodic orbit. For $\hat{d} \in (\hat{d}_l, \hat{d}_M)$, the system has no periodic orbit.

If $(b, K) \in V_1^2$, by Theorem 4.4, system (1.3) undergoes a supercritical Hopf bifurcation when $\hat{d} = \hat{d}_+$ ($\lambda = H_M$). It follows from Proposition 6.8 that the Hopf bifurcation occurs after the transcritical bifurcation involving E_μ and E_K at $\hat{d} = \hat{d}_{\mu K}$. For $\hat{d} \in (\hat{d}_+, \hat{d}_M)$, $0 < H_M < \lambda < \frac{1}{\sqrt{a}} < \mu < K$. By part 1 of Theorem 6.3, system (1.3) has neither periodic orbits nor homoclinic loops and hence has the phase portrait D.

If $(b, K) \in V_1^3$, the sequence of bifurcations is the same as in V_1^2 except that the supercritical Hopf bifurcation occurs before the transcritical bifurcation involving E_μ and E_K. Hence by part 1 of Theorem 6.3, no homoclinic bifurcations can occur.

If $(b, K) \in V_1^4$, then $0 < K < \frac{1}{\sqrt{a}}$. For $\hat{d} \in (0, \hat{d}_M)$, $K < \mu$. In this case, the transcritical bifurcation involves E_λ and E_K and occurs when $\hat{d} = \hat{d}_{\lambda K}$. Note that since $\mu > K$, no homoclinic bifurcation can occur by part 1 of Theorem 6.6. ∎

THEOREM 6.22. *Fix all parameters except* $\hat{d} > 0$. *For* $(b, K) \in V_2$, *when* $\hat{d} > 0$ *increases, the sequence of phase portraits occurring in the interior of each subregion is given in Table* 6.5. *In the table, moving from left to right as* $\hat{d}$ *increases, the phase portrait changes as a result of one of the following bifurcations:*

TABLE 6.5

Sequences of phase portraits for $(b, K) \in V_2$. In the three regions indicated by a †, above the line L, instead of a saddle-node bifurcation of limit cycles, there could be a degenerate bifurcation for $\hat{d} = 0$. In this case the sequence of phase portraits would begin with C instead of A. If more than one sequence is shown for a given region, it indicates that all of the sequences given are possible.

V_2^0	A	D	I			
	$(0, \hat{d}_{\mu K})$	$(\hat{d}_{\mu K}, \hat{d}_M)$	$(\hat{d}_M, \infty)$			
$V_2^{1\dagger}$	A	C	B	F	D	I
	$(0, \hat{d}_{sn})$	$(\hat{d}_{sn}, \hat{d}_-)$	$(\hat{d}_-, \hat{d}_{\mu K})$	$(\hat{d}_{\mu K}, \hat{d}_+)$	$(\hat{d}_+, \hat{d}_M)$	$(\hat{d}_M, \infty)$
$V_2^{2\dagger}$	A	C	B	A	D	I
	$(0, \hat{d}_{sn})$	$(\hat{d}_{sn}, \hat{d}_-)$	$(\hat{d}_-, \hat{d}_+)$	$(\hat{d}_+, \hat{d}_{\mu K})$	$(\hat{d}_{\mu K}, \hat{d}_M)$	$(\hat{d}_M, \infty)$
V_2^3	A	B	A	D	I	
	$(0, \hat{d}_-)$	$(\hat{d}_-, \hat{d}_+)$	$(\hat{d}_+, \hat{d}_{\mu K})$	$(\hat{d}_{\mu K}, \hat{d}_M)$	$(\hat{d}_M, \infty)$	
V_2^4	A	B	F	D	I	
	$(0, \hat{d}_-)$	$(\hat{d}_-, \hat{d}_{\mu K})$	$(\hat{d}_{\mu K}, \hat{d}_+)$	$(\hat{d}_+, \hat{d}_M)$	$(\hat{d}_M, \infty)$	
V_2^5	A	D	F	D	I	
	$(0, \hat{d}_{\mu K})$	$(\hat{d}_{\mu K}, d_-)$	$(\hat{d}_-, \hat{d}_+)$	$(\hat{d}_+, \hat{d}_M)$	$(\hat{d}_M, \infty)$	
V_2^6	A	D	H	F	D	I
	$(0, \hat{d}_{\mu K})$	$(\hat{d}_{\mu K}, d_{sn})$	$(\hat{d}_{sn}, \hat{d}_-)$	$(\hat{d}_-, \hat{d}_+)$	$(\hat{d}_+, \hat{d}_M)$	$(\hat{d}_M, \infty)$
	A	C	H	F	D	I
	$(0, \hat{d}_{sn})$	$(\hat{d}_{sn}, d_{\mu K})$	$(\hat{d}_{\mu K}, \hat{d}_-)$	$(\hat{d}_-, \hat{d}_+)$	$(\hat{d}_+, \hat{d}_M)$	$(\hat{d}_M, \infty)$
V_2^7	A	D	F	E	I	
	$(0, \hat{d}_{\mu K})$	$(\hat{d}_{\mu K}, \hat{d}_-)$	$(\hat{d}_-, \hat{d}_l)$	$(\hat{d}_l, \hat{d}_M)$	$(\hat{d}_M, \infty)$	
$V_2^{8a,b,c}$	A	D	H	F	E	I
	$(0, \hat{d}_{\mu K})$	$(\hat{d}_{\mu K}, \hat{d}_{sn})$	$(\hat{d}_{sn}, \hat{d}_-)$	$(\hat{d}_-, \hat{d}_l)$	$(\hat{d}_l, \hat{d}_M)$	$(\hat{d}_M, \infty)$
$V_2^{8b,c}$	A	C	H	F	E	I
	$(0, \hat{d}_{sn})$	$(\hat{d}_{sn}, \hat{d}_{\mu K})$	$(\hat{d}_{\mu K}, \hat{d}_-)$	$(\hat{d}_-, \hat{d}_l)$	$(\hat{d}_l, \hat{d}_M)$	$(\hat{d}_M, \infty)$
V_2^{8c}		C	H	F	E	I
		$(0, \hat{d}_{\mu K})$	$(\hat{d}_{\mu K}, \hat{d}_-)$	$(\hat{d}_-, \hat{d}_l)$	$(\hat{d}_l, \hat{d}_M)$	$(\hat{d}_M, \infty)$
$V_2^{9\dagger}$	A	C	B	F	E	I
	$(0, \hat{d}_{sn})$	$(\hat{d}_{sn}, \hat{d}_-)$	$(\hat{d}_-, \hat{d}_{\mu K})$	$(\hat{d}_{\mu K}, \hat{d}_l)$	$(\hat{d}_l, \hat{d}_M)$	$(\hat{d}_M, \infty)$
$V_2^{10a,b,c}$	A	D	H	G	E	I
	$(0, \hat{d}_{\mu K})$	$(\hat{d}_{\mu K}, \hat{d}_{sn})$	$(\hat{d}_{sn}, \hat{d}_l)$	$(\hat{d}_l, \hat{d}_-)$	$(\hat{d}_-, \hat{d}_M)$	$(\hat{d}_M, \infty)$
$V_2^{10b,c}$	A	C	H	G	E	I
	$(0, \hat{d}_{sn})$	$(\hat{d}_{sn}, \hat{d}_{\mu K})$	$(\hat{d}_{\mu K}, \hat{d}_l)$	$(\hat{d}_l, \hat{d}_-)$	$(\hat{d}_-, \hat{d}_M)$	$(\hat{d}_M, \infty)$
V_2^{10c}		C	H	G	E	I
		$(0, \hat{d}_{\mu K})$	$(\hat{d}_{\mu K}, \hat{d}_l)$	$(\hat{d}_l, \hat{d}_-)$	$(\hat{d}_-, \hat{d}_M)$	$(\hat{d}_M, \infty)$
V_2^{11}	A	D	G	E	I	
	$(0, \hat{d}_{\mu K})$	$(\hat{d}_{\mu K}, \hat{d}_l)$	$(\hat{d}_l, \hat{d}_-)$	$(\hat{d}_-, \hat{d}_M)$	$(\hat{d}_M, \infty)$	

- *the transcritical bifurcation involving E_μ and E_K that occurs at $\hat{d} = \hat{d}_{\mu K}$;*
- *the saddle-node bifurcation involving E_λ and E_μ that occurs at $\hat{d} = \hat{d}_M$;*
- *a Hopf bifurcation that occurs when $\hat{d} = \hat{d}_-$ or $\hat{d} = \hat{d}_+$;*
- *a homoclinic bifurcation that occurs when $\hat{d} = \hat{d}_l$;*
- *a saddle-node bifurcation of limit cycles that occurs when $\hat{d} = \hat{d}_{sn}$.*

For $(b, K) \in V_2^0$, the sequence is complete.

For $(b, K) \in V_2^2 \cup V_2^3$, the sequences are complete up to an even number of extra saddle-node bifurcations of limit cycles.

For $(b, K) \in V_2^1 \cup V_2^4 \cup V_2^5 \cup V_2^6 \cup V_2^7 \cup V_2^9$, the sequences are complete up to an even number of extra supercritical homoclinic bifurcations and an even number of extra saddle-node bifurcations of limit cycles.

For $(b, K) \in V_2^8 \cup V_2^{10} \cup V_2^{11}$, the sequences are complete up to saddle-node bifurcations of limit cycles and an even number of extra homoclinic bifurcations.

Proof. The region V_2 has 12 subregions V_2^i $(i = 0, 1, \ldots, 11)$.

1. For $(b, K) \in V_2^0$, $0 < \frac{1}{\sqrt{a}} < H_m < H_M < K$. It follows from Theorem 4.4 and Corollary 6.5 that system (1.3) does not undergo Hopf bifurcations and has neither periodic orbits nor homoclinic loops. The only bifurcations that can occur are a transcritical bifurcation involving E_μ and E_K at $\hat{d} = \hat{d}_{\mu K}$ and a saddle-node bifurcation involving E_λ and E_μ at $\hat{d} = \hat{d}_M$. The transcritical bifurcation must occur before the saddle-node bifurcation involving E_λ and E_μ.

2. For $(b, K) \in V_2^1 \cup V_2^2 \cup V_2^6$, $0 < H_m < H_M < \frac{1}{\sqrt{a}} < K$. It follows from Theorem 4.4 that there exist $\hat{d}_-$ and $\hat{d}_+$ $(\hat{d}_- < \hat{d}_+)$ such that when $\hat{d} = \hat{d}_-$, a subcritical Hopf bifurcation occurs at $(H_m, F(H_m))$, and when $\hat{d} = \hat{d}_+$, a supercritical Hopf bifurcation occurs at $(H_M, F(H_M))$.

For $(b, K) \in V_2^1$, by Proposition 6.9, the Hopf bifurcation at $(H_m, F(H_m))$ occurs before the transcritical bifurcation involving E_μ and E_K, i.e., $\hat{d}_- < \hat{d}_{\mu K} < \hat{d}_+$. For $\hat{d} \in (\hat{d}_+, \hat{d}_M)$, there are two equilibria, E_λ and E_μ, satisfying $H_M < \lambda < \frac{1}{\sqrt{a}} < \mu < K$. By Theorem 6.3, the system has neither periodic orbits nor homoclinic loops. The two equilibria E_λ and E_μ disappear through a saddle-node bifurcation when $\hat{d} = \hat{d}_M$. If (b, K) is below the line L, by Proposition 6.17, there exists a $\hat{d}_{sn} < \hat{d}_+$ such that the system undergoes a saddle-node bifurcation of limit cycles. If (b, K) is above the line L, by part 3 of Remark 6.19, this saddle node bifurcation of limit cycles may not occur; instead, two limit cycles bifurcate from $\hat{d} = 0$.

For $(b, K) \in V_2^2$, it follows from Proposition 6.9 that the only difference from the case when $(b, K) \in V_2^1$ is that the transcritical bifurcation involving E_μ and E_K occurs after the Hopf bifurcation at $\hat{d} = \hat{d}_+$.

For $(b, K) \in V_2^6$, it follows from Proposition 6.9 that the transcritical bifurcation involving E_μ and E_K occurs before both Hopf bifurcations. Since V_2^6 sits entirely below the line L, by Proposition 6.17, there must exist a $\hat{d}_{sn} \in (0, \hat{d}_-)$ such that system (1.3) undergoes a saddle-node bifurcation of limit cycles. From part 4 of Remark 6.19, it is not clear whether $\hat{d}_{sn} < \hat{d}_{\mu K}$ or $\hat{d}_{\mu K} < \hat{d}_{sn}$.

3. For $(b, K) \in V_2^3 \cup V_2^4 \cup V_2^5$, $0 < H_m < H_M < \frac{1}{\sqrt{a}} < K$. By Theorem 4.4, there are two supercritical Hopf bifurcations, the first at $\hat{d} = \hat{d}_-$ and the second at $\hat{d} = \hat{d}_+$.

It follows from Proposition 6.9 that if $(b, K) \in V_2^3$, $\hat{d}_- < \hat{d}_+ < \hat{d}_{\mu K}$; if $(b, K) \in V_2^4$, $\hat{d}_- < \hat{d}_{\mu K} < \hat{d}_+$; if $(b, K) \in V_2^5$, $\hat{d}_{\mu K} < \hat{d}_- < \hat{d}_+$.

4. For $(b, K) \in V_2^7$, $0 < H_m < \frac{1}{\sqrt{a}} < H_M < K$. By Theorem 4.4, there exists a $\hat{d}_- \in (0, \hat{d}_M)$ such that when $\hat{d} = \hat{d}_-$, system (1.3) undergoes a supercritical Hopf

TABLE 6.6
Additional homoclinic bifurcations observed for $(b, K) \in V_2^1$ as $\hat{d}$ is varied.

F	E	F
$(\hat{d}_{\mu K}, \hat{d}_{l1})$	$(\hat{d}_{l1}, \hat{d}_{l2})$	$(\hat{d}_{l2}, \hat{d}_{+})$

bifurcation. By Proposition 6.9, $\hat{d}_{\mu K} < \hat{d}_{-}$. By Theorem 6.7 and Lemma 6.10, there exists a $\hat{d}_l$ such that when $\hat{d} = \hat{d}_l$, the system undergoes a supercritical homoclinic bifurcation. The homoclinic bifurcation must destroy the periodic orbit created by the Hopf bifurcation, so $\hat{d}_l > \hat{d}_{-}$.

5. For $(b, K) \in V_2^8 \cup V_2^9 \cup V_2^{10} \cup V_2^{11}$, $0 < H_m < \frac{1}{\sqrt{a}} < H_M < K$. By Theorem 4.4, there exists a unique $\hat{d}_{-} \in (0, \hat{d}_M)$ such that when $\hat{d} = \hat{d}_{-}$, system (1.3) undergoes a subcritical Hopf bifurcation. Therefore, an unstable periodic orbit must exist when $\hat{d}_{-} - \epsilon < \hat{d} < \hat{d}_{-}$ for some ϵ. This periodic orbit must be created in either (i) a saddle-node bifurcation of limit cycles for some $\hat{d}_{sn} < \hat{d}_{-}$, or (ii) a degenerate bifurcation at $\hat{d} = 0$ creating an even number of limit cycles, or (iii) a subcritical homoclinic bifurcation. In cases (i) and (ii), the outside, asymptotically stable periodic orbit would have to be destroyed in a supercritical homoclinic bifurcation. Thus, in all three cases, there must exist a $\hat{d}_l$ at which a homoclinic bifurcation occurs.

By definition of *Dhom*, the homoclinic bifurcation changes stability and is subcritical in V_2^{11} and supercritical in $V_2^8 \cup V_2^9 \cup V_2^{10}$. By definition of HH, $\hat{d}_l < \hat{d}_{-}$ to the left of HH and $\hat{d}_l > \hat{d}_{-}$ to the right of HH. By definition of $E_{\mu K}$, $\hat{d}_{-} < \hat{d}_{\mu K}$ to the right of the branch of $E_{\mu K}$ with $b < 0$, and $\hat{d}_{-} > \hat{d}_{\mu K}$ to the left of this branch.

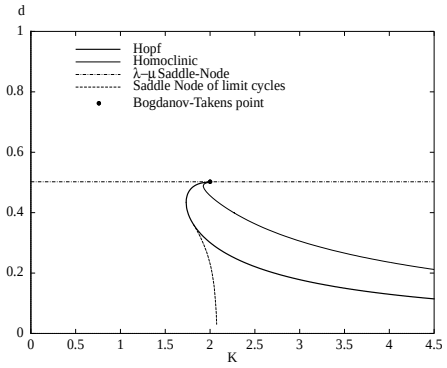
For $(b, K) \in V_2^{11}$, there is a subcritical homoclinic bifurcation at $\hat{d} = \hat{d}_l$ (i.e., case (iii) occurs) and by part 1 of Lemma 6.12, $\hat{d}_l < \hat{d}_{-}$. By part 1 of Theorem 6.6, $\hat{d}_{\mu K} < \hat{d}_l$. Therefore $0 < \hat{d}_{\mu K} < \hat{d}_l < \hat{d}_{-} < \hat{d}_M$.

For $(b, K) \in V_2^{10} \cup V_2^9 \cup V_2^8$, the homoclinic bifurcation is supercritical. In $V_2^{10a} \cup V_2^{10b} \cup V_2^{8a} \cup V_2^{8b}$ and V_2^9 below the line L case (i) occurs. In $V_2^{10c} \cup V_2^{8c}$ and V_2^9 above the line L, case (i) or (ii) occurs.

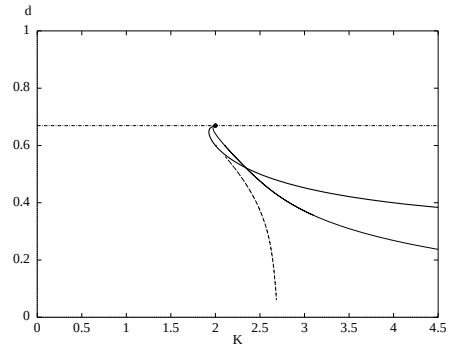
6. By part 1 of Theorem 6.6 and part 1 of Theorem 6.3, there is no homoclinic bifurcation for $(b, K) \in V_2^2 \cup V_2^3$. Solutions are bounded. There are either an even number or zero limit cycles for $\hat{d} > 0$ sufficiently small, and there are no limit cycles for $\hat{d} > \hat{d}_M$. We have considered all the other possible local and necessary global bifurcations. Hence, the sequences are complete as described in Table 6.5 up to saddle-node bifurcations of limit cycles and homoclinic bifurcations as indicated in the statement of the theorem. □

Remark 6.23. The sequence of phase portraits in Table 6.5 is complete up to saddle-node bifurcations of limit cycles and homoclinic loop bifurcations. For $(b, K) \in V_2^1$ and $\hat{d} \in (\hat{d}_{\mu K}, \hat{d}_{+})$, numerical simulations using XPPAUT [11] suggest that in addition to the critical values shown in the table, there could exist $\hat{d}_{l1}, \hat{d}_{l2} \in (\hat{d}_{\mu K}, \hat{d}_{+})$ such that when $\hat{d} = \hat{d}_{l1}$ and $\hat{d} = \hat{d}_{l2}$, system (1.3) undergoes homoclinic bifurcations and includes the subsequence of phase portraits listed in Table 6.6.

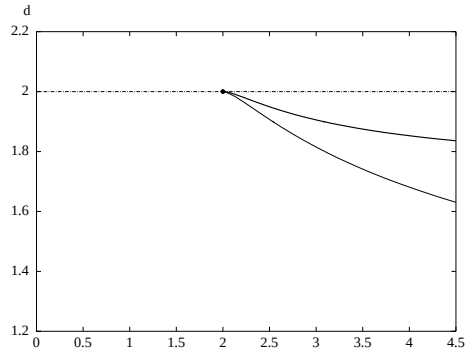
Numerical continuation of bifurcation curves carried out with the software Auto (through the XPPAUT [11] interface), supports our analysis. Fixing $a = m = c = r = 1$, which is consistent with the rescaling (1.7), we calculated the two parameter bifurcation sets in dK space, for $b = 0, -0.5, -1.5$ (Figure 6.5). Fixing K in Figure 6.5 and allowing d to vary vertically, we obtain sequences of bifurcations indicated by our analysis.



(a) $b = 0$



(b) $b = -0.5$



(c) $b = -1.5$

FIG. 6.5. *Two dimensional bifurcation diagram from XPPAUT* [11]*:* $a = m = c = r = 1$.

7. Discussion. This study was stimulated by a series of papers [13, 19, 24]. In [24], six mechanisms for periodically forcing the classical predator-prey model with Holling type II response functions were shown, surprisingly, to have topologically equivalent 2-parameter bifurcation diagrams for the associated first return map with respect to fold, flip, and Neimark–Sacker bifurcation curves of the first and second iterates and period doubling cascades. Even more unexpectedly, in [13] it was shown that the eight mechanisms for periodically forcing the analogous predator-prey model in a chemostat not only produced topologically equivalent diagrams, but these diagrams were topologically equivalent to the ones for the classical model. They conjectured a "universal diagram" for forced predator-prey systems. In [24] the authors state explicitly that it would be of interest to extend their analysis to a predator-prey model that has saddle-node bifurcations of limit cycles and homoclinic bifurcations for the unforced system. System (1.3) seemed like an ideal choice, since we were aware from previous studies [12, 30, 25, 26, 28] that it had the indicated bifurcations as the carrying capacity was varied. We felt certain that it should be possible to obtain 2-parameter bifurcation diagrams of the periodically forced version of system (1.3) that were not topologically equivalent to the ones in [13] and [24].

To our surprise, when the carrying capacity was forced, the 2-parameter bifurcation diagram was not significantly different [32]. In order to find parameters that would produce different diagrams, it was necessary to perform a more detailed bifurcation analysis on the unforced system, i.e., system (1.3), in order to understand the role of all of the parameters, and this was the motivation for this paper.

In fact, based on the results in this paper, in [32], we are able to show that for the Holling type IV response functions, different mechanisms for periodic forcing result in topologically distinct 2-parameter bifurcation diagrams and hence can be different than the postulated universal diagram.

Our work in this paper, analyzing local and global bifurcations of system (1.3), extends and complements the work in [12, 30, 25, 26, 28]. We now understand the role of perturbing each of the parameters on the dynamics. We described, for any fixed $a > 0$, $-2\sqrt{a} < b$, and $K > 0$, the sequence of bifurcations and associated phase portraits which occur as $\frac{d}{cm} > 0$ is varied. These results are summarized in Figure 6.1 and Tables 6.1–6.5 and include both local and global bifurcations. In particular, explicit regions in parameter space are provided for all of the phase portraits illustrated in Table 6.2, including regions where there are at least two limit cycles. We showed that for any $a, r > 0$ and $-2\sqrt{a} < b$ there is a Bogdanov–Takens bifurcation of codimension 2 when $d = \frac{cm}{b+2\sqrt{a}}$ and $K = \frac{2}{\sqrt{a}}$ with $b \neq -\sqrt{a}$ and a Bogdanov–Takens bifurcation of codimension 3 when $b = -\sqrt{a}$, $d = \frac{cm}{\sqrt{a}}$, and $K = \frac{2}{\sqrt{a}}$. We proved that the parameters $b,\ d,\ K$ give a versal unfolding of the codimension 3 bifurcation.

Although the model (1.3) contains seven parameters, our analysis shows that the parameter r has no effect on the existence and stability of equilibria, or of limit cycles created by Hopf bifurcations. Further, as seen in the results above, the parameters $c,\ d$, and m always occur in the ratio $\frac{d}{cm}$, and the parameter a acts as a scaling factor for b and K. These latter relations are not surprising given that $a,\ c$, and m could be removed via the rescaling (1.7), but these relations seem to indicate that this is the most "natural" way to reduce the seven parameters to four.

As discussed above, variation of the parameter d, the death rate of the predator, results in many different bifurcation sequences depending on the values of the other parameters. However, there is a common theme to all the sequences. For $d > 0$ small enough, any system starting with positive initial conditions will lead to coexistence of

the predator and the prey. For d large enough (i.e., $d > \frac{mc}{b+2\sqrt{a}}$), all initial conditions will result in extinction of the predator. In between there exists a range of values of d for which either coexistence or extinction of the predator can occur depending on the initial conditions. Most notably, if the initial prey population is large enough, then extinction of the predator will result regardless of the initial size of the predator population.

Our analysis allows us to describe the tremendous variation in the bifurcation sequences. In particular, we have shown that the coexistence of the predator and the prey can be in the form of a steady state or periodic solution, or, for some sets of parameters, both. We have not been able to exclude the possibility of further variation due to limit cycles appearing and disappearing in global bifurcations. Such variations can lead to biologically interesting sequences of bifurcations. One example, which we have observed numerically, discussed and listed in Table 6.6, involves two homoclinic bifurcations that occur in succession as d is increased. The result is that the system goes from a state where coexistence is possible to one where it is not and then back again. This gives rise to the surprising result that, for this set of parameters, *increasing* the per capita death rate of the predator actually *increases* the predator population's chance of survival (or analogously, *reducing* the per capita death rate of the predator *reduces* the predator population's chances of survival).

REFERENCES

[1] J. F. Andrews, *A mathematical model for the continuous culture of microorganisms utilizing inhibitory substrates*, Biotechnol. Bioeng., 10 (1968), pp. 707–723.

[2] A. Andronov, E. Leontovich, I. Gordon, and A. Maier, *Theory of Bifurcations of Dynamical Systems on a Plane*, Israel Program for Scientific Translations, Halstead Press, Jerusalem, 1971.

[3] A. D. Bazykin, *Nonlinear Dynamics of Interacting Populations*, World Sci. Ser. Nonlinear Sci. Ser. A Monogr. Treatises 11, World Scientific, River Edge, NJ, 1998.

[4] K.-S. Cheng, *Uniqueness of a limit cycle for a predator-prey system*, SIAM J. Math. Anal., 12 (1981), pp. 541–548.

[5] S. N. Chow and J. K. Hale, *Methods of Bifurcation Theory*, Grundlehren Math. Wiss. 251, Springer–Verlag, New York-Berlin, 1982.

[6] F. Dumortier, *Singularities of Vector Fields*, Monogr. Mat. 32, Instituto de Matemática Pura e Aplicada, Rio de Janeiro, 1978.

[7] F. Dumortier and R. Roussarie, *Étude locale des champs de vecteurs à paramètres*, in Journées Singuliéres de Dijon (Univ. Dijon, Dijon, 1978), Astérisque 59–60, Soc. Math. France, Paris, 1978, pp. 3, 7–42 (in French).

[8] F. Dumortier, R. Roussarie, and J. Sotomayor, *Generic 3-parameter families of vector fields on the plane, unfolding a singularity with nilpotent linear part. The cusp case of codimension 3*, Ergodic Theory Dynam. Systems, 7 (1987), pp. 375–413.

[9] F. Dumortier, R. Roussarie, and C. Rousseau, *Hilbert's 16th problem for quadratic vector fields*, J. Differential Equations, 110 (1994), pp. 86–133.

[10] F. Dumortier, R. Roussarie, J. Sotomayor, and H. Żoladek, *Bifurcations of Planar Vector Fields. Nilpotent Singularities and Abelian Integrals*, Lecture Notes in Math. 1480, Springer–Verlag, Berlin, 1991.

[11] B. Ermentrout, *Simulating, Analyzing, and Animating Dynamical Systems: A guide to XPPAUT for researchers and students*, SIAM, Philadelphia, 2002.

[12] H. I. Freedman and G. S. K. Wolkowicz, *Predator-prey systems with group defence: The paradox of enrichment revisited*, Bull. Math. Biol., 48 (1986), pp. 493–508.

[13] A. Gragnani and S. Rinaldi, *A universal bifurcation diagram for seasonally perturbed predator-prey models*, Bull. Math. Biol., 57 (1995), pp. 701–712.

[14] C. S. Holling, *The components of predation as revealed by a study of small-mammal predation of the European pine sawfly*, Can. Entomol., 91 (1959), pp. 293–320.

[15] C. S. Holling, *The functional response of predators to prey density and its role in mimicry and population regulation*, Mem. Entomolog. Soc. Canada, 45 (1965), pp. 3–60.

[16] N. Kasarinoff and P. van der Deiesch, *A model of predator-prey system with functional response*, Math. Biosci., 39 (1978), pp. 124–134.
[17] M. Kot, *Elements of Mathematical Ecology*, Cambridge University Press, Cambridge, UK, 2001.
[18] Y. A. Kuznetsov, *Elements of Applied Bifurcation Theory*, Appl. Math. Sci. 112, Springer–Verlag, New York, 1995.
[19] Y. A. Kuznetsov, S. Muratori, and S. Rinaldi, *Bifurcations and chaos in a periodic predator-prey model*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 2 (1992), pp. 117–128.
[20] Y. A. Kuznetsov, *CONTENT-Integrated Environment for Analysis of Dynamical Systems*, Report UPMA-98-224, Ecole Normale Supericure de Lyon, Lyon, France, 1998.
[21] P. Mardešić, *Chebyshev Systems and the Versal Unfolding of the Cusps of Order n*, Travaux en Cours 57, Hermann, Paris, 1998.
[22] J. E. Marsden and M. McCracken, *The Hopf Bifurcation and Its Applications*, Appl. Math. Sci. 19, Springer–Verlag, New York, 1976.
[23] A. Mourtada, *Cyclicité finie des polycycles hyperboliques de champs de vecteurs du plan: mise sous forme normale*, in Bifurcations of Planar Vector Fields (Luminy, 1989), Lecture Notes in Math. 1455, Springer–Verlag, Berlin, 1990, pp. 272–314.
[24] S. Rinaldi, S. Muratori, and Y. A. Kuznetsov, *Multiple attractors, catastrophes and chaos in seasonally perturbed predator-prey communities*, Bull. Math. Biol., 55 (1993), pp. 15–35.
[25] F. Rothe and D. S. Shafer, *Bifurcation in a quartic polynomial system arising in biology*, in Bifurcations of Planar Vector Fields (Luminy, 1989), Lecture Notes in Math. 1455, Springer–Verlag, Berlin, 1990, pp. 356–368.
[26] F. Rothe and D. S. Shafer, *Multiple bifurcation in a predator-prey system with nonmonotonic predator response*, Proc. Roy. Soc. Edinburgh Sect. A, 120 (1992), pp. 313–347.
[27] R. Roussarie, *A note on finite cyclicity property and Hilbert's* 16*th problem*, in Dynamical Systems, Valparaiso 1986, Lecture Notes in Math. 1331, Springer–Verlag, Berlin-New York, 1988, pp. 161–168
[28] S. Ruan and D. Xiao, *Global analysis in a predator-prey system with nonmonotonic functional response*, SIAM J. Appl. Math., 61 (2001), pp. 1445–1472.
[29] *Waterloo Maple Software* V, University of Waterloo, Waterloo, Canada, 1990.
[30] G. S. K. Wolkowicz, *Bifurcation analysis of a predator-prey system involving group defence*, SIAM J. Appl. Math., 48 (1988), pp. 592–606.
[31] D. Wrzosek, *Limit cycles in predator-prey models*, Math. Biosci., 98 (1990), pp. 1–12.
[32] H. Zhu, S. A. Campbell, and G. S. Wolkowicz, *Seasonal Forcing of a Predator-Prey Model with Nonmonotonic Functional Response*, in preparation.

SIAM J. APPL. MATH.
Vol. 63, No. 2, pp. 683–707

ON A DOUBLY NONLINEAR PARABOLIC OBSTACLE PROBLEM MODELLING ICE SHEET DYNAMICS*

N. CALVO†, J. I. DÍAZ‡, J. DURANY†, E. SCHIAVI§, AND C. VÁZQUEZ¶

Abstract. This paper deals with the weak formulation of a free (moving) boundary problem arising in theoretical glaciology. Considering shallow ice sheet flow, we present the mathematical analysis and the numerical solution of the second order nonlinear degenerate parabolic equation modelling, in the isothermal case, the ice sheet non-Newtonian dynamics. An obstacle problem is then deduced and analyzed. The existence of a free boundary generated by the support of the solution is proved and its location and evolution are qualitatively described by using a comparison principle and an energy method. Then the solutions are numerically computed with a method of characteristics and a duality algorithm to deal with the resulting variational inequalities. The weak framework we introduce and its analysis (both qualitative and numerical) are not restricted to the simple physics of the ice sheet model we consider nor to the model dimension; they can be successfully applied to more realistic and sophisticated models related to other geophysical settings.

Key words. ice sheet models, nonlinear degenerate equations, free boundaries, weak solutions, finite elements, duality methods

AMS subject classifications. 35K65, 35K85, 65C20, 65N30

PII. S0036139901385345

1. Introduction.

Modelling ice sheet dynamics has been a challenging problem since the beginning of the century, but nowadays the scientific community is showing a renewed, growing interest towards this problem. In fact, our understanding of climate system dynamics depends on the comprehension and predictability of the ice sheet dynamics. Large ice sheets influence and are influenced by climate, and their oscillations may be responsible for sudden shifts in climate in the recent geological past (Fowler [30]). The study of ice sheet models (ISM) is fundamental to the construction and comprehension of global energy balance models (EBM) and general circulation models (GCM) (see, for instance, Tarasov and Peltier [38] for a coupled ISM/EBM model). Various physically based theories have appeared during the last decades in order to explain the flow of these large ice masses, but a proper mathematical treatment is not available yet. This introduces the main aim of this paper, which is to present the mathematical and numerical analysis of an obstacle problem formulation for the study of the slow, isothermal, one-dimensional flow of ice along a rigid impermeable bed. This paper is organized as follows: after a brief description of the model equation and its strong formulation (section 2), we introduce in section 3 some weak formulations that we shall use later. The well-posedness of the model is then considered. Section 4 is devoted to the (qualitative) study of the free (moving)

*Received by the editors February 16, 2001; accepted for publication (in revised form) June 4, 2002; published electronically December 19, 2002. Partially supported by Research Projects of the D.G.E.S. (REN2000-0766) and M.C.Y.T. (BFM2001-3261-C02).
http://www.siam.org/journals/siap/63-2/38534.html

†Departamento de Matemática Aplicada, E.T.S.I. de Telecomunicaciones, Universidade de Vigo, Campus Marcosende s/n, 36280-Vigo, Spain (nati@dma.uvigo.es, durany@dma.uvigo.es).

‡Departamento de Matemática Aplicada, Facultad de Matemáticas, Universidad Complutense de Madrid, 28004-Madrid, Spain (ji_diaz@mat.ucm.es).

§Departamento de Ciencias Experimentales e Ingeniería, E.S.C.E.T. Universidad Rey Juan Carlos, E 28933, Móstoles, Madrid, Spain (eschiavi@escet.urjc.es).

¶Departamento de Matemáticas, Facultad de Informática, Universidade da Coruña, Campus Elviña s/n, 15071-A Coruña, Spain (carlosv@udc.es).

boundary defined by the model. The quantitative analysis is done in section 5, where we numerically solve the problem by using, among other techniques, the algorithm proposed in Bermúdez and Moreno [8]. Some numerical tests are then performed, and additional information is provided by the consideration of specific real data. Finally, in section 6, we discuss our results and their scope.

2. Model equation and strong formulation. The model equation is the one proposed in Fowler [29], describing the evolution of the thickness $h(t,x)$ for a two-dimensional plane ice sheet. ($z = h(t,x)$ is the top surface of the ice sheet.) Ice is taken to be incompressible, and the flow is very slow. It flows as a viscous medium under its own weight, owing to gravity. A number of additional simplifying assumptions are used in the derivation of the model we consider (isothermal flow, shallow ice approximation, a flat, rigid, and impermeable bed, etc.). We refer to [29] for details on the modelling; for a more general introduction to glaciology, see, for instance, Hutter [32], Paterson [36], and Lliboutry [33]. According Fowler [29], the local thickness of the ice sheet satisfies the following nonlinear diffusion equation:

$$h_t = \left(\frac{h^{n+2}}{n+2} |h_x|^{n-1} h_x - u_b h \right)_x + a \quad \text{on } I(t), \tag{2.1}$$

where $a = a(t,x)$ is a given scaled fixed accumulation-ablation rate function ($a < 0$ signifies ablation) and u_b is a (given) function representing the basal velocity. For each fixed t the domain $I(t)$ represents the (unknown) bounded real interval where $h(t,x) > 0$ (i.e., $I(t) := \{x \,/\, h(x,t) > 0\}$). Notice that the physically relevant rate functions $a(t,x)$ are changing sign functions, which are positive in the central (accumulation) region of the ice sheet and negative near the margins (the boundaries of $I(t)$, i.e., in the ablation region); see Fowler [30, p. 95]. The exponent n that appears in (2.1) represents the so-called Glen's exponent, and it is usually assumed that $n \approx 3$ (see Fowler [29]). We shall assume $n = 3$, but the qualitative analysis remains unchanged for any $n > 1$ (non-Newtonian case). As regards the appropriate (mechanical) boundary condition, it depends on the thermal regime which we consider at the base. There are two possible geophysical situations corresponding to slip or no slip conditions. We shall consider both of them, generalizing in this way Fowler's approach.

When basal ice reaches the melting point, there is a net heat flux arriving at the bed of the ice sheet, and consequently basal melt water is produced: the ice begins to slide. Sliding is expected only where the basal ice is at the melting point. When u_b (the sliding velocity) is a prescribed function of (t,x) (i.e., $u_b = u_b(t,x)$), this equation is a nonlinear diffusion-convection equation for h. It corresponds to slip conditions along an assumed temperate bed (warm-based ice sheet). Once the base reaches the melting point, we assume that the ice above remains cold. Our aim is to show how it is possible to solve this model for a general, prescribed velocity field. For a slow shallow flow over a flat cold base, there is no sliding (i.e., $u_b \equiv 0$), and the isothermal ice sheet equation (2.1) becomes just

$$h_t = \left(\frac{h^{n+2}}{n+2} |h_x|^{n-1} h_x \right)_x + a \quad \text{on } I(t). \tag{2.2}$$

We shall refer to (2.2) as the pure diffusive case. As discussed in Fowler [29], when a cold-based flow regime is considered and the no slip condition is prescribed (due to infinite slope), singularities appear at the margins during the advance of fronts of

a land-based ice sheet (such as the one which covered North America in the last ice age). Classical (finite-differences) numerical methods can fail. A further complication arises due to the fact that the domain where the equation holds is unknown. In fact it has to be determined as part of the solution.

The original *strong* formulation can be stated in the following terms: let $T > 0$, $L > 0$ be positive fixed real numbers, and let $\Omega = (-L, L)$ be an open bounded interval of $\mathbb{R}$ (a sufficiently large, fixed spatial domain). Given an accumulation-ablation rate function $a = a(t, x)$, an eventually zero sliding velocity function $u_b = u(t, x)$ defined on a large fixed parabolic domain $Q = (0, T) \times (-L, L)$, and an initial thickness $h_0 = h_0(x) \geq 0$ (bounded and compactly supported) on Ω, find two curves $S_+, S_- \in C^0([0, T])$, with $S_-(t) \leq S_+(t)$, a set $I(t) := (S_-(t), S_+(t)) \subset \Omega$ for any $t \in [0, T]$, and a sufficiently smooth function $h(t, x)$ defined on the set $Q_T := \bigcup_{t\in(0,T)} I(t)$ such that

$$(2.3) \qquad (\mathrm{SF}) := \begin{cases} h_t = \left[\dfrac{h^{n+2}}{n+2}|h_x|^{n-1}h_x - u_b h\right]_x + a & \text{in} \quad Q_T, \\ h = 0 \quad \text{on} \quad \{S_-(t)\} \cup \{S_+(t)\}, & t \in (0, T), \\ h = h_0 & \text{on} \quad I(0), \end{cases}$$

and $h(t, x) > 0$ on Q_T. Notice that, for each fixed $t \in (0, T)$, $I(t) = (S_-(t), S_+(t)) = \{x \in \Omega : h(t, x) > 0\}$ denotes the ice-covered region. The curves $S_\pm(t)$ are called the interface curves or free boundaries associated with the problem and are defined by

$$S_-(t) = \mathrm{Inf}\{x \in \Omega : h(t, x) > 0\}, \qquad S_+(t) = \mathrm{Sup}\{x \in \Omega : h(t, x) > 0\}.$$

These curves define the interface separating the regions in which $h(t, x) > 0$ (i.e, ice regions) from those in which $h(t, x) = 0$ (i.e., ice-free regions). In the physical context they represent the propagation fronts of the ice sheet. The above formulation needs two different refinements. First, we have to prescribe some additional information on the spatial derivative of h at the free boundary. (We shall assume that the ice flux is zero there; see (3.2).) To introduce the other refinement it is useful to recall that many other examples of degenerate equations are typical of slow phenomena and satisfy the finite speed of propagation property (see, e.g., Díaz [17]).

Assuming $a \equiv 0$, for instance, if h_0 has a compact support, then $h(t, .)$ also has a compact support in $\mathbb{R}$ for any $t \in [0, T]$. So, if $a \equiv 0$, the domain Q_T can be found through the support of the solution $h(t, x)$ of the doubly nonlinear parabolic equation over the whole space $(0, T) \times \mathbb{R}$ and satisfying the initial condition $h(0, x) = h_0(x)$, $x \in \mathbb{R}$. Unfortunately, the physically relevant case, $a \not\equiv 0$, is much more complicated. Indeed, the finite speed of propagation still holds if $a(t, .)$ has compact support in $\mathbb{R}$ (for fixed $t \in (0, T)$). Moreover, in that case it can be shown that $\mathrm{supp}(h(t, .)) \subset \mathrm{supp}(a(t, .))$, and therefore $a(t, .)$ vanishes on the free boundary. Nevertheless, in glaciology models it is well known (see Fowler [30, p. 95]) that $a(t, .) < 0$ near the free boundaries (i.e., the margins of the ice sheet), and so there must exist another reason (other than the degenerate character of the equation) justifying the occurrence of the free boundaries $S_-(t), S_+(t)$. This is a mathematical modelling problem. We must make sure that the mathematical solutions are nonnegative compactly supported solutions (i.e., physically admissible). In short, for a sufficiently large fixed spatial domain, the physically admissible solutions are compactly supported nonnegative bounded functions such that $a < 0$, where $h = 0$ (in

particular, in the free boundary), and this is not predicted by the solutions of the diffusion equation (despite its degenerate character). Mathematically it is then possible (for special choices of the accumulation/ablation rate function) to have negative (no physically admissible) solutions corresponding to negative thickness! A practical way to overcome these inconsistencies is proposed in the following section.

3. Weak formulations. In this section we show that proper mathematical modelling in terms of weak formulations of the physical problem must be considered if the assumed physics have to be respected (i.e., if just physically admissible solutions have to be computed).

Let T, L, and Ω be as before and set $Q := (0,T) \times \Omega \subset \mathbb{R}^2$. The new model we present is based upon the fact that we can extend the function $h(t,x)$ outside of Q_T (the ice-covered regions) by zero on $Q \setminus Q_T = [(0,T) \times \Omega] \setminus Q_T$ (the complementary ice-free region). This extension still satisfies a nonlinear equation (this time of multivalued type) having the great advantage of being defined on an a priori known domain $Q = (0,T) \times \Omega$ (whose parabolic boundary is $\Sigma = \partial Q = (0,T) \times \partial\Omega$). This type of problem is known in the literature as an *obstacle problem*. In our case the obstacle function is $\psi \equiv 0$, the null function. Obstacle problems arise in many contexts related to friction, elasticity, thermodynamics, and so on (see, e.g., Duvaut and Lions [27] for further details). The multivalued formulation we propose appeared first in Díaz and Schiavi [21] (where the no slip condition was considered) to describe the slow, isothermal, one-dimensional flow of *cold* ice (i.e., all the ice is below melting point and the melting point is reached only at the bed) along a *hard* (i.e., rigid, impermeable) bed. Our results can be generalized to deal with the two-dimensional case that describes the evolution of a three-dimensional ice sheet. Here we extend that model to consider (prescribed) sliding along a temperate base. This introduces a nonlinear convective term into the multivalued equation which describes the movement of the ice masses. In order to properly characterize the behavior of h near the free boundary, we assume that the ice flux is not singular in the sense that

$$\frac{h^{n+2}}{n+2}|h_x|^{n-1}h_x - u_b h \in L^1(Q). \tag{3.1}$$

Notice that, formally, this implies

$$(h^m)_x = 0 \qquad \text{on} \quad \{S_-(t)\} \cup \{S_+(t)\}, \quad t \in (0,T), \tag{3.2}$$

where $m = 2(n+1)/n$. Introducing the maximal monotone graph of $\mathbb{R}^2$, β, defined by

$$\beta(r) = \emptyset \text{ (the empty set) if } \quad r < 0, \qquad \beta(0) = (-\infty, 0], \qquad \beta(r) = 0 \text{ if } \quad r > 0, \tag{3.3}$$

the *obstacle formulation* (written in terms of a multivalued equation) is the following: given a bounded, sufficiently large interval $\Omega = (-L, L) \subset \mathbb{R}$, a rate function $a \in L^\infty(Q)$, a sliding velocity $u_b \in L^\infty(Q)$, and a compactly supported initial data $h_0 \in L^\infty(\Omega)$, find a sufficiently smooth function $h(t,x)$ which is a solution of

$$(\mathrm{MF}) := \begin{cases} h_t - \left(\dfrac{h^{n+2}}{n+2}|h_x|^{n-1}h_x - u_b h\right)_x + \beta(h) \ni a(t,x) & \text{in} \quad Q, \\ h(t,x) = 0 & \text{on} \quad \Sigma, \\ h(0,x) = h_0(x) & \text{on} \quad \Omega. \end{cases} \tag{3.4}$$

Notice that β is multivalued just where h is zero, i.e., at the free boundaries. Moreover, by definition (3.3), we have that $0 \in \beta(0)$. Now, let h be a solution (in a weak sense to be specified later) of (3.4), for almost every $x \in \Omega$ and $\forall\, t \in (0,T)$. It is clear that in the null set $Q \setminus Q_T$ we must have $\beta(0) \ni a(t,x)$. This condition shows that, if β is multivalued at the origin, then it is possible to have solutions with a nonempty null set (i.e., $Q \setminus Q_T \neq \emptyset$) corresponding to equations in which $a \neq 0$ on $Q \setminus Q_T$, and thus new results are possible with respect to the single valued case ($\beta \equiv 0$).

Details on this kind of (multivalued) formulations, (MF), and maximal monotone graphs can be found in Brezis [11]. It is well known that the multivalued equation (3.4) can be written in terms of the so-called *complementarity formulation* for obstacle problems, which states: given Ω, a, u_b, and h_0 as before, find a sufficiently smooth function h such that

$$
\text{(CF)} := \begin{cases}
h_t - \left(\dfrac{h^{n+2}}{n+2} |h_x|^{n-1} h_x - u_b h \right)_x - a \geq 0 & \text{in} \quad Q, \\
\left[h_t - \left(\dfrac{h^{n+2}}{n+2} |h_x|^{n-1} h_x - u_b h \right)_x - a \right] h = 0 & \text{in} \quad Q, \\
h \geq 0 & \text{in} \quad Q, \\
h = 0 & \text{on} \quad \Sigma, \\
h = h_0(x) & \text{on} \quad \Omega.
\end{cases} \tag{3.5}
$$

It is obvious that if a regular function h verifies the strong formulation, then its extension by zero over $Q \setminus Q_T$ (which we will denote again by h) satisfies trivially the complementarity formulation, assuming that $a(t,x)$ satisfies the condition

$$
a(t,x) \leq 0 \qquad \text{on} \quad Q \setminus Q_T.
$$

A more general framework is obtained if we define $\phi(r) = |r|^{n-1} r$, $r \in \mathbb{R}$, $n > 1$, and $\psi(s) = s^m$, with $s \geq 0$ and $m = 2(n+1)/n > 1$. (In fact, the existence and the uniqueness of solutions and some qualitative properties remain true if we replace ϕ by any real continuous strictly increasing convex function such that $\phi(0) = 0$, and β as before; see (3.3).) Introducing the new variable $u = u(t,x)$ and the real function $b(s)$ in form

$$
u := h^m = \psi(h) \qquad \Longrightarrow \qquad u^{1/m} = h = \psi^{-1}(u) := b(u), \tag{3.6}
$$

we have $\phi(\psi(h)_x) = \phi(u_x) = |u_x|^{p-2} u_x$, where $p = n+1$. The previous *multivalued formulation* is the following: given Ω, a, u_b, and $u_0 = \psi(h_0)$ as before and a constant $\mu = n^n/[2^n (n+1)^n (n+2)]$, determine a function $u(t,x) = \psi(h(t,x))$ which is the solution of

$$
\text{(GF)} := \begin{cases}
b(u)_t - [\mu \phi(u_x) - u_b b(u)]_x + \beta(u) \ni a(t,x) & \text{in} \quad Q, \\
u(t,x) = 0 & \text{on} \quad \Sigma, \\
b(u(0,x)) = b(u_0(x)) & \text{on} \quad \Omega.
\end{cases} \tag{3.7}
$$

This *general formulation*, (GF), is the one that we shall use to deal with the well-posedness of the model problem (3.4). Notice that we can write $\beta(u)$ instead of $\beta(b(u))$ because $\beta(u) \equiv \beta(h) := \beta(b(u))$ in Q. The equivalence is readily understood, observing that h (i.e., the original variable) and u have exactly the same support. The same remark applies to the boundary condition on Σ.

3.1. On the existence and uniqueness of weak solutions. Problem (3.7) admits various notions of solutions according to the required spatial and time regularity. In any case we must start by assuming some regularity on the data $a(t,x)$, $u_b(t,x)$, and $u_0(x)$. In our case it will be enough to assume that

$$a \in L^\infty(Q), \qquad u_b \in L^\infty(0,T : W^{1,\infty}(\Omega)), \qquad \text{and} \quad u_0 \in L^\infty(\Omega). \tag{3.8}$$

Motivated by (3.1), a natural notion of weak solution is the following.

DEFINITION 3.1. *A function $u \in L^1(Q)$ is a weak solution of* (3.7) *if $u \in L^p(0,T : W_0^{1,p}(\Omega))$, $b(u) \in L^1(Q)$, $u(t,x) \geq 0$ a.e. $(t,x) \in Q$, and there exists a function $j \in L^1(Q)$ such that $j(x,t) \in \beta(u(t,x))$ a.e. $(t,x) \in Q$ and*

$$\int_Q (\zeta_t b(u) - \zeta j + \zeta a) dxdt + \int_\Omega \zeta(0,\cdot) b(u_0) dx = \int_Q \zeta_x [\mu \phi(u_x) - u_b b(u)] dxdt$$

for any $\zeta \in L^p(0,T : W_0^{1,p}(\Omega)) \cap L^\infty(Q)$, $\zeta_t \in L^\infty(Q)$, and $\zeta(T,\cdot) = 0$.

Assuming (3.8) and that b, ϕ are the power functions indicated above, the existence of a weak solution can be obtained by different methods, for instance, by an easy modification of Theorem 2.3 and Proposition 3.2 of Benilan and Wittbold [4]. In fact, in order to check assumption (H1) of [4] it is useful to replace function b by its truncation

$$b_M(r) := \begin{cases} b(r) & \text{if } r \in [0,M], \\ b(M) & \text{if } r \in [M,+\infty), \end{cases}$$

with $M > 0$ an upper bound of any weak solution. (We shall come back to this point later; see Proposition 3.3.) Proving the uniqueness of (and the comparison principle for) weak solutions is a more delicate task due to the presence of the nonlinear term $b(u)$. This type of result is well known (see, for instance, Díaz and de Thelin [18]) in the case in which we additionally know that the weak solution is differentiable with respect to time in the sense that $b(u)_t \in L^1(Q)$. (We recall that from the definition of weak solutions we know merely that $b(u)_t \in L^{p'}(0,T : W^{-1,p'}(\Omega)) + L^1(Q)$ with $p' := p/(p-1)$.) In order to get such results, a weaker notion was introduced in previous works by different authors (see Boccardo et al. [10] for the case of $b(u) = u$, and Carrillo and Wittbold [15] for a general nondecreasing function $b(u)$): the notion of a *renormalized solution*, coming originally from a different context (Di Perna and Lions [24]). In fact both notions coincide in the class of bounded functions $u \in L^\infty(Q)$, which is our case, as we shall prove in this section.

For the sake of simplicity in the exposition we assume that

$$u_b(t,\cdot) \quad \text{is spatially constant.} \tag{3.9}$$

So, by some trivial modifications of the results of Carrillo and Wittbold [15] we arrive at the following result.

THEOREM 3.2. *Assume a_i, u_b, and $u_{0,i}$ satisfying* (3.8) *and* (3.9) *for $i = 1,2$. Let u_i be weak solutions of problem* (3.7) *associated with the data a_i and $u_{0,i}$, respectively. Then for any $t \in [0,T]$*

$$\int_\Omega [b(u_1(t,\cdot)) - b(u_2(t,\cdot))]_+ dx \leq \int_\Omega [b(u_{0,1}) - b(u_{0,2})]_+ dx + \int_0^t \int_\Omega [a_1(s,x) - a_2(s,x)]_+ dxds,$$

where $[f]_+ = \max(f,0)$. In particular, $b(u_{0,1}) \leq b(u_{0,2})$ and $a_1(t,x) \leq a_2(t,x)$, on their respective domains of definition, implies that $b(u_1(t,x)) \leq b(u_2(t,x))$ for any

$t \in [0,T]$ *and a.e.* $x \in \Omega$. *Consequently, there is at most one weak solution of problem* (3.7).

We point out that the above comparison remains true even if the functions u_i are not homogeneous at the boundary but satisfy $u_1(t,x) \le u_2(t,x)$ for any $t \in [0,T]$ and a.e. $x \in \partial\Omega$. This is again a trivial modification of the result by Carrillo and Wittbold [15], which will be used in the following section.

The boundedness of the associated weak solutions can be deduced from the above comparison principle as follows.

PROPOSITION 3.3. *Let* u *be any weak solution of problem* (3.7)*; then*

$$\|u\|_{L^\infty(Q)} \le M_0, \tag{3.10}$$

with

$$M_0 := b^{-1}\left(\max\{\|b(u_0)\|_{L^\infty(\Omega)}, 1\} \exp T\left\{\left[\frac{\|a\|_{L^\infty(Q)}}{\max\{\|b(u_0)\|_{L^\infty(\Omega)}, 1\}} + \|(u_b)_x\|_{L^\infty(\Omega)}\right]\right\}\right).$$

Proof. We take as a candidate *supersolution* a spatially constant function of the form $u_2(t,x) := b^{-1}(Ce^{\lambda t})$ for some $C > 0$ and $\lambda \in \mathbb{R}$ to be determined. Then $u_1(t,x) \le u_2(t,x)$ for any $t \in [0,T]$ and a.e. $x \in \partial\Omega$ and $b(u_{0,1}) \le b(u_{0,2})$ holds if

$$C = \max\{\|b(u_0)\|_{L^\infty(\Omega)}, 1\}.$$

Finally, by substituting u by u_2 in (3.7), it is easy to check that

$$a_2(t,x) := \lambda C e^{\lambda t} + (u_b)_x(x) C e^{\lambda t},$$

and so condition $a_1(t,x) \le a_2(t,x)$ is satisfied if, for instance,

$$\lambda = C^{-1}\|a\|_{L^\infty(Q)} + \|(u_b)_x\|_{L^\infty(\Omega)},$$

which implies the result. □

We point out that although the application of the present version of the comparison principle given in the above theorem requires condition (3.9), the a priori estimate (3.10) can be obtained without using a comparison principle (see, for instance formula (13) of Benilan and Wittbold [4]), and so the boundedness of u remains true also when $(u_b)_x \ne 0$, as we shall consider later.

4. On the free boundary. In this section we shall study both thermal regimes at the base. In the first case the bed is assumed to be cold (below melting point). No sliding is prescribed (i.e., $u_b \equiv 0$), and the pure diffusive case is analyzed. Next, we assume the ice sheet to be warm-based; the bed is then temperate, and sliding is prescribed (i.e., $u_b = u_b(t,x)$). Here we are not concerned with the switching mechanism between cold-temperate dynamics. (Results in that direction can be found in Fowler and Schiavi [31] and Díaz and Schiavi [22].) Our aim is to qualitatively describe the behavior of the free boundaries by means of a priori estimates on the support of the solution.

4.1. The no slip condition (pure diffusive case): Existence of the free boundary and the waiting time property. In this section we shall prove the existence of a nonempty null set

$$N(h(t,\cdot)) := \left\{(t,x) \in \{t\} \times \frac{\Omega}{h(t,x)} = 0\right\}$$

for the (unique) solution $h(t,x)$ of the problem

$$\text{(4.1)}\qquad \begin{cases} h_t - \mu\phi(\psi(h)_x)_x + \beta(h) \ni a(t,x) & \text{in} \quad Q, \\ h(t,x) = 0 & \text{on} \quad \Sigma, \\ h(0,x) = h_0(x) & \text{on} \quad \Omega, \end{cases}$$

which can be deduced from the general formulation (GF) written in terms of the original variable h and of the functions ϕ and ψ introduced before. Assuming extra regularity of the solution (i.e, $h \in C(\bar{Q})$), we are able to analyze a great number of geophysical phenomena related to location and evolution of the free boundary and associated with the behavior of the function $a(t,x)$.

We shall now deal with the existence and location of the free boundary defined by problem (4.1). To show the existence of a free boundary as well as to locally estimate its location, we will use a technique based on the comparison result of section (3.1), which consists of the construction of appropriate local super-sub solutions having compact support. Thus, for all $\epsilon > 0$, we define the set

$$\text{(4.2)}\qquad N_\epsilon(a(t,\cdot)) := \left\{ (t,x) \in \{t\} \times \frac{\Omega}{a(t,x)} \leq -\epsilon \right\}$$

and the set $S_\epsilon(a(t,.)) = Q \setminus N_\epsilon(a(t,\cdot)$. Then we have the following result.

THEOREM 4.1. *Let $h \in C(\bar{Q})$, $h \geq 0$, be a solution of* (4.1)*, and let ϵ be a small real positive number such that the set $N_\epsilon(a(t,.))$ is not empty. Then there exist $R > 0$ and $T_0 \geq 0$ such that $\forall t \geq T_0$ we have*

$$N(h(t,\cdot)) \supset \{(t_0,x_0) \in N_\epsilon(a(t_0,.)) : d(x_0, S_\epsilon(a(t,.))) \geq R\}.$$

Proof. The proof is based on an original idea of Evans and Knerr [28], which applies when $n = 1$ and $a(t,x) \equiv 0$. See also Díaz and Hernández [20] for its adaptation to the case $n > 1$. In our multivalued case, with $n > 1$ but $a(t,x) \not\equiv 0$, we argue as follows. We consider the set $N_\epsilon(a(t,.))$ and define the function

$$\tilde{h}(t,x) = \psi^{-1}\left(\eta(|x - x_0|) + \psi(U(t))\right),$$

where

$$\text{(4.3)}\qquad \eta(r) = c\, r^{\frac{p}{p-1}}, \qquad c = \frac{p-1}{p}\left(\frac{\epsilon}{2}\right)^{\frac{1}{p-1}},$$

and $U(t)$ is the (unique) solution of the initial value problem

$$\text{(4.4)}\qquad \begin{cases} U' + \frac{1}{2}\beta(U) & \ni & -\frac{\epsilon}{2}, \\ U(0) & = & ||h_0||_{L^\infty(\Omega)}. \end{cases}$$

It is easy to state that $U(t) = [-\frac{\epsilon}{2}t + ||h_0||_{L^\infty}]^+$, whence

$$U(t) \equiv 0 \qquad \forall t \geq T_0 = \frac{2}{\epsilon}||h_0||_{L^\infty(\Omega)}.$$

On the other hand, as by construction $\phi(\psi(\tilde{h})_x)_x = \phi(\eta_x)_x = \epsilon/2$, we have (in $N_\epsilon(a(t,.))$)

$$\tilde{h}_t - \mu\phi(\psi(\tilde{h})_x)_x + \beta(\tilde{h})$$

$$\equiv \frac{d}{dt}\left[\psi^{-1}\left(\eta(|x-x_0|)+\psi(U(t))\right)\right]-\mu\phi(\eta_x(|x-x_0|))_x+\beta\left(\psi^{-1}\left(\eta(|x-x_0|)+\psi(U(t))\right)\right)$$

$$\supseteq \frac{\psi'(U)}{\psi'\left(\psi^{-1}\left(\eta(|x-x_0|)+\psi(U(t))\right)\right)}U'-\mu\phi(\eta_x(|x-x_0|))_x+\frac{1}{2}\beta(\psi^{-1}(\eta(|x-x_0|)))+\frac{1}{2}\beta(U)$$

$$\supseteq U'+\frac{1}{2}\beta(U)-\mu\phi(\eta_x(|x-x_0|))_x+\frac{1}{2}\beta\left(\psi^{-1}(\eta(|x-x_0|))\right)\ni-\epsilon\geq a(t,x).$$

Using the comparison principle written in terms of the original variable h, the following estimate holds (see Benilan and Wittbold [4]):

$$||h||_{L^\infty(Q)}\leq||h_0||_{L^\infty(\Omega)}+\int_0^t||a||_{L^\infty(\Omega)}=M(t).$$

Then

$$||h||_{L^\infty(Q)}\leq M(t)\leq\tilde{h}(t,\cdot)\qquad\text{on } N_\epsilon(a(t,\cdot))$$

iff $\psi^{-1}\left(\eta((|x-x_0|))+\psi(U(t))\right)\geq M(t)$; i.e., $\eta(|x-x_0|)+\psi(U(t))\geq\psi(M(t))$. In particular, this is true if $c\,|x-x_0|^{\frac{p}{p-1}}\geq\psi(M(t))$; by (4.3) the above reads

$$|x-x_0|\geq\frac{\psi(M(T))^{\frac{p-1}{p}}}{(\frac{p-1}{p})^{\frac{p-1}{p}}(\frac{\epsilon}{2})^{\frac{1}{p}}}=R,\tag{4.5}$$

and (4.5) implies that $\tilde{h}\geq h$ on $\partial N_\epsilon(a(t,.))$. At $t=0$ we use the monotonicity of ψ^{-1}:

$$\tilde{h}(0,x)=\psi^{-1}\left(\eta(|x-x_0|)+\psi(U(0))\right)=\psi^{-1}\left(\eta(|x-x_0|)+\psi(||h_0||_{L^\infty})\right)$$

$$\geq\psi^{-1}\left(\psi(||h_0||_{L^\infty})\right)=||h_0||_{L^\infty}\geq h_0(x)\geq 0.$$

Summarizing, if $(t,x)\in N_\epsilon(a(t,\cdot))$ is such that $|x-x_0|\geq R$, then

$$h_t-\mu\phi(\psi(h)_x)_x+\beta(h)\ni a\leq\inf\left(\tilde{h}_t-\mu\phi(\psi(\tilde{h})_x)_x+\beta(\tilde{h})\right)\qquad\text{in}\quad N_\epsilon(a(t,\cdot)),$$

$$h(t,x)\leq\tilde{h}(t,x)\qquad\text{on}\quad\partial N_\epsilon(a(t,\cdot)),$$

$$h_0(x)\leq\tilde{h}(0,x)\qquad\text{on}\quad N_\epsilon(a(0,\cdot)).$$

Next, from the comparison result (Theorem 3.2) it follows that

$$0\leq h(t,x)\leq\tilde{h}(t,x)\qquad\text{in}\quad N_\epsilon(a(t,.)),$$

and we end up observing that $h(t,x_0)=0\ \forall t\geq T_0=\frac{2}{\epsilon}||h_0||_{L^\infty}$ and x_0 satisfies inequality (4.5), i.e., $(t,x_0)\in\{N_\epsilon(a(t,.))/|x-x_0|\geq R\}$. □

We shall now analyze the so-called waiting time property. As discussed in Fowler [30, p. 95], the slope of the surface is singular in advance but finite in retreat. This distinction causes the degenerate diffusion equation above to have waiting-time behavior, because following a retreat, the margin slope must rebuild itself before another advance it possible. The following property applies if the initial data is sufficiently *flat* in the *ablation region.*

THEOREM 4.2. *Let $h \in C(\bar{Q})$, $h \geq 0$, be a solution of problem (4.1). Let $\delta = \eta^{-1}(\psi(M))$ and $B_\delta^+(x_0) = \{x \in \Omega / \ x \in [x_0, x_0 + \delta)\}$, with $M = ||h||_{L^\infty(Q)}$, $x_0 = S_+(0)$, $\tilde{c} = (\frac{p-1}{p})\epsilon^{\frac{1}{p-1}}$, and $\eta(|x - x_0|) = \tilde{c}\,|x - x_0|^{\frac{p}{p-1}}$. Assume that there exists $T^* > 0$ such that $a(t,x) \leq -\epsilon$ a.e. $x \in B_\delta^+(x_0)$ and $t \in (0, T^*)$. If $x_0 \in \Omega$ satisfies $0 \leq h_0(x_0) \leq \psi^{-1}(\eta(|x - x_0|))$, then*

$$\exists\, t^*, \quad 0 < t^* \leq T^*, \quad \textit{such that} \quad S_+(0) = S_+(t) \quad \forall t \in (0, t^*).$$

Proof. We define the function

$$\tilde{h}(x) := \psi^{-1}(\eta(|x - x_0|)) \qquad \text{in} \quad B_\delta^+(x_0) \times (0, T^*).$$

Then

$$h_t - \mu\phi(\psi(h)_x)_x + \beta(h) \ni a \leq -\epsilon \leq \inf\big(\tilde{h}_t - \mu\phi(\psi(\tilde{h})_x\big)_x + \beta(\tilde{h})) \quad \text{in } B_\delta^+(x_0) \times (0, T^*).$$

On $\partial B_\delta^+(x_0) \times [0, t^*]$ we have to verify that $h \leq M \leq \tilde{h} = \psi^{-1}(\eta)$, and this is iff $\psi(M) \leq \eta = \tilde{c}|x - x_0|^{\frac{p}{p-1}}$. On ∂B_δ^+ this reads as $\psi(M) \leq \tilde{c}\delta^{\frac{p}{p-1}}$. Using that $\delta = \eta^{-1}(\psi(M))$,

$$h \leq M \leq \tilde{h} \quad \Longleftrightarrow \quad \psi(M) \leq \tilde{c}[\eta^{-1}(\psi(M))]^{\frac{p}{p-1}}$$

$$\Longleftrightarrow \quad \left[\frac{\psi(M)}{\tilde{c}}\right]^{\frac{p-1}{p}} \leq \eta^{-1}(\psi(M)) \quad \Longleftrightarrow \quad \eta\left(\left[\frac{\psi(M)}{\tilde{c}}\right]^{\frac{p-1}{p}}\right) \leq \psi(M),$$

and this is always verified as can be deduced by applying the definition of the function η. Then we have

$$h_t - \mu\phi(\psi(h)_x)_x + \beta(h) \ni a \leq \inf(\tilde{h}_t - \mu\phi(\psi(\tilde{h})_x)_x + \beta(\tilde{h})) \text{ in } B_\delta^+(x_0) \times (0, t^*),$$

$$h(x_0, 0) = h_0(x_0) \leq \tilde{h}(x) = \psi^{-1}(\eta(|x - x_0|)) \quad \text{on } B_\delta^+(x_0),$$

$$h(t,x) \leq M \leq \tilde{h}(x) \quad \text{on } \partial B_\delta^+(x_0) \times (0, t^*).$$

Finally, the comparison result shows that $0 \leq h(t,x) \leq \tilde{h}(x)$, and so $h(t, x_0) \equiv 0$ $\forall t \in (0, t^*)$. ∎

4.2. The slip condition (diffusive-convective case): Existence of the free boundary and the waiting time property. In this section we shall consider the general formulation (GF), assuming that $u_b \not\equiv 0$ and without assuming (3.9). Even if Theorem 3.2 can be extended to cover cases in which (3.9) is not satisfied, the presence of the convection term makes the method of super- and subsolutions very hard to apply. Thus, in order to prove the existence of the free boundary, we shall use a different technique called the *energy method*. It has been developed by different authors in the last twenty years for the study of nonlinear problems for which the maximum principle fails (see, for instance, the monograph of Antontsev, Díaz, and Shmarev [2]). In fact, although this energy method can be applied in different ways, we shall follows the ideas introduced in Díaz and Galiano [19] in order to apply the method to some fluid dynamics problems. We start by pointing out that the equation of problem (3.7) can be written in terms of a nonconservative transport multivalued equation in the form

$$b(u)_t + u_b b(u)_x - \mu\phi(u_x)_x + (u_b)_x b(u) + \beta(u) \ni a(t,x) \qquad \text{in} \quad Q. \tag{4.6}$$

In this way, the equation involves the material derivative $b(u)_t + u_b b(u)_x$, which can be associate to a *virtual non-Newtonian fluid with a reactive term* $(u_b)_x b(u)+\beta(u)$. We shall prove the existence of the free boundary in terms of the so-called *finite speed of propagation* near a given point x_0.

In the next results we shall assume that u_b is a globally Lipschitz continuous function. Thus, we can define the characteristics of the associate flow by

$$\left\{ \begin{array}{rcll} \dfrac{d}{dt}X(t,x) & = & u_b(t,X(t,x)) & \text{on} \quad (0,T), \\ X(0,x) & = & x. & \end{array} \right. \tag{4.7}$$

As usual in continuum mechanics, given a ball $B_\rho(x_0) = \{x \in \mathbb{R} : |x - x_0| \le \rho\}$, we denote the transformed set by

$$B_\rho(x_0)_t = \{\, y \in \mathbb{R} : y = X(t,x) \quad \text{for some} \quad x \in B_\rho(x_0)\}.$$

THEOREM 4.3. *Let b, ϕ, β, a, and u_0 be as in section 3. Let u_b be a globally Lipschitz continuous function on Q. For $\epsilon \ge 0$ let $N_\epsilon(a(t,\cdot)) := \{(t,x) \in \{t\} \times \Omega / \, a(t,x) \le -\epsilon\}$. Assume also that $\epsilon = 0$ if $m(p-1) > 1$, and $\epsilon > 0$ if $m(p-1) \le 1$. Let $u_0 = 0$ on a ball $B_{\rho_0}(x_0)$ for some x_0 such that $(t, B_{\rho_T}(x_0)) \subset N_\epsilon(a(t,\cdot))$ for any $t \in [0,T]$ and some $L \ge \rho_0$. Then there exists a $T_\epsilon \in (0,T]$ and a function $\rho : [0,T_\epsilon] \to [0,\rho_0]$ such that $u(t,x) = 0$ a.e. $x \in B_{\rho(t)}(x_0)$ for any $t \in [0,T_\epsilon]$.*

Proof. We introduce the change of variable $b(w(t,x)) = b(u(t,x))e^{\lambda t}$. Then, it is easy to prove that w satisfies the equation

$$\begin{array}{ll} b(w)_t + u_b b(w)_x - \mu e^{\lambda t(1-(p-1)m)}\phi(w_x)_x + [(u_b)_x + \lambda]b(w) + \beta(w) & \\ \ni a(t,x)e^{\lambda t} \qquad \text{in} \quad Q. & \end{array} \tag{4.8}$$

Thus, by taking $\lambda > 2C$ with $C = \|(u_b)_x\|_{L^\infty(Q)}$ (which is finite, since u_b is a globally Lipschitz continuous function), we have that $[(u_b)_x + \lambda] \ge C > 0$. By multiplying, formally, by w (i.e., by some arguments of regularization, localization, and passing to the limit, as in Díaz and Veron [23]), we get that if $\rho \le L$, then

$$\int_{B_\rho(x_0)_t} \frac{\partial}{\partial t}\Psi(w)dx + \int_{B_\rho(x_0)_t} u_b \Psi(w)_x dx + \mu e^{\lambda t(1-(p-1)m)} \int_{B_\rho(x_0)_t} |w_x|^p \, dx$$

$$\le \mu e^{\lambda t(1-(p-1)m)} w(t,\cdot)\, |w_x(t,\cdot)|^{p-1}\, w_x(t,\cdot)\, \big|_{\partial B_\rho(x_0)_t} - \epsilon \int_{B_\rho(x_0)_t} w dx,$$

where

$$\Psi(w) := wb(w) - \int_0^w b(s)ds \tag{4.9}$$

and we used that $w \ge 0$ and that $\beta(w)w = \{0\}$. Now, by using the Reynolds transport lemma,

$$\int_{B_\rho(x_0)_t} \frac{\partial}{\partial t}\Psi(w) + \int_{B_\rho(x_0)_t} u_b \Psi(w)_x = \frac{d}{dt}\int_{B_\rho(x_0)_t} \Psi(w(t,y))dy.$$

Thus, integrating in $(0,t)$ and using the information on u_0, we get that

$$\int_{B_\rho(x_0)_t} \Psi(w(t,y))dy + C_1 \int_0^t \int_{B_\rho(x_0)_t} |w_x|^p \, dyds$$

$$\leq C_2 \int_0^t w(s,\cdot) \left| |w_x(s,\cdot)|^{p-1} w_x(s,\cdot) \big|_{\partial B_\rho(x_0)_t} \right| ds - \epsilon \int_0^t \int_{B_\rho(x_0)_t} w dx ds, \tag{4.10}$$

with

$$C_1 = \mu \min_{t\in[0,T]} e^{\lambda t(1-(p-1)m)}, \qquad C_2 = \mu \max_{t\in[0,T]} e^{\lambda t(1-(p-1)m)}.$$

Assume now, for the moment, that $1 < (p-1)m$ and $\epsilon = 0$. Then we define the energies

$$B(t,\rho) = \sup_{0\leq s\leq t} \int_{B_\rho(x_0)s} \Psi(w(s,y))dy, \qquad E(t,\rho) = \int_0^t \int_{B_\rho(x_0)_t} |w_x|^p \, dyds. \tag{4.11}$$

Using Hölder inequality and the interpolation-trace inequality of Díaz and Veron [23], we get that

$$B + E \leq K \left(\frac{\partial E}{\partial \rho} \right)^\omega \tag{4.12}$$

for some positive constant K and some $\omega > 1$, and the result follows in a standard way (see, e.g., Díaz and Veron [23], or Antontsev, Díaz, and Shmarev [2]). In the case $1 \geq (p-1)m$ and $\epsilon > 0$ we pass the term $\epsilon \int_0^t \int_{B_\rho(x_0)_t} w dx ds$ to the left-hand side of the inequality (4.10), and we introduce the additional energy function defined as

$$C(t,\rho) = \int_0^t \int_{B_\rho(x_0)_t} |w| \, dyds.$$

(Remember that $|w| = w$.) Then, we can apply Theorem 1 of Antontsev, Díaz, and Shmarev [1], with $\lambda = 0$ since the interpolation-trace inequality (2.6) of that paper applies also to the limit case $\lambda = 0$. Thus, we arrive at the inequality

$$E + C \leq K \left(\frac{\partial (E+C)}{\partial \rho} \right)^\omega \tag{4.13}$$

for some positive constant K and some $\omega > 1$, and the theorem holds. □

Remark 4.1. We point out that, due to the presence of the convective term and the specific exponents involved in (4.6), the statement of the parabolic part of Díaz and Veron [23] is not directly applicable, and this is the reason for using the characteristic transformation argument. Notice, also, that in contrast with the case $u_b = 0$, now it may occur that $T_\epsilon < T$ for any $\epsilon \geq 0$, and notice too that the energy method allows the consideration of the case $\epsilon = 0$ when $m(p-1) > 1$. Moreover, any estimate of the function $\rho(t)$ automatically gives an estimate on the location of the free boundary. Finally, we indicate that it is possible to get global consequences of the above result by estimating (globally) the energies introduced in (4.11). (For some related arguments, see, e.g., Díaz and Veron [23] or Antontsev, Díaz, and Shmarev [2].) Unfortunately the above information on the free boundary is quite implicit and difficult to manage. This also justifies the use of numerical methods.

The waiting time property can also be studied by energy methods once it is reformulated in terms of the characteristics associated with u_b. Notice that if $u_b \equiv 0$, then the characteristics are vertical lines.

THEOREM 4.4. *Let b, ϕ, β, a, u_b, ϵ, $N_\epsilon(a(t,\cdot))$, and x_0 be as in the previous theorem but now with $L > \rho_0$. Let $u_0(x) = 0$ on a ball $B_{\rho_0}(x_0)$ for some x_0 and satisfying that*

$$\int_{B_\rho(x_0)_t} \Psi(u_0(y))dy \le \theta[(\rho-\rho_0)^+]^{\omega/(\omega-1)} \qquad \text{for any} \quad \rho_0 \le \rho \le L \tag{4.14}$$

for some small enough $\theta > 0$ and some $L > \rho_0$, where Ψ is defined by (4.9) *and $\omega > 1$ is the exponent given in* (4.12) *or* (4.13). *Then, there exists $T_0 \in (0,T]$ such that $u(t,x) = 0$ a.e. $x \in B_{\rho_0}(x_0)_t$ for any $t \in [0,T_0]$, where $B_{\rho_0}(x_0)_t = \{y \in \mathbb{R} : y = X(t,x)$ for some $x \in B_{\rho_0}(x_0)\}$, with $X(t,x)$ the characteristics defined by* (4.7).

Proof. The proof follows from the same arguments as those used by Antontsev, Díaz, and Shmarev [1], but adapted to our framework. Thus, the integration by parts formula (4.10) must be replaced by

$$\begin{aligned} &\int_{B_\rho(x_0)_t} \Psi\,(w(t,y))dy + C_1 \int_0^t \int_{B_\rho(x_0)_t} |w_x|^p\, dyds \\ &\qquad \le \int_0^t w(s,\cdot) \left| |w_x(s,\cdot)|^{p-1}\, w_x(s,\cdot)\,\big|_{\partial B_\rho(x_0)_t} \right| ds \\ &\qquad\quad - \epsilon \int_{B_\rho(x_0)_t} w dx + \int_{B_\rho(x_0)_t} \Psi(u_0(y))dy. \end{aligned} \tag{4.15}$$

In particular, inequality (4.12) becomes the nonhomogeneous one,

$$B + E \le K \left(\frac{\partial E}{\partial \rho}\right)^\omega + \theta(\rho-\rho_0)_+^{\omega/(\omega-1)},$$

and the conclusion holds thanks to a technical lemma (see, e.g., Lemma 1 of Antontsev, Díaz, and Shmarev [1]). ∎

5. Numerical solution. This section is devoted to the numerical solution of the ice sheet moving boundary problem whose multivalued formulation (MF) is stated in (3.4). We first introduce the total derivative notation in conservative form,

$$\frac{Dh}{Dt} = \frac{\partial h}{\partial t} + \frac{\partial}{\partial x}(u_b h),$$

so that the complementarity formulation (CF) given by (3.5) can be posed as

$$\begin{cases} \dfrac{Dh}{Dt} - \left(\dfrac{h^{n+2}}{n+2}|h_x|^{n-1}h_x\right)_x - a \ge 0 & \text{in} \quad Q, \\ \left[\dfrac{Dh}{Dt} - \left(\dfrac{h^{n+2}}{n+2}|h_x|^{n-1}h_x\right)_x - a\right] h = 0 & \text{in} \quad Q, \\ h \ge 0 & \text{in} \quad Q, \\ h = 0 & \text{on} \quad \Sigma, \\ h = h_0(x) & \text{on} \quad \Omega. \end{cases} \tag{5.1}$$

An overview of different numerical strategies for solving free boundary problems (fixed domain methods, front-tracking, and front-fixing methods, adaptative algorithms, and others) can be found in [34]. Our approach is based on fixed domain

methods, upwinding time discretization, and duality methods for nonlinearities. More precisely, in the recent paper [13] a first attempt was made to solve (5.1) by combining this approach with a fixed point method for the nonlinear diffusive term. Indeed, the linearization process was based on freezing the nonlinear diffusive term at each step of the algorithm.

In the subsequent paper [14], in order to solve a temperature-profile coupled problem, this method was combined with the algorithm proposed in [12] to approximate ice sheet temperature distribution. However, this profile approximation procedure requires extremely small time steps and, consequently, it leads to very high computing times to obtain an accurate stationary solution.

In the present work, to overcome the drawbacks of the previously described numerical approach, we express the nonlinear diffusive term by means of a monotone operator. As usual in glaciology (see section 2), let $n = 3$. Hence, $m = 2(n+1)/n = 8/3 > 1$ and $p = n + 1 = 4$. Then, by using (3.6), we introduce the following new variable u,

$$u(t,x) = h^{8/3}(t,x), \tag{5.2}$$

so that problem (5.1) can be written in terms of u as

$$\begin{cases} \dfrac{D}{Dt}(u^{3/8}) - \mu\left(|u_x|^2 u_x\right)_x - a \geq 0 & \text{in} \quad Q, \\ \left[\dfrac{D}{Dt}(u^{3/8}) - \mu\left(|u_x|^2 u_x\right)_x - a\right] u = 0 & \text{in} \quad Q, \\ u \geq 0 & \text{in} \quad Q, \\ u = 0 & \text{on} \quad \Sigma, \\ u = u_0(x) = h_0^{8/3}(x) & \text{on} \quad \Omega, \end{cases} \tag{5.3}$$

where the constant μ takes the value $\mu = \frac{(3/8)^3}{5}$. Notice that formulation (5.3) allows us to introduce a maximal monotone operator to express the nonlinear diffusive term and implicitly contains a convective term.

In view of the particular nonlinear diffusive term in (5.3), the classical method analyzed in the framework of linear diffusion problems ($p = 2$) by Nochetto and Verdi [35] cannot be applied. In the same manner, the nonlinear diffusive terms treated in the recent paper [16] do not cover the case of (5.3).

In fact, the nonlinear p-Laplacian term has been numerically studied in [6] and the references therein, but without including either convection or free boundary aspects.

The combination of characteristic methods with duality algorithms for solving (5.3) is justified by its previous validation in analogous free boundary problems in other topics (lubrication [3, 25, 26], phase change [9], and gas flow [7], for example).

One goal of this work is to propose a numerical solution method for approximating the ice sheet profile for prescribed accumulation-ablation rates and the sliding velocity of ice. A further goal is to illustrate the qualitative properties that were analyzed in section 4.

5.1. Time semidiscretization. As in previous works in the glaciology setting [13, 14], problem (5.3) is discretized in time using the scheme of characteristics. For this, let T and M be fixed positive real numbers, and let Δt be the time step so that $T = M\Delta t$. In short, this upwinded time scheme is based on the approximation

of the total derivative; see Pironneau [37] for linear convection-diffusion equations. Thus, in our particular nonlinear convection case, for $m = 0, 1, \ldots, M(M = T/\Delta t)$, we consider the approximation

$$\frac{D}{Dt}(u^{3/8})((m+1)\Delta t, x) \approx \frac{(u^{m+1})^{3/8}(x) - J^m(x)\,(u^m)^{3/8}(\chi^m(x))}{\Delta t}, \tag{5.4}$$

where

$$u^{m+1}(x) = u((m+1)\Delta t, x) \qquad \text{in} \quad \Omega \tag{5.5}$$

and $J^m(x)$ is obtained by numerical quadrature techniques in the expression

$$J^m(x) = J(t^{m+1}, x; t^m) = 1 - \int_{t^m}^{t^{m+1}} (u_b(\tau, \chi(x, t^{m+1}; \tau)))_x \, d\tau,$$

where J is the Jacobian associated with the change of variable mapping $x \to \chi(t, x; \tau)$. Notice that the presence of J arises from the application of the characteristics method when the convection is written in conservative form (see Bercovier, Pironneau, and Sastri [5] for details).

The value $\chi^m(x)$ is given by $\chi^m(x) = \chi((m+1)\Delta t, x; m\Delta t)$, χ being the solution of the final value problem

$$\left\{ \begin{array}{rcl} \dfrac{d\chi(t, x; s)}{ds} & = & u_b(s, \chi(x, t; s)), \\ \chi(t, x; t) & = & x. \end{array} \right. \tag{5.6}$$

The next step consists of the substitution of the approximation (5.4) into (5.3) to obtain the following sequence of nonlinear elliptic complementarity problems.

For $m = 0, 1, 2, \ldots, M$, *find* u^{m+1} *such that*

$$\left\{ \begin{array}{l} \dfrac{(u^{m+1})^{3/8} - J^m((u^m)^{3/8} \circ \chi^m)}{\Delta t} - \mu \dfrac{\partial}{\partial x}\left(|u_x^{m+1}|^2 u_x^{m+1}\right) - a^{m+1} \geq 0 \quad \text{in } \Omega, \\ \qquad\qquad u^{m+1} \geq 0 \quad \text{in } \Omega, \\ \left[\dfrac{(u^{m+1})^{3/8} - J^m((u^m)^{3/8} \circ \chi^m)}{\Delta t} - \mu \dfrac{\partial}{\partial x}\left(|u_x^{m+1}|^2 u_x^{m+1}\right) - a^{m+1}\right] u^{m+1} = 0 \quad \text{in } \Omega, \\ \qquad\qquad u^{m+1} = 0 \quad \text{in } \partial\Omega, \\ \qquad\qquad u^0(x) = h_0 = (h_0)^{8/3} \quad \text{in } \Omega, \end{array} \right. \tag{5.7}$$

where $a^{m+1}(x) = a((m+1)\Delta t, x)$ *and "*$\circ$*" denotes the composition symbol.*

5.2. Spatial discretization. First, in order to solve the nonlinear complementarity problem (5.7) to obtain u^{m+1}, we pose the following equivalent variational inequality formulation.

Find $u^{m+1} \in K$ *such that*

$$\begin{array}{l} \dfrac{1}{\Delta t}\displaystyle\int_\Omega (u^{m+1,})^{3/8}\,(\varphi - u^{m+1})\,dx + \mu \int_\Omega |\, u_x^{m+1}\,|^2\, u_x^{m+1}(\varphi - u^{m+1})_x dx \\ \geq \dfrac{1}{\Delta t}\displaystyle\int_\Omega J^m\,((u^m)^{3/8} \circ \chi^m)(\varphi - u^{m+1})\,dx + \int_\Omega a^{m+1}\,(\varphi - u^{m+1})dx \\ \forall \varphi \in K, \end{array} \tag{5.8}$$

where $K = \{\varphi \in W_0^{1,4}(\Omega) \,/\, \varphi \geq 0 \; a.e. \; in \; \Omega\}$.

Next, the duality algorithm proposed in Bermúdez and Moreno [8] is applied to the variational inequality (5.8). For this, (5.8) is expressed in terms of the indicatrix function I_K of the convex K in the following form.

Find $u^{m+1} \in W_0^{1,4}(\Omega)$ *such that*

$$\frac{1}{\Delta t}\int_\Omega (u^{m+1})^{3/8}(\varphi - u^{m+1})dx + \mu\int_\Omega \mid u_x^{m+1} \mid^2 u_x^{m+1}(\varphi - u^{m+1})_x dx + I_K(\varphi) - I_K(u^{m+1})$$
$$\geq \frac{1}{\Delta t}\int_\Omega J^m\left((u^m)^{3/8} \circ \chi^m\right)(\varphi - u^{m+1})dx + \int_\Omega a^{m+1}(\varphi - u^{m+1})dx \qquad \forall \varphi \in W_0^{1,4}(\Omega). \tag{5.9}$$

Moreover, the use of subdifferential calculus leads to the equivalent formulation

$$\xi_1^{m+1} = -(\mathcal{A}(u^{m+1}) - f^m) \in \partial I_K(u^{m+1}), \tag{5.10}$$

where $\partial I_K(u)$ denotes the subdifferential of the convex function I_K at the point u; see Brezis [11] for more details. Moreover, the operator $\mathcal{A} : W_0^{1,4}(\Omega) \to W^{-1,4/3}(\Omega)$ is defined by

$$\langle \mathcal{A}(\varphi), \psi\rangle = \frac{1}{\Delta t}\int_\Omega \varphi^{3/8}\psi dx + \mu\int_\Omega \mid \varphi_x \mid^2 \varphi_x\psi_x dx,$$

and the element $f^m \in W^{-1,4/3}(\Omega)$ is defined by

$$\langle f^m, \psi\rangle = \int_\Omega a^{m+1}\psi dx + \frac{1}{\Delta t}\int_\Omega J^m((u^m)^{3/8} \circ \chi^m)\psi dx.$$

Therefore, (5.10) is equivalent to the following problem.

Find $u^{m+1} \in W_0^{1,4}(\Omega)$ *such that*

$$\frac{1}{\Delta t}\int_\Omega (u^{m+1})^{3/8}\; \psi \; dx + \int_\Omega \xi_1^{m+1}\; \psi \; dx + \mu\int_\Omega \xi_2^{m+1}\psi_x dx$$
$$-\frac{1}{\Delta t}\int_\Omega J^m((u^m)^{3/8} \circ \chi^m)\psi dx = \int_\Omega a^{m+1}\; \psi \; dx \qquad \forall \psi \in W_0^{1,4}(\Omega), \tag{5.11}$$

$$\xi_1^{m+1} \in \partial I_K\left[u^{m+1}\right], \tag{5.12}$$

$$\xi_2^{m+1} = \Lambda\left(\frac{\partial u^{m+1}}{\partial x}\right), \tag{5.13}$$

where $\Lambda(v) = \mid v \mid^2 v = v^3$.

The application of the Bermúdez–Moreno algorithm [8] to solving the nonlinear problem (5.11)–(5.13) introduces the following new unknowns (multipliers) q_1^{m+1} and q_2^{m+1},

$$q_1^{m+1} \in \partial I_K\left[u^{m+1}\right] - \omega_1 u^{m+1}, \tag{5.14}$$

$$q_2^{m+1} = \Lambda\left(\frac{\partial u^{m+1}}{\partial x}\right) - \omega_2\frac{\partial u^{m+1}}{\partial x}, \tag{5.15}$$

defined in terms of the positive parameters ω_1 and ω_2. So, (5.11) is equivalent to

$$\frac{1}{\Delta t}\int_\Omega (u^{m+1})^{3/8}\psi dx + \int_\Omega (q_1^{m+1}+\omega_1 u^{m+1})\psi\, dx + \mu\int_\Omega \left(q_2^{m+1}+\omega_2\frac{\partial u^{m+1}}{\partial x}\right)\frac{\partial \psi}{\partial x}\, dx$$
$$= \int_\Omega a^{m+1}\psi\, dx + \frac{1}{\Delta t}\int_\Omega J^m\left((u^m)^{3/8}\circ\chi^m\right)\psi dx \qquad \forall \psi \in W_0^{1,4}(\Omega). \tag{5.16}$$

Now, since ∂I_K and Λ are maximal monotone operators, the definitions given by (5.14) and (5.15) can be characterized by their respective identities (see [8]):

$$q_1^{m+1} = (\partial I_K)_{\lambda_1}^{\omega_1}\left[u^{m+1}+\lambda_1 q_1^{m+1}\right], \tag{5.17}$$

$$q_2^{m+1} = \Lambda_{\lambda_2}^{\omega_2}\left[\frac{\partial u^{m+1}}{\partial x}+\lambda_2 q_2^{m+1}\right]. \tag{5.18}$$

In (5.17) and (5.18), the functions $(\partial I_K)_{\lambda_1}^{\omega_1}$ and $\Lambda_{\lambda_2}^{\omega_2}$ denote the Yosida approximations for the operators $(\partial I_K - \omega_1 I)$ and $(\Lambda - \omega_2 I)$ with positive parameters λ_1 and λ_2, respectively (see Brezis [11], for example). Next, to discretize (5.16)–(5.18) in space, we consider piecewise linear Lagrange finite elements. More precisely, for a given positive parameter h we build a uniform finite element mesh τ_h for Ω. Thus, let $x_i = (i-1)h$, $i = 1, \ldots, N+1$, be the mesh nodes. Now, we introduce the classical finite elements spaces and sets:

$$\begin{aligned} V_h &= \{\varphi_h \in C^0(\Omega) \,/\, \varphi_h\,|_E \in P_1 \quad \forall E \in \tau_h\},\\ V_{0h} &= \{\varphi_h \in V_h \,/\, \varphi_h\,|_{\partial\Omega} = 0\},\\ K_h &= \{\varphi_h \in V_{0h} \,/\, \varphi_h(x_i) \geq 0, \quad i = 1, \ldots, N+1\}, \end{aligned} \tag{5.19}$$

where E denotes a standard finite element interval.

Then, the fully discretized problem can be posed as follows.

Find $u_h^{m+1} \in K_h$ such that

$$\begin{aligned} &\frac{1}{\Delta t}\int_\Omega (u_h^{m+1})^{3/8}\psi_h\, dx + \omega_1\int_\Omega u_h^{m+1}\psi_h\, dx + \mu\omega_2\int_\Omega \frac{\partial u_h^{m+1}}{\partial x}\frac{\partial \psi_h}{\partial x}\, dx\\ &= \int_\Omega a_h^{m+1}\,\psi_h\, dx + \frac{1}{\Delta t}\int_\Omega J^m((u_h^m)^{3/8}\circ\chi^m)\psi_h\, dx - \int_\Omega q_{1,h}^{m+1}\psi_h\, dx\\ &- \mu\int_\Omega q_{2,h}^{m+1}\frac{\partial \psi_h}{\partial x}\, dx \quad \forall \psi_h \in V_{0h}. \end{aligned} \tag{5.20}$$

Thus, by treating the first term in (5.20) in explicit form at each step of the inner multipliers loop, the numerical algorithm for solving the fully discretized problem (5.20), (5.18), and (5.17) can be sketched as follows.

Step 0 : Initialize $(u_h^{m+1})_0$ (equal to u_h^m, for example)

Step j : For a given$(u_h^{m+1})_j$, compute $(u_h^{m+1})_{j+1} \in V_{0h}$ by solving the linear problem

$$\begin{aligned}
&\omega_1 \int_\Omega (u_h^{m+1})_{j+1}\psi_h \, dx + \mu\omega_2 \int_\Omega \frac{\partial (u_h^{m+1})_{j+1}}{\partial x}\frac{\partial \psi_h}{\partial x} \, dx \\
&= -\frac{1}{\Delta t}\int_\Omega (u_h^{m+1})_j^{3/8} \, \psi_h \, dx \, - \int_\Omega (q_{1,h}^{m+1})_j \, \psi_h \, dx - \mu \int_\Omega (q_{2,h}^{m+1})_j \frac{\partial \psi_h}{\partial x} dx \\
&+ \int_\Omega a_h^{m+1}\psi_h \, dx + \frac{1}{\Delta t}\int_\Omega J^m \left(((u_h^m)_j)^{3/8} \circ \chi^m\right) \psi_h \, dx \qquad \forall \psi \in V_{0h}.
\end{aligned} \tag{5.21}$$

The multipliers updating (indexed by j) is provided by the following expressions:

$$(q_{1,h}^{m+1})_{j+1} = (\partial I_K)_{\lambda_1}^{\omega_1}[(u_h^{m+1})_{j+1} + \lambda_1 (q_{1,h}^{m+1})_j], \tag{5.22}$$

$$(q_{2,h}^{m+1})_{j+1} = \Lambda_{\lambda_2}^{\omega_2}\left[\frac{\partial}{\partial x}(u_h^{m+1})_{j+1} + \lambda_2 (q_{2,h}^{m+1})_j\right]. \tag{5.23}$$

The convergence of the duality method is established in Bermúdez and Moreno [8] and Bermúdez [7] under the technical constraint $\lambda_i \omega_i = 0.5$ for $i = 1, 2$. For this particular choice of the parameters, the Yosida approximations can easily be computed and are given by

$$(\partial I_K)_{\frac{1}{2\omega_1}}^{\omega_1}(r) = -2\omega_1 \mid r \mid,$$

$$\Lambda_{\frac{1}{2\omega_2}}^{\omega_2}(r) = 2\,\Lambda_{\frac{1}{\omega_2}}(2r) - 2\,\omega_2\, r\,,$$

where $\Lambda_\lambda(r) = (r - s)/\lambda$, the value s being the real solution of the nonlinear equation $\lambda s^3 + s = r$, which has been solved for each r by using Cardano's formulae.

5.3. Numerical results: Comparison tests. In order to validate the correct performance of our numerical approach, we have considered a first test (Test 1), which presents a closed form stationary solution. It corresponds to a no sliding case (i.e., $u_b = 0$) and is adapted from Paterson [36]. More precisely, in Test 1, for a sufficiently large time interval $(0, T)$, we consider the open set $\Omega = (-L, L)$ and the following piecewise constant accumulation-ablation function:

$$a(x) = \begin{cases} a_1 & \text{if } 0 \le |x| < R, \\ -a_2 & \text{if } R \le |x| \le L, \end{cases} \tag{5.24}$$

where $L > 1$, $a_1 > 0$, $a_2 > 0$, and $R \in (0, 1)$. Moreover, we assume that

$$a_1 R = a_2(1 - R) \tag{5.25}$$

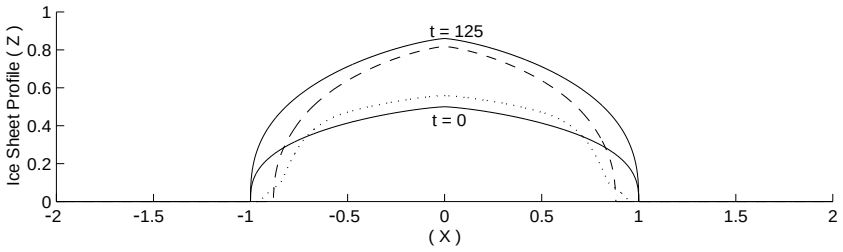


FIG. 5.1. *Computed numerical solution of Test* 1*;* $t = 0(-), t = 5(\cdots), t = 75(--)$, *stationary* $(-)$.

holds. Thus, for the particular values $a_1 = 0.01$ and $a_2 = 0.03$, we have the steady state solution

$$\overline{\eta}(x) = \begin{cases} H\left[1 - \left(1 + \dfrac{a_1}{a_2}\right)^{1/3} \left(\dfrac{|x|}{L}\right)^{4/3}\right]^{3/8} & \text{if } |x| \le R, \\ H\left(1 + \dfrac{a_2}{a_1}\right)^{1/8} \left(1 - \dfrac{|x|}{L}\right)^{1/2} & \text{if } R \le |x| \le 1, \\ 0 & \text{if } 1 \le |x| \le L, \end{cases} \tag{5.26}$$

where $H = (40\, a_1\, R\,)^{1/8}$ represents the thickness at $x = 0$ (the *ice divide*).

Moreover, in Test 1 the values $L = 2$ and $R = 0.75$ have been chosen so that $H = 0.86$. As initial conditions for the evolutive problem, we have considered

$$\eta_0(x) = \begin{cases} c\,(1 - |x|^{4/3})^{3/8} & \text{if } \ |x| \le 1, \\ 0 & \text{if } \ 1 \le |x| \le 2, \end{cases} \tag{5.27}$$

with $c = 0.5$.

For the numerical solution a uniform finite element mesh with $N = 4001$ nodes and a time step $\Delta t = 1$ have been taken.

In Figure 5.1 we present the initial profile ($t = 0$), the computed solutions for $t = 5$ and $t = 75$, and the stationary exact solution (which matches the numerical approximation for $t = 125$) for Test 1. Figure 5.1 is obtained with the described Bermúdez–Moreno method with $\omega_1 = 15$ and $\omega_2 = 30$. The computed results agree with the same test example solved with another numerical approach in Calvo, Durany, and Vázquez [13], but computing time is highly reduced (by about 99 per cent). Notice that in Figure 5.1 for $t = 5$ the ice sheet is retreating. The ice mass shrinks until $t = 25$, and then it expands with time until reaching the stationary solution given by expression (5.26). The initial contraction is mainly due to the fact that accumulation taking place at the center cannot balance the initial effect of ablation near the margins.

Test 2 is proposed to simulate the behavior of the concave profile when increasing the sliding velocity. In this case, we take $L = 2$ and $R = 1$ in (5.24), so that (5.25) is not verified. Thus, in Figures 5.2 and 5.3 several examples are presented by considering the velocity field

$$u_b(t,x) = \begin{cases} C\,x^2 & \text{if } \ x \ge 0, \\ -C\,x^2 & \text{if } \ x < 0 \end{cases} \tag{5.28}$$

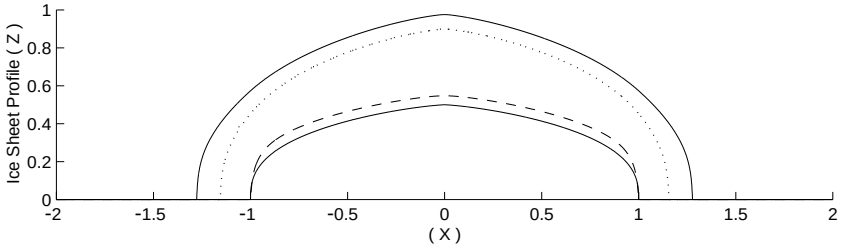


FIG. 5.2. *Numerical solution of Test* 2 *in the case* $C = 0.005$*;* $t = 0(-), t = 5(--)$, $t = 50(\cdots), t = 90(-)$.

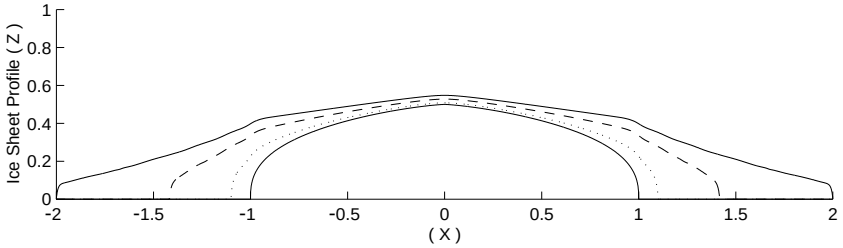


FIG. 5.3. *Numerical solution of Test* 2 *in the case* $C = 0.1$*;* $t = 0(-), t = 1(\cdots), t = 3(--), t = 5(-)$.

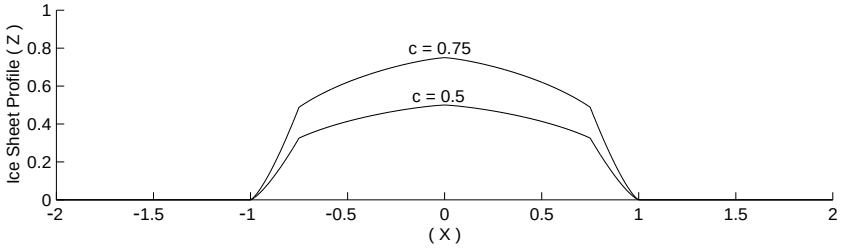


FIG. 5.4. *Convex-concave initial condition* $\overline{\eta}_0$ *for* $c = 0.5$ *and* $c = 0.75$.

and the initial condition (5.27). More precisely, Figure 5.2 shows the results obtained for $t = 5$, $t = 50$, and $t = 90$ in the case $C = 0.005$, and Figure 5.3 presents the computed profiles for $t = 1$, $t = 3$, and $t = 5$ when $C = 0.1$. These figures illustrate how concave profiles disappear in the presence of enough convection ($C = 0.1$). The values of time have been chosen to present the profiles that most emphasize this realistic phenomenon. The time step, the number of nodes, and the parameters in the Bermúdez–Moreno algorithm are the same as those used in Test 1.

Test 3 has been designed to illustrate the *waiting time property* discussed in section 4. The idea is to show how when the initial condition of the problem has a sufficiently flat convex-concave shape (see Figure 5.4), then the displacement of the initial free boundary ($S_+(t_0)$, for example) starts after a certain time (the *waiting time*). Nevertheless, an instantaneous displacement occurs for a concave initial condition (as (5.27), for example).

More precisely, in order to illustrate this so-called *waiting time property*, the numerical solutions obtained from the initial condition (5.27) and the following alter-

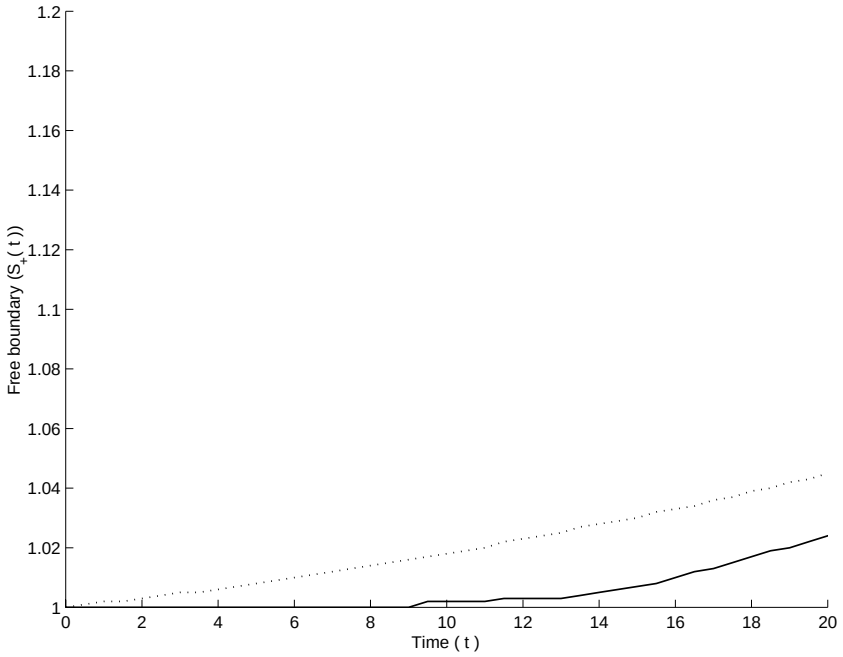


FIG. 5.5. *Moving boundary $S_+(t)$ in Test 3, with convex-concave $\overline{\eta}_0$ $(-)$ and purely concave η_0 $(\cdots)$ initial conditions for $c = 0.5$.*

native one,

$$\overline{\eta}_0(x) = \begin{cases} c\,(1-|x|^{4/3})^{3/8} & \text{if} \quad -0.75 \le x \le 0.75, \\ 16.77\,c\left(\dfrac{a_2}{2}\right)^{1/3} |x-1|^{4/3} & \text{if} \quad 0.75 \le x \le 1, \\ 16.77\,c\left(\dfrac{a_2}{2}\right)^{1/3} |x+1|^{4/3} & \text{if} \quad -1 \le x \le -0.75, \\ 0 & \text{otherwise,} \end{cases} \tag{5.29}$$

are compared for different values of c. Thus, in Figures 5.5 and 5.6 the moving boundaries are compared for the initial conditions (5.27) and (5.29) for $c = 0.5$ and $c = 0.75$, respectively. Notice that the theoretical result stated in section 4 about the *waiting time* property is a local one, while numerical observations yield a global *waiting time*. To our knowledge the proof of global *waiting time* properties is an open and difficult question.

Next we illustrate the relation between the *waiting time* and the initial ice mass stated in the theoretical analysis. As the ice mass associated with an initial condition such as (5.29) depends linearly on the parameter c, in Figure 5.7 we compare the moving boundary evolution for different values of c in the absence of convection. Notice how the *waiting time* decreases when the initial ice mass increases.

Finally, in Figure 5.8 the influence of convection for a fixed initial ice mass associated with $c = 0.75$ is illustrated. Thus, an increasing basal sliding reduces the *waiting time* as expected in real situations.

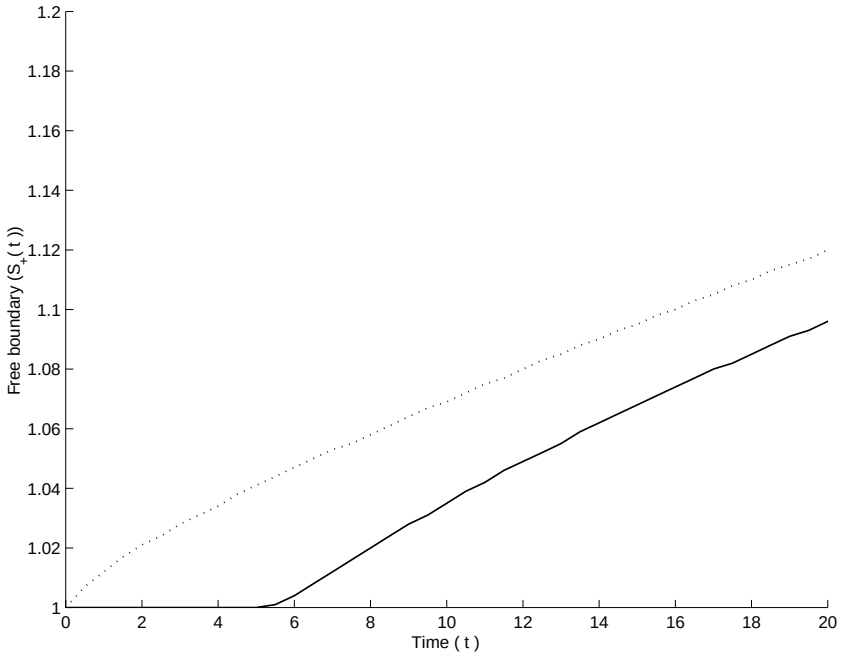


FIG. 5.6. *Moving boundary $S_+(t)$ in Test 3, with convex-concave $\overline{\eta}_0$ $(-)$ and purely concave η_0 $(\cdots)$ initial conditions for $c = 0.75$.*

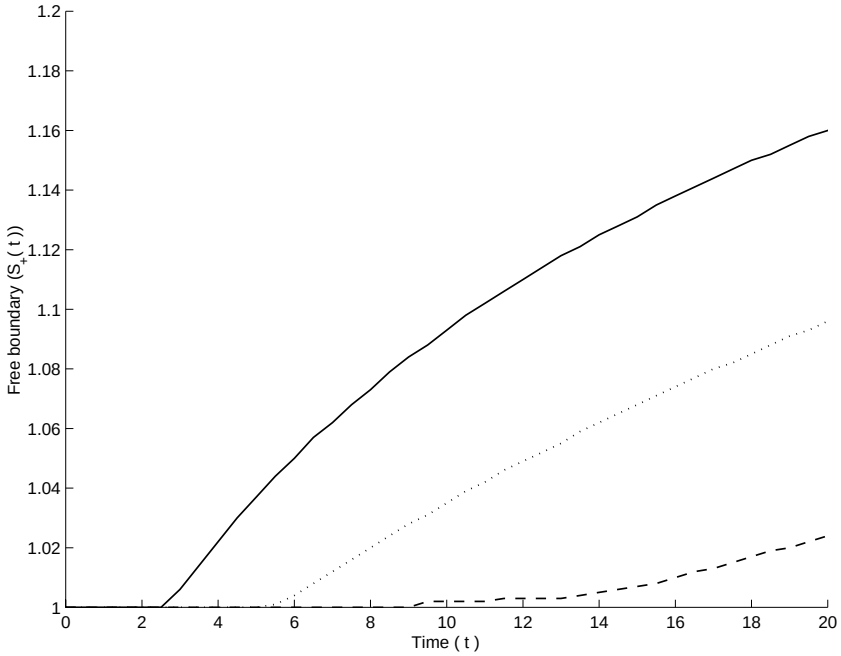


FIG. 5.7. *Moving boundary $S_+(t)$ in Test 3, with convex-concave $\overline{\eta}_0$ function and $C = 0$ for $c = 0.5(--), c = 0.75(\cdots), c = 0.9(-)$.*

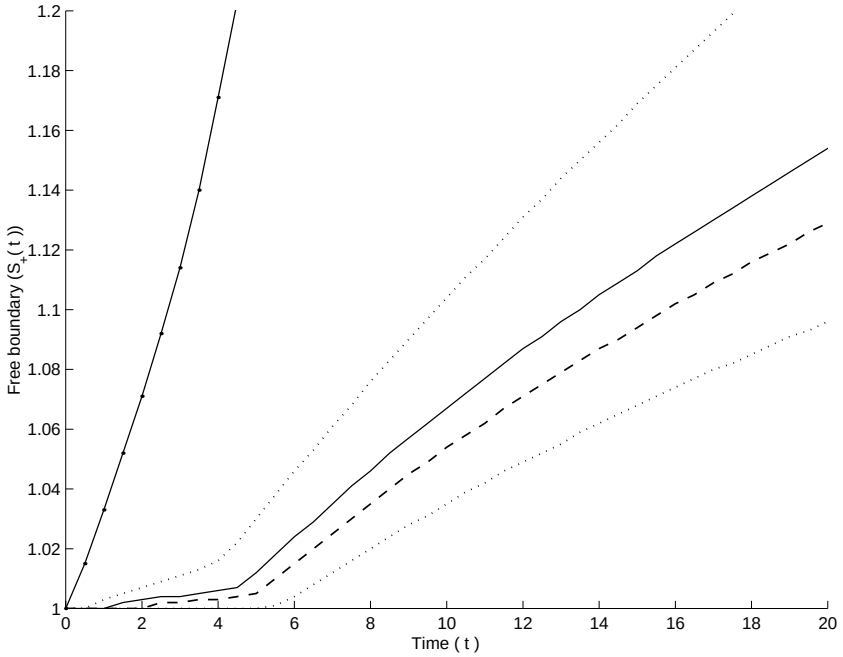


FIG. 5.8. *Moving boundary* $S_+(t)$ *in Test* 3*, with* $\overline{\eta}_0$ *and* $c = 0.75$*;* $C = 0(\cdots), C = 0.003(--), C = 0.005(-), C = 0.01(\cdots), C = 0.05(.\,-\,.-)$.

6. Discussion. In this paper we have used different but equivalent weak formulations, expressed in terms of either multivalued equations or variational inequalities when the complementarity formulation is considered for numerical purposes. This approach makes more precise the original doubly nonlinear formulation of Fowler [29], converting it into an obstacle problem for the associated operator. Assuming some extra regularity properties of the solution, we have given sufficient conditions (in terms of a, the accumulation rate, and h_0, the initial thickness) for the existence of the free moving boundary and its spatial location. For this, we employed two different methods: a comparison principle, combined with the construction of suitable barrier functions in the case $u_b \equiv 0$, and a local energy method if $u_b \neq 0$. In both cases, we prove rigorously the possible existence of a waiting time in the dynamics of the free boundary, whose location and evolution can be qualitatively described as long as suitable and physically admissible hypotheses on the data of the problem hold. From the numerical point of view, the main advantage of the proposed new approach follows from the introduction of a maximal monotone operator for the nonlinear diffusive term that had already been treated in explicit form [13]. Thus, a duality method can also be applied to greatly improve the speed of convergence with respect to the previous work. In order to verify the good performance of the new algorithm as well as the computational cost reduction, a problem with a closed form solution has been tested. Moreover, to complete the theoretical results and reflect some realistic situations, several test examples illustrate some qualitative properties of the ice profile and the associated free boundary.

REFERENCES

[1] S. N. Antontsev, J. I. Díaz, and S. I. Shmarev, *The support shrinking properties for solutions of quasilinear parabolic equations with strong absorption terms*, Ann. Fac. Sci. Toulouse Math., 4 (1995), pp. 5–30.

[2] S. N. Antontsev, J. I. Díaz, and S. I. Shmarev, *Energy Methods for Free Boundary Problems*, Birkäuser-Boston, Cambridge, MA, 2001.

[3] G. Bayada, M. Chambat and C. Vázquez, *Characteristics method for the formulation and computation of a free boundary cavitation problem*, J. Comput. Appl. Math., 98 (1998), pp. 191–212.

[4] Ph. Benilan and P. Wittbold, *On mild and weak solutions of elliptic-parabolic problems*, Adv. Differential Equations, 1 (1996), pp. 1053–1073.

[5] M. Bercovier, O. Pironneau, and V. Sastri, *Finite elements and characteristics for some parabolic-hyperbolic problems*, Appl. Math. Model., 7 (1983), pp. 89-96.

[6] R. Bermejo and J.-A. Infante, *A multigrid algorithm for the p-Laplacian*, SIAM J. Sci. Comput., 21 (2000), pp. 1774–1789.

[7] A. Bermúdez, *Un método numérico para la resolución de ecuaciones con varios términos no lineales. Aplicación a un problema de flujo de gas en un conducto*, Rev. R. Acad. Cienc. Exactas Fís. Nat. (Esp.), 78 (1981), pp. 485–495.

[8] A. Bermúdez and C. Moreno, *Duality methods for solving variational inequalities*, Comput. Math. Appl., 7 (1981), pp. 43–58.

[9] A. Bermúdez, M. C. Muñiz, and P. Quintela, *Numerical solution of a three-dimensional thermoelectric problem taking place in an aluminium electrolytic cell*, Comput. Methods Appl. Mech. Engrg., 106 (1993), pp. 129–142.

[10] L. Boccardo, D. Giachetti, J. I. Díaz, and F. Murat, *Existence of a solution for a weaker form of a nonlinear elliptic equation*, in Recent Advances in Nonlinear Elliptic and Parabolic Problems, Pitman Res. Notes Math. Ser. 208, Longman Scientific Technical, Harlow, UK, 1989, pp. 229–246.

[11] H. Brezis, *Opérateurs Maximaux Monotones et Semigroupes de Contractions dans les Espaces de Hilbert*, North–Holland, Amsterdam, 1973.

[12] N. Calvo, J. Durany, and C. Vázquez, *Numerical approach of temperature distribution in a free boundary polythermal ice sheet*, Numer. Math., 83 (1999), pp. 557–580.

[13] N. Calvo, J. Durany, and C. Vázquez, *Numerical computation of ice sheet profiles with free boundary problems*, Appl. Numer. Math., 35 (2000), pp. 111–128.

[14] N. Calvo, J. Durany, and C. Vázquez, *Numerical approach of thermomechanical coupled problems with moving boundaries in glaciology*, Math. Models Methods Appl. Sci., 12 (2002), pp. 229–249.

[15] J. Carrillo and P. Wittbold, *Uniqueness of renormalized solutions of degenerate elliptic-parabolic problems,* J. Differential Equations, 156 (1999), pp. 93–121.

[16] Z. Chen, R. H. Nochetto, and A. Schmidt, *A characteristics Galerkin method with adaptive error control for the continuous casting problem*, Comput. Methods Appl. Mech. Engrg., 189 (2000), pp. 249-276.

[17] J.I. Díaz, *Elliptic and parabolic quasilinear equations giving rise to a free boundary: The boundary of the support of the solution*, in Nonlinear Functional Analysis and Applications, F. E. Browder, ed., Proc. Sympos. Pure Math. 45.2, AMS, Providence, RI, 1986, pp. 381–393.

[18] J. I. Díaz and F. de Thelin, *On a nonlinear parabolic problem arising in some models related to turbulent flows*, SIAM J. Math. Anal., 25 (1994), pp. 1085–1111.

[19] J. I. Díaz and G. Galiano, *On the Boussinesq system with nonlinear thermal diffusion*, Nonlinear Anal., 30 (1997), pp. 3255–3263.

[20] J. I. Díaz and J. Hernández, *Qualitative properties of free boundaries for some non linear degenerate parabolic equations*, in Nonlinear Parabolic Equations: Qualitative Properties of Solutions, L. Boccardo and A. Tesei, eds., Pitman Res. Notes Math. Ser. 149, Longman Scientific and Technical, London, 1989, pp. 85–93.

[21] J. I. Díaz and E. Schiavi, *Tratamiento matemático de una ecuación parabólica cuasilineal degenerada en Glaciología*, in Electronic Proceedings of the XIV CEDYA-IV Congreso de Matemática Aplicada, Vic, Barcelona, 1995, http://www.ma1.upc.es/cedya/cedya.html.

[22] J. I. Díaz and E. Schiavi, *On a degenerate parabolic/hyperbolic system in glaciology giving rise to a free boundary*, Nonlinear Anal., 38 (1999), pp. 649–673.

[23] J. I. Díaz and L. Veron, *Local vanishing properties of solutions to elliptic and parabolic equations*, Trans. Amer. Math. Soc., 290 (1985), pp. 787–814.

[24] R. J. Di Perna and P. L. Lions, *On the Cauchy problem for Boltzmann equations: Global existence and weak stability,* Ann. of Math., 130 (1989), pp. 321–366.

[25] J. Durany, G. García, and C. Vázquez, *An elastohydrodynamic coupled problem between a piezoviscous equation and a hinged plate model*, M2AN Math. Model. Numer. Anal., 31 (1997), pp. 495–516.
[26] J. Durany, G. García, and C. Vázquez, *Numerical simulation of a lubricated Hertzian contact problem under imposed load*, Finite Elem. Anal. Des., 38 (2002), pp. 645–658.
[27] G. Duvaut and J. L. Lions, *Les inéquations en Mécanique et en Physique*, Dunod, Paris, 1972.
[28] L. C. Evans and B. F. Knerr, *Instantaneous shrinking of the support of nonnegative solutions to certain nonlinear parabolic equations and variational inequalities*, Illinois J. Math., 23 (1979), pp. 153–166.
[29] A. C. Fowler, *Modelling ice sheet dynamics*, Geophys. Astrophys. Fluid Dyn., 63 (1992), pp. 29–65.
[30] A. C. Fowler, *Mathematical Models in the Applied Sciences*, Cambridge University Press, Cambridge, UK, 1997.
[31] A. C. Fowler and E. Schiavi, *A theory of ice-sheet surges*, J. Glaciology, 44 (1998), pp. 104–118.
[32] K. Hutter, *Theoretical Glaciology*, Reidel, Dordrecht, The Netherlands, 1981.
[33] L. Lliboutry, *Very Slow Flows of Solids*, Martinus Nijhoff, Dordrecht, The Netherlands, 1987.
[34] R. H. Nochetto, *Numerical methods for free boundary problems*, in Free Boundary Problems: Theory and Applications, K.H. Hoffman and J. Sprekels, eds., Pitman Res. Notes Math. Ser. 185–186, Longman Scientific and Technical, London, 1990, pp. 555–566.
[35] R. H. Nochetto and C. Verdi, *Approximation of degenerate parabolic problems using numerical integration*, SIAM J. Numer. Anal., 25 (1988), pp. 784–814.
[36] W. S. B. Paterson, *The Physics of Glaciers*, Pergamon, Oxford, 1981.
[37] O. Pironneau, *On the transport-diffusion algorithm and its application to Navier-Stokes equation*, Numer. Math., 38 (1982), pp. 309–332.
[38] L. Tarasov and W. R. Peltier, *Terminating the* 100 *kyr ice age cycle*, J. Geophys. Res., 102 (1997), pp. 21665–21693.

SIAM J. APPL. MATH.
Vol. 63, No. 2, pp. 708–721

HYPERBOLIC PHASE TRANSITIONS IN TRAFFIC FLOW*

RINALDO M. COLOMBO†

Abstract. This paper provides a mathematical model of the phenomenon of phase transitions in traffic flow. The model consists of a scalar conservation law coupled with a 2×2 system of conservation laws. The coupling is achieved via a free boundary, where the phase transition takes place. For this model, the Riemann problem is stated and globally solved. The Cauchy problem is proved to admit a solution defined globally in time without any assumption about the smallness of the initial data or the number of phase boundaries. Qualitative properties of real traffic flow are shown to agree with properties of the solutions of the model.

Key words. hyperbolic conservation laws, phase transitions, macroscopic vehicular traffic model, hyperbolic systems, partial differential equations

AMS subject classifications. 35L65, 90B20, 76T99

PII. S0036139901393184

1. Introduction.

We are concerned with phase transitions in hyperbolic systems of conservation laws. By *phase transition* we mean a discontinuity separating states belonging to different *phases*, i.e., having deep qualitative differences. Equivalently, a phase boundary can be seen as a *free boundary* separating two different models and whose evolution is determined by the solution on both of its sides. In fact, besides the usual phase transitions in fluids that motivate the term, other phenomena are described within this framework. Examples include those from nonlinear elastodynamics [1] and from combustion theory [9, 20].

The present work deals with phase transitions in traffic flow. In the specialized literature (see [15] and the many references therein), it has been shown that traffic flow admits two distinct behaviors, depending on whether it is *free* or *congested*. In this context, "mathematical models and theories ... still cannot explain and predict the experimental features of the phase transitions in real traffic flow which have been found out ..." [15, p. 257]. The model presented below gives such an explanation, showing that "real traffic flow" can be described within the mathematical framework provided here, based on hyperbolic conservation laws that develop phase transitions.

From the analytical point of view, we study a scalar conservation law coupled with a 2×2 system of conservation laws. The coupling is achieved via the phase boundary, i.e., a free boundary whose evolution is regulated by the Rankine–Hugoniot conditions. The whole model admits a bounded variation (**BV**) weak solution defined globally in time. Note that the total variation of the initial data is required merely to be bounded.

The presence of phase transitions usually leads to a lack of uniqueness in the solution to Riemann problems. A standard way out of this dilemma is the introduction of suitable admissibility conditions [1, 20]. Here, we do not need this provision and, assigning the *structure* of the solution, select a unique solution to any Riemann problem.

*Received by the editors July 30, 2001; accepted for publication (in revised form) July 18, 2002; published electronically December 19, 2002.
http://www.siam.org/journals/siap/63-2/39318.html
†Department of Mathematics, Brescia University, Via Branze 38, 25123 Brescia, Italy (rinaldo@ing.unibs.it).

From the point of view of traffic flow theory, the present model satisfies the requirements stated in [2, 11] and is able to describe various phenomena reported in Kerner's review [15]. Indeed, Kerner's work inspired the present paper, and we shall often refer to it. More precisely, here we present a PDE-based model that can be rigorously stated and studied. The qualitative properties of its solutions fully agree with the qualitative "three phase theory" given by Kerner [14] and supported by him [13] with empirical studies. Furthermore, the mathematical techniques here adopted are those typical of several phase transition models of continuum dynamics, providing a mathematical justification of Kerner's choice of the term "phase transitions" referring to traffic flow.

The present work, while being in full agreement with Kerner's observations, suggests a different choice of terms in the distinction between *synchronized flow* and *wide moving jams*, called by Kerner distinct *phases*. Here, both these configurations are present and have the same properties underlined in [15]. But they essentially correspond to *waves* of two different characteristic families, rather than to different *phases*; see section 4 for a more detailed discussion.

A further example of qualitative agreement between the present model and the observed phenomena is shown in Figure 1.1. Here, the *fundamental diagram*, i.e., the "curve in the flow-density plane which gives a correspondence of the vehicle density to the flow rate in traffic flow" [15, p. 254] is compared with the domain in which the present model is defined and studied. Note that while the former picture consists of measured data, the second is determined uniquely as an invariant domain for a set of PDEs; indeed it displays the sets Ω_f and Ω_c defined in (2.5). These sets are characterized as being *invariant* for the system of PDEs (2.3) introduced below; for a characterization of invariant domains in systems of conservation laws, see [12].

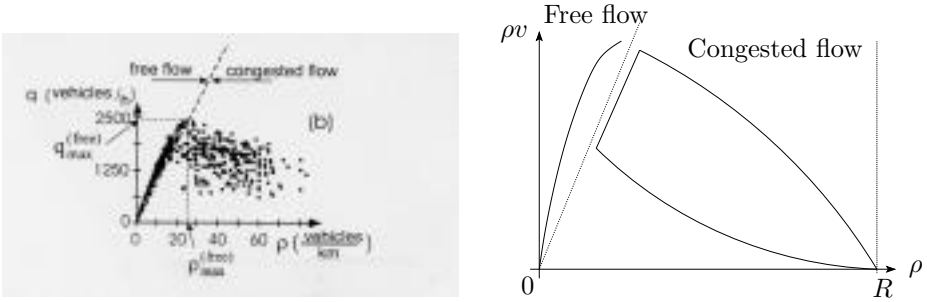


FIG. 1.1. *Left, an experimental fundamental diagram (from* [15]; used with permission) *and, right, the one obtained from the model presented here as an invariant set for* (2.3)*; see* (2.5).

In section 4, further relations between this model and features of real traffic flow will be considered. In particular, the classical Lighthill–Whitham [17] and Richards [18] (LWR) model (2.1) gives a good solution of the *traffic light problem* (see [21, p. 71]. In section 4 we shall consider the solution of the same problem, as provided by the model presented here, obtaining some reasonable improvements. Quantitative tests on the present model are under investigation.

This paper is organized as follows. The next section deals with the statement of the model. Then, section 3 is devoted to the Riemann problem, while the Cauchy problem is left to section 5. The final section 6 contains the technical details.

Throughout, we focus on analytically tractable realistic situations rather than aiming at maximal generality.

2. The model. The two phases we shall be concerned with are *free flow* and *congested flow*. In the case of free flow the classical LWR model

$$\partial_t \rho + \partial_x (\rho \cdot v) = 0, \qquad v = v_f(\rho), \tag{2.1}$$

is applicable. Here, ρ is the car density, while $v = v_f(\rho)$ is a suitable smooth decreasing function giving the car speed v at traffic density ρ.

As the car density ρ increases, the assumption that the speed v is a function of only ρ is no longer acceptable, as clearly shown by several observations (see [15], the references therein, and Figure 1.1, left). As soon as v crosses a certain threshold, the density-flow points are scattered in a cloud rather than along a line. This is *congested flow*. Here, the standard LWR model (2.1) is inadequate. We propose to complete it by means of the 2×2 system (see [8])

$$\begin{cases} \partial_t \rho + \partial_x(\rho \cdot v) = 0, \\ \partial_t q + \partial_x((q - q_*) \cdot v) = 0, \end{cases} \qquad v = v_c(\rho, q), \tag{2.2}$$

where ρ and v appear as *independent* variables. The weighted flow q is a variable originally motivated by the linear momentum in gas dynamics; v_c is given by the speed law (2.4). The threshold parameter q_* distinguishes between possible behaviors of the flow; see [8]. The well-posedness of (2.2) in the sense of the theory of standard Riemann semigroups [5, 6] follows from [3].

Thus, we propose the following model:

$$\begin{array}{ll} \text{Free flow: } (\rho, q) \in \Omega_f, & \text{Congested flow: } (\rho, q) \in \Omega_c, \\ \partial_t \rho + \partial_x [\rho \cdot v] = 0, & \begin{cases} \partial_t \rho + \partial_x [\rho \cdot v] = 0, \\ \partial_t q + \partial_x \left[(q - q_*) \cdot v\right] = 0, \end{cases} \\ v = v_f(\rho), & v = v_c(\rho, q), \end{array} \tag{2.3}$$

where Ω_f and Ω_c denote the *free* and the *congested* phases, respectively. In Ω_f the only variable is the car density ρ. Here, the car speed v is assumed to be a known function v_f of the car density: $v = v_f(\rho)$. In Ω_c the variables are the car density ρ and the car speed v or, equivalently, ρ and the weighted flow q; see [8]. Thus, at a fixed density, different speeds are admissible, as raw data observations require. Note that there may well be car densities at which the flow may be either free or congested.

In what follows we assume that the speed laws in the two phases are

$$v_f(\rho) = \left(1 - \frac{\rho}{R}\right) \cdot V \quad \text{and} \quad v_c(\rho, q) = \left(1 - \frac{\rho}{R}\right) \cdot \frac{q}{\rho}\,. \tag{2.4}$$

The former relation is the simplest standard linear choice (see, for instance, [22]), while the latter was introduced in [8]. Here, R is the maximal possible car density and V is the maximal possible speed. More general expressions can be considered, but we limit ourselves to the expressions above in order to keep the formal analytical difficulties at a minimum while maintaining all the more interesting qualitative features.

In the present model, the road is translation invariant both spatially and temporally, since none of the various functions appearing in (2.3)–(2.4) explicitly depends on x or t. As a consequence, the model may not foresee where and when new queues might form. To this aim, either time/space dependent terms should be introduced, or nondeterministic disturbances should be added. Both possibilities are far from the scope of the present paper. As a consequence, we impose the condition that if (2.3)

is assigned an initial datum contained in a single phase, then the solution will be contained in the same phase for all times. In other words, if the traffic flow is free (or congested) on all the real line, it will remain free (or congested) for all times. This leads us to require that Ω_f be invariant with respect to (2.1) and, similarly, that Ω_c be invariant with respect to (2.2). We are thus lead to define (see Figure 2.1)

$$\begin{aligned} \Omega_f &= \{(\rho, q) \in [0, R] \times [0, +\infty[: v_f(\rho) \geq \hat{V}_f,\ q = \rho \cdot V\}, \\ \Omega_c &= \left\{(\rho, q) \in [0, R] \times [0, +\infty[: v_c(\rho.q) \leq \hat{V}_c,\ \frac{q - q_*}{\rho} \in \left[\frac{Q_1 - q_*}{R}, \frac{Q_2 - q_*}{R}\right]\right\}. \end{aligned} \tag{2.5}$$

Here, $\hat{V}_f$ and $\hat{V}_c$ are the threshold speeds; i.e., above $\hat{V}_f$ the flow is free, while below $\hat{V}_c$ the flow is congested. We impose that the two phases do not intersect and hence that $\hat{V}_f > \hat{V}_c$; see Remark 2 in section 6. Moreover, $\hat{V}_f \leq V$ since $\frac{d}{d\rho}(\rho v(\rho)) \geq 0$ in the free phase. $Q_1 \in]0, q_*[$ and $Q_2 \in]q_*, +\infty[$ depend on the various environmental conditions and determine the width of the "cloud" in the congested phase. The case in which one of the equalities $Q_1 = q_*$ or $Q_2 = q_*$ holds is simpler (see Remark 1 in section 6). Finally, we let $(\check{Q}_f - q_*)/\check{R}_f = (Q_2 - q_*)/R$, as is suggested from the experimental data (see Figure 1.1) and is consistent with the "capacity drop" [15, p. 254] that takes place when passing from the free phase to the congested one.

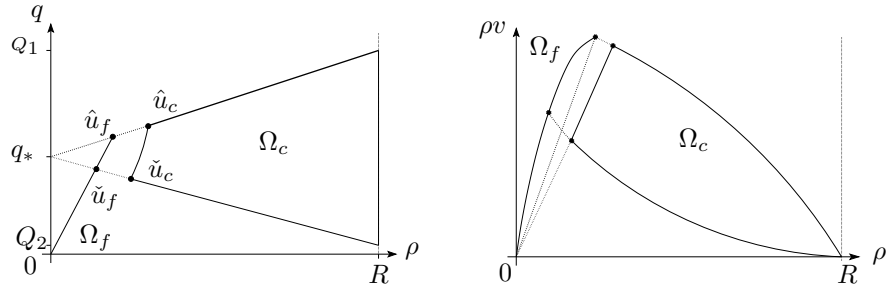


FIG. 2.1. *Left, the fundamental diagram in the (ρ, q)-plane and, right, in the $(\rho, \rho v)$-plane.*

Note that $\max_{\Omega_f} \rho v \geq \max_{\Omega_c} \rho v$, according to the phenomenon of *capacity drop*. The fundamental diagram $\Omega_f \cup \Omega_c$ depends on the specific road to which the model is applied and plays a central role, for the Riemann solver itself depends on its choice.

Note that (2.3) can be formally restated as a 2×2 system of conservation laws, for instance by introducing the map

$$f(\rho, q) = \begin{cases} (\rho \cdot v_f(\rho), q \cdot v_f(q)) & \text{if } (\rho, q) \in \Omega_f, \\ (\rho \cdot v_c(\rho, q), (q - q_*) \cdot v_c(\rho, q)) & \text{if } (\rho, q) \in \Omega_c, \end{cases} \tag{2.6}$$

and writing $\partial_t(\rho, q) + \partial_x[f(\rho, q)] = 0$. However, the solutions to (2.1)–(2.2) defined below are *not* standard Lax solutions [6, 10, 16] to any 2×2 systems of hyperbolic conservation laws. This fact will clearly follow from the next section.

We conclude this section with a remark on the parameters defining the model. The maximal density R, the maximal speed V, and the threshold densities $\hat{V}_f$ and $\hat{V}_c$ have a clear physical meaning and can be estimated as in other traffic models. Q_1 and Q_2 depend on the various environmental conditions. A possible way to determine them consists of imposing the requirement that all measured data fall inside the domain

$\Omega_f \cup \Omega_c$. The situation with q_* is more delicate. A posteriori, in section 4, we shall see a property of the line $q = q_*$ that, indirectly, leads to a possible way to determine q_*. A theory of the inverse problem for systems of hyperbolic conservation laws would give better tools but is at present not available.

3. The Riemann problem. In the free phase the characteristic speed is $\lambda(\rho) = V \cdot (1 - 2\frac{\rho}{R})$, while in the congested phase, from [8] we obtain the characteristic speeds λ_1, λ_2, the eigenvectors r_1, r_2, the i-Lax curve $q = q_i(\rho; \rho_o, q_o)$ exiting (ρ_o, q_o), and the Riemann invariants w_1, w_2. The following display summarizes this information:

$$
\begin{array}{ll}
r_1(\rho, q) = \begin{bmatrix} \rho \\ q - q_* \end{bmatrix}, & r_2(\rho, q) = \begin{bmatrix} R - \rho \\ \frac{R}{\rho} q \end{bmatrix}, \\
\lambda_1(\rho, q) = \left(\dfrac{2}{R} - \dfrac{1}{\rho} \right) \cdot (q_* - q) - \dfrac{q_*}{R}, & \lambda_2(\rho, q) = v_c(\rho, q), \\
\nabla \lambda_1 \cdot r_1 = 2 \dfrac{q_* - q}{R}, & \nabla \lambda_2 \cdot r_2 = 0, \\
q_1(\rho; \rho_o, q_o) = q_* + \dfrac{q_o - q_*}{\rho_o} \rho, & q_2(\rho; \rho_o, q_o) = \dfrac{\rho}{\rho_o} \dfrac{R - \rho_o}{R - \rho} q_o, \\
w_1 = v_c(\rho, q), & w_2 = \dfrac{q - q_*}{\rho}.
\end{array}
\tag{3.1}
$$

In the (ρ, q)-plane the 1-Lax curves are the straight half-lines exiting $(0, q_*)$, while the 2-Lax curves are convex, exit $(0, 0)$, increase monotonically, and $\lim_{\rho \to R} q_2(\rho, \rho_o, q_o) = +\infty$ for all (ρ_o, q_o). Shock and rarefaction curves coincide, but the 2-Lax curves are *not* straight lines. Hence, (2.2) is *not* a Temple system in the sense defined in [19]. The first family is not genuinely nonlinear, for $\nabla \lambda_1 \cdot r_1$ vanishes along the characteristic line $q = q_*$.

By the Riemann problem we mean (2.3) together with the initial datum

$$
(\rho, q)(0, x) = \begin{cases} (\rho^l, q^l) & \text{if } x < 0, \\ (\rho^r, q^r) & \text{if } x > 0. \end{cases}
\tag{3.2}
$$

If (ρ^l, q^l) and (ρ^r, q^r) are in the same phase, then the usual Lax solution to (2.1) or to (2.2) will be used. Note that this solution attains values in the same phase of the initial data, and no phase transition may arise.

If $(\rho^l, q^l) \in \Omega_f$ and $(\rho^r, q^r) \in \Omega_c$, then an *admissible solution* to (2.3), (3.2) is a self-similar function $u \colon [0, +\infty[\times \mathbf{R} \mapsto \Omega_f \cup \Omega_c$ such that there exists a $\Lambda \in \mathbf{R}$ with

1. $u(t,]-\infty, \Lambda t[) \subseteq \Omega_f$ and $u(t,]\Lambda t, +\infty[) \subseteq \Omega_c$;
2. the functions u^l and u^r, respectively defined by

$$
u^l(t, x) = \begin{cases} u(t, x) & \text{if } x < \Lambda \cdot t, \\ u(t, \Lambda t-) & \text{if } x > \Lambda \cdot t, \end{cases}
$$

$$
u^r(t, x) = \begin{cases} u(t, \Lambda t+) & \text{if } x < \Lambda \cdot t, \\ u(t, x) & \text{if } x > \Lambda \cdot t, \end{cases}
$$

   are restrictions of Lax solutions to Riemann problems for (2.1) and (2.2), respectively;
3. the Rankine–Hugoniot conditions

$$
\Lambda \cdot (\rho(t, \Lambda t+) - \rho(t, \Lambda t-)) = \rho(t, \Lambda t+) \cdot v_c(t, \Lambda t+) - \rho(t, \Lambda t-) \cdot v_f(t, \Lambda t-)
$$

   are satisfied for all positive t.

In the other case $(\rho^l, q^l) \in \Omega_c$ and $(\rho^r, q^r) \in \Omega_f$, conditions entirely analogous to the ones above are required.

Statement 3 above ensures that the total number of cars is conserved.

Note that *nucleation* is not possible, since no phase boundary may arise in a Riemann problem with data in a single phase. On the other hand, in the case of the Cauchy problem, zones of free (resp., congested) traffic may well disappear and turn into zones of congested (resp., free) traffic; see Figure 4.3.

We show below that it is possible to define a global Riemann solver for (2.3), (3.2). Furthermore, this solver enjoys properties underlined in the literature [2, 11] as *not* satisfied in several common models. Recall that a *Riemann solver* is a map assigning to any pair of states u^l, u^r an admissible **BV** self-similar solution to (2.3), (3.2).

The main tool in the construction of the Riemann solver is its consistency. Namely, let $\mathcal{R}\colon (u^l, u^r) \mapsto \mathcal{R}(u^l, u^r)$ denote a Riemann solver; i.e., $x \mapsto \mathcal{R}(u^l, u^r)(x)$ is the solution computed at time 1 of the Riemann problem with data u^l, u^r. $\mathcal{R}$ is *consistent* if the following two conditions hold for all u^l, u^m, u^r, and $\bar{x}$:

$$\textbf{I.} \qquad \mathcal{R}(u^l, u^r)(\bar{x}) = \bar{u} \qquad \Rightarrow \qquad \begin{array}{l} \mathcal{R}(u^l, \bar{u}) = \left\{ \begin{array}{ll} \mathcal{R}(u^l, u^r) & \text{if } x \le \bar{x}, \\ \bar{u} & \text{if } x > \bar{x}, \end{array} \right. \\ \mathcal{R}(\bar{u}, u^r) = \left\{ \begin{array}{ll} \bar{u} & \text{if } x < \bar{x}, \\ \mathcal{R}(u^l, u^r) & \text{if } x \ge \bar{x}, \end{array} \right. \end{array}$$

$$\textbf{II.} \quad \left. \begin{array}{l} \mathcal{R}(u^l, u^m)(\bar{x}) = u^m \\ \mathcal{R}(u^m, u^r)(\bar{x}) = u^m \end{array} \right\} \quad \Rightarrow \quad \mathcal{R}(u^l, u^r) = \left\{ \begin{array}{ll} \mathcal{R}(u^l, u^m) & \text{if } x < \bar{x}, \\ \mathcal{R}(u^m, u^r) & \text{if } x \ge \bar{x}. \end{array} \right.$$

This property is enjoyed by the standard Lax solver [16] and is a necessary condition for the well-posedness of the Cauchy problem.

PROPOSITION 3.1. *Assume that the fundamental diagram is as in Figure* 1.1*, with* $\hat{V}_f \ge \hat{V}_c$*. Then, there exists a Riemann solver assigning a unique self-similar admissible solution to any Riemann problem* (2.3), (3.2) *for all pairs of initial states in* $\Omega_f \cup \Omega_c$*. Moreover,*

(0) *the Riemann solver is consistent;*
(1) *no wave travels faster than the cars;*
(2) *any solution attains values in* $\Omega_f \cup \Omega_c$*, i.e., in a compact set where both car density and car speed are bounded and nonnegative;*
(3) *if the initial data are in the same phase, the solution also attains values in that phase;*
(4) *no* 2×2 *system of conservation laws may have as its Lax solutions those defined by this Riemann solver.*

Note that the *vacuum* state $\rho = 0$ is treated as any other state and does not lead to any instability, unlike what happens in [2].

Aiming at the later use of this proposition for the Cauchy problem, we proceed by first generalizing the usual Lax curves of the first and second families exiting a fixed $u^l \in \Omega_f \cup \Omega_c$. Furthermore, if u^r is on the generalized curve exiting u^l, we will suitably define the *size* $\Sigma_i(u^l, u^r)$ of the wave connecting u^l to u^r so that

$$c \cdot \|u^r - u^l\| \le |\Sigma_i(u^l, u^r)| \le C \cdot \|u^r - u^l\|, \tag{3.3}$$

C and c being suitable positive constants independent from u^l, u^r, and i. The role of Σ_i will be essential in the study of the Cauchy problem.

Proof of Proposition 3.1. The proof is achieved through three steps.

A. Generalized Riemann coordinates. Let $\check{u} = (\check{R}, \check{Q})$ be the point in Ω_f defined by $\check{Q} = \check{R} \cdot V$ and $\frac{\check{Q}-q_*}{\check{R}} = \frac{Q_1-q_*}{R}$; see Figure 2.1, left. For $(\rho, q) \in \Omega_f \cup \Omega_c$, define the generalized Riemann coordinates (w_1, w_2) as

$$w_1 = \begin{cases} v_c(\rho, q) & \text{if } (\rho, q) \in \Omega_c, \\ \hat{V}_f & \text{if } (\rho, q) \in \Omega_f, \end{cases}$$

(3.4)

$$w_2 = \begin{cases} \frac{q-q_*}{\rho} & \text{if } (\rho, q) \in \Omega_c, \\ \frac{q-q_*}{\rho} & \text{if } (\rho, q) \in \Omega_f, \rho \geq \check{R}, \\ v_f(\rho) - v_f(\check{R}) + \frac{\check{Q}-q_*}{\check{R}} & \text{if } (\rho, q) \in \Omega_f, \rho \leq \check{R}. \end{cases}$$

The construction of the solution to the Riemann problem (2.3), (3.2) will be carried out in the w coordinates. Define $W_i = \frac{Q_i - q_*}{R}$; see Figure 3.1, right.

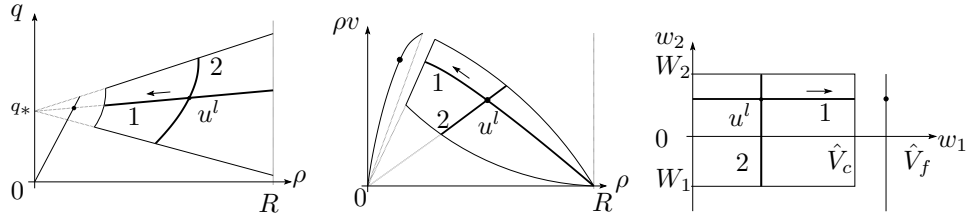


FIG. 3.1. *Generalized Lax curves exiting u^l, with $u^l \in \Omega_c$.*

B. Generalized Lax curves and simple waves. First let $w^l \in \Omega_c$. Then the generalized 1-curve exiting w^l is the segment $w_2 = w_2^l$, $w_1 \in [0, \hat{V}_c]$, to which we add the point $(w_1^l, \hat{V}_f)$; see Figure 3.1. Assume now that w^r is on this generalized 1-curve; i.e., $w_2^r = w_2^l$. Then the solution to (2.3), (3.2) is the standard Lax solution as long as $w^r \in \Omega_c$. If $w^r \in \Omega_f$, the solution to (2.3), (3.2) is as follows:

- if $w_2^l > 0$, a rarefaction followed by a phase transition;
- if $w_2^l = 0$, a single phase transition acting as a contact discontinuity;
- if $w_2^l < 0$, a single shock-like phase transition.

In all cases, the speed of the phase boundary is assigned by the Rankine–Hugoniot conditions. The size of this wave is $\Sigma_1(w^r, w^l) = w_1^r - w_1^l$.

The generalized 2-curve through w^l, for $w^l \in \Omega_c$, coincides with the standard Lax curve; see (3.1). Hence, if w^r lies on the 2-curve exiting w^l, the solution to (2.3), (3.2) is the standard Lax solution, and to the wave connecting w^l to w^r we assign size $\Sigma_2(w^r, w^l) = w_2^r - w_2^l$.

Now let $w^l \in \Omega_f$. We formally assign the standard Lax curves of the scalar equation (2.1) to the second family, defining the wave size as $\Sigma_2(w^r, w^l) = w_2^r - w_2^l$.

Concerning the generalized 1-curve through w^l, with $w^l \in \Omega_f$, the following cases are all the possibilities:

1. If $w_2^l \geq \check{w}_2$, the generalized 1-curve consists of w^l and of the segment $w_2 = w_2^l$ in Ω_c. If w^r belongs to it, the wave size is measured by $\Sigma_1(w^r, w^l) = w_2^r - w_2^l$, and the solution to (2.3), (3.2) is
   - if $w_2^l \in [0, \check{w}_1]$, a shock-like phase transition;
   - if $w_2^l = 0$, a single phase boundary acting as a contact discontinuity;
   - if $w_2^l < 0$, a phase boundary followed by a rarefaction wave.
2. If $w_2^l < \check{w}_2$, then the generalized 1-curve is the lower side of Ω_c, i.e., the segment $w_2 = \frac{Q_1 - q_*}{R}$, $w_1 \in [0, \hat{V}_c]$. Let $q_m(\rho) = q_* + \frac{\rho}{R}(Q_1 - q_*)$ be the equation

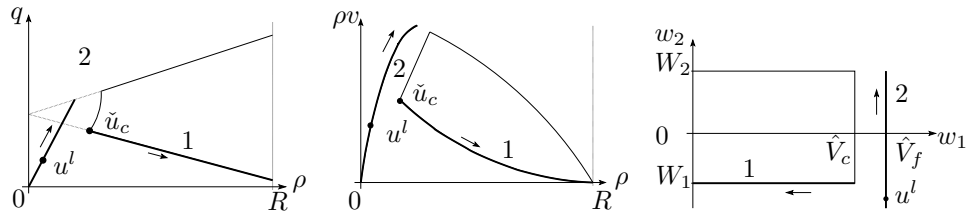


FIG. 3.2. *Generalized Lax curves exiting u^l, with $u^l \in \Omega_f$.*

of this segment in the (ρ, q) coordinates, and consider the two functions

$$\bar{\Lambda}(\rho) = \frac{\rho^l v^l - \rho v_c(\rho, q_m(\rho))}{\rho^l - \rho} \quad \text{and} \quad \bar{\lambda}_1(\rho) = \lambda_1(\rho, q_m(\rho)).$$

The former is the speed of the phase boundary joining $u^l \in \Omega_f$ to $(\rho, q_m(\rho)) \in \Omega_c$, while the latter is the first characteristic speed of $(\rho, q_m(\rho)) \in \Omega_c$. Consider the following three subcases:

- $\bar{\Lambda}(\check{R}_c) \le \bar{\lambda}_1(\check{R}_c)$. Here the solution to (2.3), (3.2) is a phase transition from u^l to $\check{u}_c$, followed by a 1-Lax rarefaction from $\check{u}_c$ to u^r.
- $\bar{\Lambda}(\rho) > \bar{\lambda}_1(\rho)$ for all ρ with $\check{R}_c \le \rho \le \rho^r$. In this case the solution consists of a shock-like phase transition.
- Otherwise, let ρ_m be the smallest car density such that $(\rho_m, q_m(\rho_m)) \in \Omega_c$ with $\bar{\Lambda}(\rho_m) = \bar{\lambda}_1(\rho_m)$. The solution to (2.3), (3.2) is a compound wave composed first by a phase transition and, attached to it, by a 1-Lax rarefaction in Ω_c.

This 1-wave (see Figure 3.2) is assigned the total size

$$\Sigma_1(w^r, w^l) = w_2^r - w_2^l + w_1^l - w_1^r. \tag{3.5}$$

C. The Riemann solver. From the above construction it follows that, given any two points w^l, w^r in $(\Omega_f \cup \Omega_c)^2$, the generalized curves defined above intersect in a single point $w^m \in \Omega_f \cup \Omega_c$, proving (2). The invariance of Ω_f (resp., Ω_c) with respect to (2.1) (resp., (2.2)) ensures (3). Furthermore, the wave speeds are well ordered, in the sense that any 1-generalized wave (eventually containing the phase boundary) propagates more slowly than the generalized 2-wave that follows, since in Ω_c

$$\lambda_2(\rho, q) - \lambda_1(\rho, q) = \frac{q}{R} + \left(\frac{1}{\rho} - \frac{1}{R}\right) q_* > 0, \tag{3.6}$$

which implies (0). Moreover, in the $(\rho, \rho v)$-plane, the Rankine–Hugoniot speed of the phase boundary connecting u^l to u^r is the slope of the segment joining $(\rho^l, \rho^l v^l)$ to $(\rho^r, \rho^r v^r)$. This shows that also, in the case of phase boundaries, any 1-wave reaching u_m is slower than any 2-wave exiting from w^m.

Note that the fastest wave in the solution to (2.3), (3.2) has maximal speed bounded above by the characteristic speed of the state on the right, i.e., by the traffic speed, which shows (1).

Finally, (4) follows immediately from the above construction—for example, from the presence of 3 waves in the solutions to some Riemann problems; see Figure 3.3. ∎

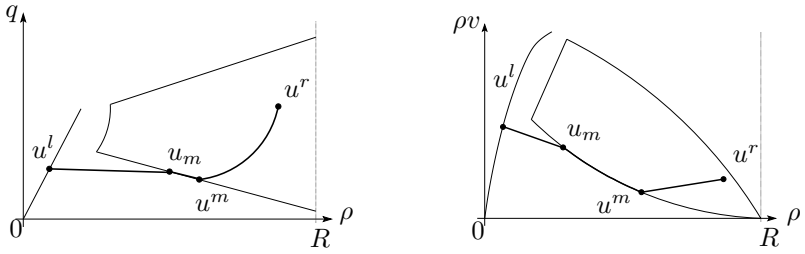


FIG. 3.3. *The solution to a sample Riemann problem.*

4. Qualitative properties of the solutions. In [21] a traffic light at $x = 0$ that turns red is simulated through the restriction to the quadrant $t \geq 0$, $x \leq 0$ of the solution to the Riemann problem for (2.1) with initial data

$$\rho(0,x) = \begin{cases} \rho_i & \text{if } x < 0, \\ R & \text{if } x > 0. \end{cases}$$

The solution is a shock with negative propagation speed. The location of this shock is the end of the queue of cars; each driver, as soon as he reaches it, brakes and immediately stops the car. See Figure 4.1, left.

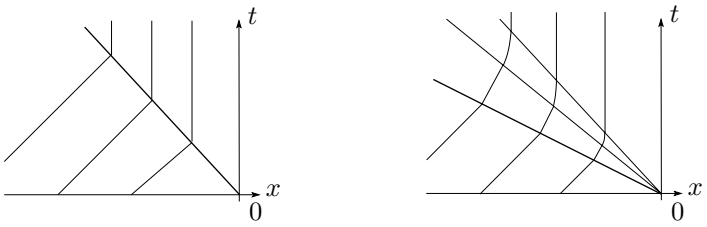


FIG. 4.1. *Car paths when the traffic light turns red.*

We now follow the same scheme but in the framework of (2.3). Assume $Q_1 < q_*$ and assign the initial data

$$(\rho, q)(0,x) = \begin{cases} (\rho_i, q_i) & \text{if } x < 0, \\ (R, q) & \text{if } x > 0, \end{cases}$$

with $(\rho_i, q_i) \in \Omega_f$ and $q \in [Q_1, Q_2]$. Computations based on the solution to Riemann problems defined above show that there exist two threshold parameters ρ_i^- and ρ_i^+ such that

- if $\rho_i \leq \rho_i^-$ or $\rho_i \geq \rho_i^+$, the same behavior as in (2.1) is obtained (see Figure 4.1, left);
- if $\rho_i^- < \rho_i < \rho_i^+$, the solution consists of a phase transition followed by a rarefaction possibly attached to it (see Figure 4.1, right).

Reasonably enough, the LWR description of drivers braking suddenly to zero speed works when there are either many slow cars or few fast cars. In the middle situation, drivers brake and the traffic flow enters the congested region. Here, the car speed continues to diminish, but smoothly.

It is remarkable to note that the description above holds *independently* from the value of $q \in [Q_1, Q_2]$ assigned to the state on the right.

In the case of the traffic light turning green, the present model gives a description similar to that provided by the LWR model.

In Kerner's paper [15] there is a further distinction within the congested phase, namely "synchronized traffic flow" and "wide moving traffic jams." However, he also claims, "On the flow density plane measured points related to fronts of wide moving jams usually cannot be separated from points related to synchronized traffic flow" [15, p. 265]. Usually, in the hyperbolic phase transition models common in continuum dynamics, the term *phase* characterizes suitable subsets of the space where the conserved quantities (or equivalent coordinates) may vary, with *different* phases corresponding to *disjoint* sets.

On the other hand, the observations by Kerner do show the existence of two qualitatively deeply distinct behaviors. This distinction is present also in (2.3): it corresponds to the distinction between waves of the first and second families.

More precisely, "homogeneous-in-speed-states" [15, p. 260], i.e., a type of synchronized flow, are described by states separated by a 2-contact discontinuity in the congested phase. In fact, recall that the traffic speed does not change across a 2-contact discontinuity; see (3.1).

Furthermore, the solid line J shown in Figure 4.2, left, which represents in [15] the "wide moving traffic jam," is here replaced by the 1-Lax curves near the line $q = q_*$, shown in Figure 4.2, right. In fact, along $q = q_*$ we have $\nabla\lambda_1 \cdot r_1 \equiv 0$ (see (3.1)), and the first characteristic family is linearly degenerate. Hence, near to that line, where $\nabla\lambda_1 \cdot r_1$ is small, a sharp increase in the density may persist for a long time.

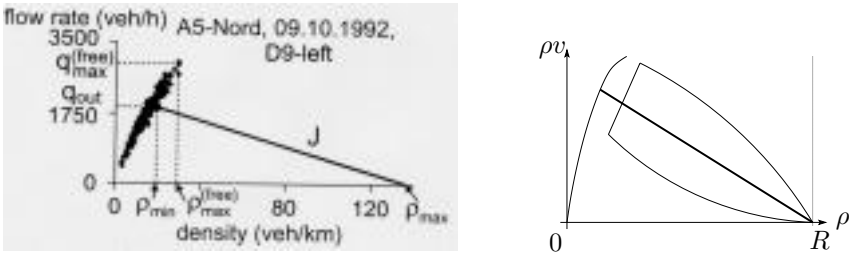


FIG. 4.2. *The transition from free to wide jam. Left, measured data from* [15] *(used with permission) and, right, the line where* $\nabla\lambda_1 \cdot r_1 = 0$ *in the model described here.*

Finally, let us remark that the present model (2.3) describes how a congested zone disappears into a free one due to the interaction of two phase boundaries. This may happen only if behind the congested zone the density is very low, i.e., $\rho^l < \check{R}_f$; see Figure 4.3. Analogously, an interaction between two phase boundaries where $\rho > \check{R}_f$ may lead to a free zone disappearing.

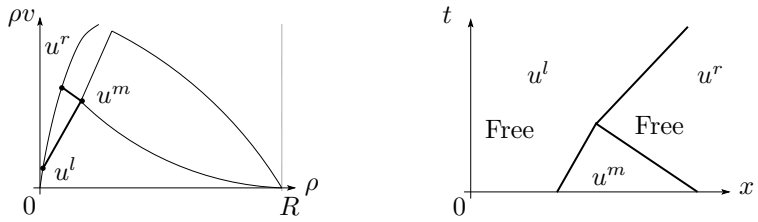


FIG. 4.3. *An interaction between phase boundaries.*

5. The Cauchy problem. Consider the set $\mathcal{U}$ of $\mathbf{BV} \cap \mathbf{L^1}$ functions $u \colon \mathbf{R} \mapsto \Omega_f \cup \Omega_c$ such that there exists a partition of $\mathbf{R}$ into a finite number of disjoint intervals $I_1, \dots, I_n$, with $u(I_i) \subseteq \Omega_f$ and $u(I_{i+1}) \subseteq \Omega_c$ for all i. The more general case in which $(u - u_{-\infty}) \in \mathbf{L^1}$ and $(u - u_{+\infty}) \in \mathbf{L^1}$ for some fixed states $u_{-\infty}$, $u_{+\infty}$ is a straightforward extension involving only formal complications.

Given a function $u \in \mathcal{U}$, we say that there is a *change of phase* at all points $\xi \in \mathbf{R}$ where $u(\xi-)$ and $u(\xi+)$ belong to different phases.

Let $u_o \in \mathcal{U}$. Generalizing the corresponding definition in [6], we call an *admissible solution* of the Cauchy problem for (2.3) on the time interval $[0, T]$, with $0 < T \leq +\infty$, any function $u \colon [0, T] \mapsto \mathcal{U}$ such that

1. u is continuous with respect to the $\mathbf{L^1}$-norm;
2. for all test functions $\phi \in \mathbf{C^1}(\mathbf{R} \mapsto \mathbf{R})$ with compact support contained in $u^{-1}(\Omega_f)$

$$\int_0^T \int_{\mathbf{R}} \left(\rho \cdot \partial_t \phi + \rho \cdot v_f(\rho) \cdot \partial_x \phi \right) \, dx \, dt + \int_{\mathbf{R}} \rho_o(x) \cdot \phi(0, x) \, dx = 0;$$

3. for all test functions $\phi \in \mathbf{C^1}(\mathbf{R} \mapsto \mathbf{R}^2)$ with compact support contained in $u^{-1}(\Omega_c)$

$$\int_0^T \int_{\mathbf{R}} \left(\begin{bmatrix} \rho \\ q \end{bmatrix} \partial_t \phi + \begin{bmatrix} \rho \cdot v_c(\rho, q) \\ (q - q_*) \cdot v_c(\rho, q) \end{bmatrix} \partial_x \phi \right) dx \, dt + \int_{\mathbf{R}} \begin{bmatrix} \rho_o(x) \\ q_o(x) \end{bmatrix} \phi(0, x) \, dx = 0;$$

4. the set of points at which there is a change of phase is the union of a finite number of Lipschitz curves $p_i \colon [0, T] \mapsto \mathbf{R}$ such that if $p_i(\tau) = p_j(\tau)$, then $p_i(t) = p_j(t)$ for all $t \in [\tau, T]$;
5. at all points $(\bar t, \bar x)$ where there is a change of phase, i.e., $\bar x = p_i(\bar t)$, let $\Lambda = \dot p_i(t+)$ and introduce the left and right flows at $(\bar t, \bar x)$ as

$$F^l = \begin{cases} \rho(\bar t, \bar x-) \cdot v_f(\rho(\bar t, \bar x-)) & \text{if } \rho(\bar t, \bar x-) \in \Omega_f, \\ \rho(\bar t, \bar x-) \cdot v_c(\rho(\bar t, \bar x-), q(\bar t, \bar x-)) & \text{if } \rho(\bar t, \bar x-) \in \Omega_c, \end{cases}$$

$$F^r = \begin{cases} \rho(\bar t, \bar x+) \cdot v_f(\rho(\bar t, \bar x+)) & \text{if } \rho(\bar t, \bar x+) \in \Omega_f, \\ \rho(\bar t, \bar x-) \cdot v_c(\rho(\bar t, \bar x+), q(\bar t, \bar x+)) & \text{if } \rho(\bar t, \bar x+) \in \Omega_c. \end{cases}$$

We require that

$$\Lambda \cdot (\rho(\bar t, \bar x+) - \rho(\bar t, \bar x-)) = F^r - F^l.$$

We prove below that (2.3) admits a solution for all initial data in $\mathbf{BV}$.

THEOREM 5.1. *For all $u_o \in \mathbf{BV}$, the problem* (2.3) *admits a solution $u \colon [0, +\infty[\times \mathbf{R} \mapsto \Omega_f \cup \Omega_c$ such that $u(0, x) = u_o(x)$.*

The proof follows the currently standard wave-front tracking machinery already used in various works; see [4, 7] and the references in [6]. We briefly recall below the general construction, underlining only those details specific to the present construction.

For all $u_o \in \Omega_f \cup \Omega_c$, let $\Psi_i(u_o, \sigma)$ denote the point u on the generalized i-curve exiting u_o as defined in section 3 such that the wave from u_o to u has size $\sigma = \Sigma_i(u, u_o)$.

Fix a large $n \in \mathbf{N}$ and introduce the mesh sizes

$$\delta_1^n = \frac{1}{2^n} \cdot \hat{V}_c, \qquad \delta_2^n = \frac{1}{2^n} \cdot \frac{Q_2 - Q_1}{R}, \qquad \bar{\delta}_2^n = \frac{1}{2^n} \cdot \left(v_f(0) - \check{w}_2\right).$$

Let

$$\begin{aligned}
\Omega_f^n &= \left\{u \in \Omega_f \colon w_1 = V \text{ and } w_2 = W_1 + k\delta_2^n \text{ or } w_2 = v_f(\check{\rho}) + k\bar{\delta}_2^n \,;\, k \in \mathbf{N}\right\},\\
\Omega_c^n &= \left\{u \in \Omega_f \colon w_1 = h\delta_1^n \text{ and } w_2 = W_1 + k\delta_2^n \,;\, h, k \in \mathbf{N}\right\},\\
\Omega^n &= \Omega_f^n \cup \Omega_c^n.
\end{aligned}$$

Approximate the initial data u_o by means of a piecewise constant initial datum u_o^n attaining values in Ω^n and such that $\|u_o^n - u_o\|_{\mathbf{L^1}} \leq 2^{-n}$. At every point of jump of u_o^n we approximately solve the Riemann problem within the class of piecewise constant functions. To this aim, we define an approximate solver as follows. Shocks, contact discontinuities, and phase boundaries are solved exactly, i.e., they are given the exact Rankine–Hugoniot speed. Centered rarefaction waves are approximated through rarefaction fans. Let $u^l \in \Omega_n$ and $u^r \in \Omega_n$ be connected by a 1-rarefaction, i.e., $u^r = \Psi_1(u^l, k2^{-n})$, with $\lambda_1(u^l) < \lambda_1(u^r)$; then the approximate solution to the Riemann problem with data u^l, u^r is a fan of rarefaction shock attaining values in Ω^n. Differently from the usual wave-front tracking algorithms, to each of these "shocklets" we assign the Rankine–Hugoniot speed $\lambda = \frac{[\rho v]}{[\rho]}$. The approximation of compound waves is obtained by gluing the solutions above. We thus obtain an approximate solution

$$u^n(t) = \sum_\alpha u_\alpha \cdot \chi_{]x_{\alpha-1}(t), x_\alpha(t)]} \quad \text{with} \quad u_{\alpha+1} = \Psi_2(\Psi_1\left(u_\alpha, \sigma_{1,\alpha}\right), \sigma_{2,\alpha}), \tag{5.1}$$

where $u_\alpha \in \Omega^n$, $x_{\alpha-1}(t) < x_\alpha(t)$ for all α and t. $\sigma_{i,\alpha}$ is the size of the i-wave in the solution of the Riemann problem (2.3) with data u_α, $u_{\alpha-1}$.

A global approximate solution is obtained as soon as suitable bounds on the total variation and on the number of discontinuities are provided. To this aim we introduce the following Glimm functional defined on all functions of the type (5.1):

$$\mathcal{V}(u) = \sum_\alpha \sum_{i=1}^{2} \left|\sigma_{i,\alpha}\right|. \tag{5.2}$$

The interaction estimates provided in section 6 lead to the following result.

PROPOSITION 5.2. *Fix $n \in \mathbf{N}$. Then for all piecewise constant initial data u_o attaining values in Ω_n such that $\mathrm{TV}(u_o) < +\infty$ there exists an approximate solution $u^n \colon [0, +\infty[\times \mathbf{R} \mapsto \Omega_n$ with the following properties:*

(1) *the function $t \mapsto \mathcal{V}(u^n(t))$ is nonincreasing;*

(2) *any strip of the form $[0, T] \times \mathbf{R}$ contains finitely many interaction points.*

The proof follows from Lemmas 6.1 and 6.2 below.

A standard [5, 6, 7, 10] limiting procedure based on Helly's compactness theorem leads to the existence of a global weak solution to (2.3).

6. Technical details. We collect here technical results useful in the preceding sections.

LEMMA 6.1. *Assume that the approximate solution u^n is defined up to time T. Then the map $t \mapsto \mathcal{V}(u^n(t))$ is nonincreasing on the interval $[0, T]$.*

Proof. Assume that only two waves interact. Consider the following cases:

1. The interacting waves σ', σ'' both belong to the second family. Then there is a single outgoing wave σ^+, and $\sigma^+ = \sigma' + \sigma''$.
2. The interacting waves σ', σ'' both belong to the first family, and no 2-wave exits the interaction. Then, again, the resulting 1-wave satisfies $\sigma^+ = \sigma' + \sigma''$.
3. The interacting waves σ', σ'' both belong to the first family, and the resulting wave σ^+ is a 2-wave. Then the choice (3.5) again implies that $\sigma^+ = \sigma' + \sigma''$.
4. The interacting waves σ_1^- and σ_2^- belong to different families. Then the resulting waves σ_1^+, σ_2^+ satisfy $\sigma_i^+ = \sigma_i^-$.

In all the cases above, $\mathcal{V}$ does not increase at the interaction time.

The general case of an interaction in which several waves take part follows by standard arguments. □

The presence of the line $q = q_*$, where $\nabla\lambda_1 \cdot r_1$ vanishes, makes the control of the number of discontinuities more delicate. Indeed, the interaction of two waves, a 1-contact discontinuity, and a 2-shock may produce a new 1-contact and $\mathcal{O}(2^n)$ rarefaction wavelets of the first family.

LEMMA 6.2. *Given an approximate solution defined up to time T, any strip of the form $[0,T] \times \mathbf{R}$ contains finitely many interaction points.*

Proof. The proof follows from Lemma 6.1, since the number of waves in u^n at time t is bounded by $\mathcal{O}(1) \cdot 2^n \cdot \mathcal{V}(u^n(t))$. □

Remark 1. In the case $Q_1 \geq q_*$, the solution to the Riemann problem (2.3), (3.2) is simpler. In fact, all the phase transitions along the lower border of Ω_c behave like shocks if $Q_1 > q_*$, and like contact discontinuities if $Q_1 = q_*$. Furthermore, the number of discontinuities may not increase at any interaction.

Remark 2. Allowing $\hat{V}_f = \hat{V}_c$ leads to the non–well-posedness of the Riemann problem. Let the left state be $\check{u}_c$ and the right state be $\hat{u}_f = \hat{u}_c$; see Figure 2.1. Then, a first solution to (2.3), (3.2) consists of a single Lax 2-wave in the congested phase. Another solution consists of a phase transition from $\check{u}_c$ to $\check{u}_f$ followed by a Lax 2-wave in the free phase.

Acknowledgments. The present work was partly accomplished while the author was visiting the Department of Mathematics at Stanford University. The author thanks Barbara Lee Keyfitz for her very careful editing work.

REFERENCES

[1] R. ABEYARATNE AND J. K. KNOWLES, *Kinetic relations and the propagation of phase boundaries in solids*, Arch. Ration. Mech. Anal., 114 (1991), pp. 119–154.
[2] A. AW AND M. RASCLE, *Resurrection of "second order" models of traffic flow*, SIAM J. Appl. Math., 60 (2000), pp. 916–938.
[3] S. BIANCHINI, *The semigroup generated by a Temple class system with non-convex flux function*, Differential Integral Equations, 13 (2000), pp. 1529–1550.
[4] A. BRESSAN, *Global solutions of systems of conservation laws by wave-front tracking*, J. Math. Anal. Appl., 170 (1992), pp. 414–432.
[5] A. BRESSAN, *The unique limit of the Glimm scheme*, Arch. Ration. Mech. Anal., 130 (1995), pp. 205–230.
[6] A. BRESSAN, *Hyperbolic Systems of Conservation Laws*, Oxford University Press, London, UK, 2000.
[7] A. BRESSAN AND R. M. COLOMBO, *The semigroup generated by 2×2 conservation laws*, Arch. Ration. Mech. Anal., 133 (1995), pp. 1–75.
[8] R. M. COLOMBO, *On a 2×2 hyperbolic traffic flow model*, Math. Comput. Modelling, 35 (2002), pp. 683–688.
[9] R. M. COLOMBO AND A. CORLI, *Sonic hyperbolic phase transition and Chapman–Jouguet detonations*, J. Differential Equations, 184 (2002), pp. 321–347.

[10] C. M. Dafermos, *Hyperbolic Conservation Laws in Continuum Physics*, 1st ed., Springer-Verlag, New York, 2000.
[11] C. F. Daganzo, *Requiem for high-order fluid approximations of traffic flow*, Transportation Res., 29 (1995), pp. 277–287.
[12] D. Hoff, *Invariant regions for systems of conservation laws*, Trans. Amer. Math. Soc., 289 (1985), pp. 591–610.
[13] B. Kerner, *Experimental features of self-organization in traffic flow*, Phys. Rev. Lett., 81 (1998), pp. 3797–3800.
[14] B. Kerner, *The physics of traffic*, Physics World, 12 (1999), pp. 25–30.
[15] B. Kerner, *Phase transitions in traffic flow*, in Traffic and Granular Flow '99, D. Helbing, H. Hermann, M. Schreckenberg, and D. E. Wolf, eds., Springer-Verlag, New York, 2000, pp. 253–283.
[16] P. D. Lax, *Hyperbolic systems of conservation laws.* II, Comm. Pure Appl. Math., 10 (1957), pp. 537–566.
[17] M. J. Lighthill and G. B. Whitham, *On kinematic waves.* II. *A theory of traffic flow on long crowded roads*, Proc. Roy. Soc. London. Ser. A., 229 (1955), pp. 317–345.
[18] P. I. Richards, *Shock waves on the highway*, Oper. Res., 4 (1956), pp. 42–51.
[19] D. Serre, *Systems of Conservation Laws, Vol.* 1. *Hyperbolicity, Entropies, Shock Waves*, Cambridge University Press, Cambridge, 1999; translated from the 1996 French original by I. N. Sneddon.
[20] M. Slemrod, *Admissibility criteria for propagating phase boundaries in a van der Waals fluid*, Arch. Ration. Mech. Anal., 81 (1983), pp. 301–315.
[21] G. B. Whitham, *Linear and Nonlinear Waves*, John Wiley and Sons, New York, 1999.
[22] E. C. Zachmanoglou and D. W. Thoe, *Introduction to Partial Differential Equations with Applications*, 2nd ed., Dover Publications, New York, 1986.

SIAM J. APPL. MATH.
Vol. 63, No. 2, pp. 722–744

THIN FILM TRAVELING WAVES AND THE NAVIER SLIP CONDITION*

ROBERT BUCKINGHAM†, MICHAEL SHEARER‡, AND ANDREA BERTOZZI§

Abstract. We consider the lubrication model for a thin film driven by competing gravitational forces and thermal gradients on an inclined plane. We are interested in the general traveling wave problem when the Navier slip boundary condition is used. We contrast (1) gravity dominated flow, (2) Marangoni dominated flow, and (3) flow in which the two driving effects balance. For a "singular slip" model we show that when Marangoni forces are present the resulting traveling wave ODE reduces locally near the contact line to a case not considered previously in the literature. We compute an asymptotic expansion of the solution near the contact line and compare with numerical simulations of the full problem. Using numerical simulations and phase space analysis involving Poincaré sections, we show that for all three problems there is a finite range of admissible contact angles for which traveling wave solutions exist. Even in the well-studied case (1), this is a new observation that has ramifications for the use of constitutive laws at the contact line in the case of singular slip. For case (3) multiple traveling wave solutions are observed with the same contact angle.

Key words. thin liquid films, contact lines, traveling waves, nonlinear partial differential equations

AMS subject classifications. 35K55, 35Q35, 76D08, 76D45, 34C37

PII. S0036139902401409

Introduction. Dynamic contact lines occur at the leading edge of a layer of fluid coating a dry solid surface. Understanding how contact lines move has been the subject of intense interest for several decades. In particular, it was shown by Dussan and Davis [13] that motion necessarily implies a singularity of stress at the contact line if the usual no-slip boundary condition is imposed between the fluid and the solid surface. Two approaches to removing this singularity emerged early on, namely (i) the precursor layer model and (ii) the Navier slip condition.

In 1964, Bascom, Cottington, and Singleterry [1] reported experimental observations of contact lines for thin liquid films. A very thin film was observed spreading ahead of the thicker film, beyond the apparent location of the contact line. Based on these and similar observations, one reasonable model is to assume that there is a very thin layer of fluid ahead of the contact line. The contact line itself is then replaced by a rapid transition from the thicker layer to the very thin layer. This is the basis for the so-called *precursor model* studied in various contexts over a number of years. While this is an attractive and tractable way to remove the singularity associated with the film thickness going all the way to zero, modeling the very thin precursor layer using hydrodynamics can be questionable since its thickness is only a

*Received by the editors January 23, 2002; accepted for publication (in revised form) June 24, 2002; published electronically January 14, 2003.
http://www.siam.org/journals/siap/63-2/40140.html
†Department of Mathematics, Duke University, Durham, NC 27708 (robbie@math.duke.edu). The research of this author was supported by NSF grants DMS-9983320 and DMS-0074049.
‡Center for Research in Scientific Computation and Department of Mathematics, North Carolina State University, Raleigh, NC 27695-8205 (shearer@math.ncsu.edu). The research of this author was supported by National Science Foundation grant DMS-0073841 and by Army Research Office grant DAAG55-98-10128.
§Departments of Mathematics and Physics, Duke University, Durham, NC 27708 (bertozzi@math.duke.edu). The research of this author was supported by NSF grant DMS-0074049 and ONR grant N000140110290.

few Ångstroms. Within a continuum framework, this can be approached by including additional physics, such as long range van der Waals forces [12, 34, 41] in a dynamic model of the precursor layer itself (as in [16]).

The second approach that has received extensive study is to keep the distinct contact line but remove the stress singularity by modifying the boundary condition between the liquid and the solid surface at or near the contact line. The most promising way to do this is to introduce the *Navier slip condition*, proposed by Navier [32] in 1832 during the long debate over whether a fluid can slide over a solid surface [15]. The Navier condition was apparently first invoked in the context of lubrication theory by Greenspan [17] and has since been included in numerous studies of contact lines [2, 22, 31, 38, 39]. Again, this is an attempt to resolve the issue using only hydrodynamics, rather than dealing with the atomic scale forces that are undoubtedly significant near the contact line. Nonetheless, it is plausible to believe that the effect of these forces at the macroscopic scale could be captured in an empirical law like the Navier slip condition. Once a choice of slip model is made, there is still the question of the need for a boundary condition at the contact line. Greenspan [17] proposed that the speed of the contact line is related to the contact angle. This was further considered for spreading drops by Haley and Miksis [18] and Ehrhard and Davis [14]. Hocking [23] considers using the static contact angle for the dynamic problem. In the case of complete wetting, a zero contact angle solution is preferred. For the general lubrication PDE, existence of such "zero contact angle" solutions with slip was proved rigorously in one space dimension [3, 8]. In this case, the zero contact angle condition replaces the boundary condition or fixed contact angle condition at the contact line. A natural question, which we address, is whether the PDE admits traveling wave solutions with a prescribed nonzero contact angle condition.

In this paper, we consider a thin film being driven up an inclined solid surface by a surface or Marangoni stress. These driven films have been studied extensively theoretically, experimentally, and numerically [5, 6, 10, 11, 27, 28, 29, 37, 36]. In particular, in a series of papers on the precursor model, we found interesting novel structures for traveling waves and their stability [4, 6, 7]. Our purpose in this paper is to explore the Navier slip model as an alternative to the precursor model for Marangoni driven films. Among other conclusions, we find that for a given slip length and film thickness, there is a finite range of contact angles that admit traveling wave solutions.

In section 2, we give a brief derivation of the fourth order nonlinear PDE governing the evolution of film height, using the lubrication approximation to the Navier–Stokes equations for two-dimensional incompressible flow. This derivation shows how the Navier slip condition enters the thin film PDE. Also in section 2 we show how traveling wave solutions with a contact line satisfy a third order ODE, in which the traveling wave speed is determined by the upstream height. The PDE and associated traveling wave ODE can be used to study contact lines under three scenarios, each of which we consider in this paper:

I. Flow in which gravity is the only driving force. For example, a layer of fluid wetting a dry surface as it slides down the surface under the action of gravity.

II. Flow in which the Marangoni force dominates gravity.

III. Flow in which Marangoni force and the force due to gravity are balanced.

The main results of this paper, both analytical (in section 3) and numerical (in section 4) concern flow in which forces are balanced, but our numerical results add something to the understanding of the first two scenarios as well.

In section 3 we analyze the traveling wave ODE near the contact line when Marangoni forces are present, under the Navier slip condition. Curiously, the leading order terms of the ODE are independent of the precise form of the Navier slip condition; we are led to the problem of finding a function $y = y(x)$ (representing the film height) that is positive for $x > 0$ and satisfies

$$yy''' = 1, \quad x > 0, \quad y(0) = 0. \tag{1.1}$$

This equation would appear to fall under the extensive classification of solutions of the more general third order ODE $y^n y''' = 1$ contained in the paper of Boatto, Kadanoff, and Olla [9]. However, we observe that the case $n = 1$ is special, and we find a different structure for the solutions. In section 3.1, we find a two-parameter family of asymptotic solutions

$$y(x) = ax + \frac{1}{2a}x^2 \log x + bx^2 + \text{h.o.t.}, \tag{1.2}$$

where h.o.t. denotes higher order terms and $a > 0$ and b are free parameters. We show how the series can be continued indefinitely, and in section 3.1 we perform a reduction to a polynomial planar vector field that establishes the dimension of the solution set.

In section 4 we present various results of numerical integration of the third order traveling wave ODE. The technique we use is similar to that in earlier studies of the precursor layer model [6], but here we show how the entire phase space can be understood by focusing on the structure of stable and unstable invariant manifolds associated with equilibria. Phase portraits are three-dimensional, since the ODE is third order, so we visualize invariant manifolds through their intersection with carefully chosen Poincaré sections.

The numerical solutions are compared with the asymptotic form (1.2) near the contact line. The numerical results highlight various interesting issues. We find a finite range of contact angles for each wave speed. This has relevance for the use of a boundary condition relating contact angle and wave speed, as considered in [14, 18], or for the case of a fixed contact angle condition, as considered in [31, 38] for singular slip and in [20, 21, 23] for nonsingular slip. Moreover, at a given speed and at a given contact angle in this range, there may be several different traveling waves. The latter property is specific to the case in which gravity and Marangoni effects balance and is related to the nonconvexity of the flux function in the lubrication model. This particular effect is well understood for the same problem in which a simpler precursor film model is used to remove the contact line singularity [6].

2. The lubrication approximation and traveling waves. In section 2.1, we outline the lubrication approximation and formulate the PDE that governs the motion of the thin liquid layer, including the Navier slip condition. In section 2.2, we derive a third order ODE whose solutions are traveling wave solutions of the PDE.

2.1. The thin film PDE. Consider a thin liquid film moving slowly up a flat solid surface, inclined at an angle α to the horizontal. The film is driven by a constant surface stress τ, and gravity also acts on the film, as indicated in Figure 2.1. We shall consider the film to be uniform in the transverse direction. This means that the transverse velocity is zero, the in-plane velocity (u, v) and pressure p are functions of x, z, and time t, and the free surface is given by $z = h(x, t)$, where h is a function to be determined. The lubrication approximation reduces the description of the flow to a PDE for $h(x, t)$.

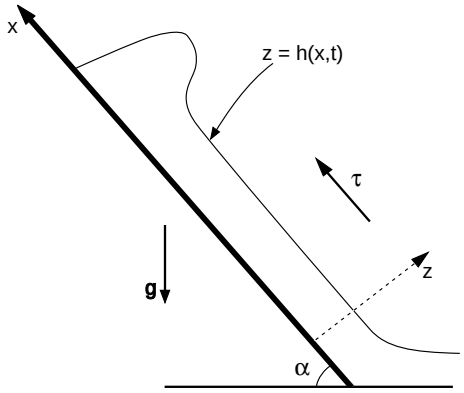


FIG. 2.1. *Thin film propagating up an inclined solid surface.*

The Navier–Stokes equations for the two-dimensional flow of the figure are as follows:

$$\begin{aligned} &\text{(a)} && \rho(u_t + uu_x + vu_z) = -p_x + \mu(u_{xx} + u_{zz}) - \rho g \sin\alpha, \\ &\text{(b)} && \rho(v_t + uv_x + vv_z) = -p_z + \mu(v_{xx} + v_{zz}) - \rho g \cos\alpha, \\ &\text{(c)} && u_x + v_z = 0. \end{aligned} \tag{2.1}$$

Here, ρ is the density, taken to be constant, consistent with the incompressibility condition (2.1c), and μ is the viscosity.

In the lubrication approximation, we exploit two small quantities to reduce the complexity of the equations, keeping only leading order terms but maintaining a balance between terms that are significant, namely surface stresses and viscous forces. Let H be a typical thickness of the film, say a maximum thickness. This is assumed small compared to a typical length scale L along the solid surface. The other small parameter is the Reynolds number $Re = \rho UH/\mu$, calculated with respect to the thickness length scale, but where U is a typical velocity in the x-direction parallel to the solid surface.

The lubrication approximation that emerges as the leading order terms consists of two equations, with unknowns velocity u parallel to the solid surface and pressure p (the normal velocity v is given separately by the incompressibility condition (2.1c)):

$$\begin{aligned} &\text{(a)} && p_x = \mu u_{zz} - \rho g \sin\alpha, \\ &\text{(b)} && p_z = -\rho g \cos\alpha. \end{aligned} \tag{2.2}$$

To this system we add boundary conditions at $z = 0$ and $z = h$:

$$\begin{aligned} &\text{(a)} && p = p_A - \gamma h_{xx} && \text{on} \quad z = h, \\ &\text{(b)} && \mu u_z = \tau && \text{on} \quad z = h, \\ &\text{(c)} && k(h)u_z = u && \text{on} \quad z = 0. \end{aligned} \tag{2.3}$$

Here, p_A denotes atmospheric pressure, γ is the coefficient of surface tension, taken

to be constant, τ is a constant surface stress[1] coupled to the flow by the rate of shear strain appearing on the left-hand side of (2.3b). (Other mechanisms produce a surface stress, such as gradients in the concentration of a surfactant or airflow over the surface.) The boundary condition (2.3c) expresses the Navier slip condition, in which $k(h)$ is a coefficient with dimension of length that becomes essentially zero away from the contact line (i.e., for $h > h_0$ where h_0 is small). As in earlier work on the Navier slip condition, we take $k(h)$ to be a smooth and positive function; it would be possible to cut $k(h)$ off at a specified distance from an advancing contact line. For simplicity we choose only smooth functions $k(h)$ as in [17, 18, 39, 31]. Specific forms for $k(h)$ will be given later.

To derive an equation for $h(x,t)$, we first integrate system (2.2) using the boundary conditions (2.3) to obtain an expression for u in terms of z and h. Then u is averaged across the film to get an average velocity Q, expressed entirely in terms of h and derivatives of h. Finally, this formula for Q is substituted into the equation

$$h_t + (hQ)_x = 0$$

expressing conservation of mass. This procedure, explained in detail and in greater generality elsewhere [17, 31], leads to the single equation

$$h_t + \left\{ \frac{\tau}{2\mu} q(h) - \frac{\rho g \sin\alpha}{3\mu} c(h) - \frac{\rho g \cos\alpha}{3\mu} c(h) h_x + \frac{\gamma}{3\mu} c(h) h_{xxx} \right\}_x = 0, \tag{2.4}$$

in which

$$q(h) = h^2 + 2hk(h), \quad c(h) = h^3 + 2h^2 k(h) \tag{2.5}$$

would be quadratic and cubic functions (respectively) were it not for the modifications from the Navier slip condition.

From (2.4) we can realize the three cases of the introduction:

1. Gravity dominates: $\tau = 0$.
2. Marangoni forces dominate: $\tau >> \rho g \sin\alpha$.
3. Gravity and Marangoni effects are in balance: $\tau \sim \rho g \sin\alpha$.

In each case, a slightly different scaling of the variables leads to nondimensional equations. We give the details in the third case, rescaling the variables as in [6]. We introduce length scales H, L and a corresponding time scale T:

$$h = H\hat{h}, \quad x = \hat{x}L, \quad \text{and} \quad t = T\hat{t}. \tag{2.6}$$

Balancing the competing convective effects of gravity and Marangoni forces in (2.4) gives $H = \frac{3\tau}{2\sin\alpha\rho g}$. Setting L to be the capillary length on which surface tension balances the driving forces gives $L = (\frac{2\gamma H^2}{3\tau})^{1/3} = (\frac{3\gamma\tau}{2\rho^2 g^2 \sin^2\ a})^{1/3}$. The time scale is then chosen to be the one on which all three of these effects balance, $T = 2\frac{\mu}{\tau^2}(\frac{4}{9}\tau\gamma\rho g\sin\alpha)^{1/3}$.

This leads to the equation

$$\begin{aligned} h_t + \left(\left(h^2 + \frac{2}{3} h K(h) \right) - (h^3 + h^2 K(h)) \right)_x \\ = D\left((h^3 + h^2 K(h)) h_x \right)_x - \left((h^3 + h^2 K(h)) h_{xxx} \right)_x, \end{aligned} \tag{2.7}$$

[1]Here, we assume a constant surface tension gradient, proportional to the constant temperature gradient in experiments [37, 36], in the regime in which surface tension depends linearly upon temperature.

where $D = \frac{\rho g \cos\alpha\, T H^3}{3\mu L}$ and

$$K(h) = \frac{3k(Hh)}{H}. \tag{2.8}$$

We remark that D is typically small and is zero for a vertical plane ($\alpha = \pi/2$). It will be convenient to label the flux $f(h)$ on the left-hand side of (2.7),

$$f(h) = \left(h^2 + \frac{2}{3}hK(h)\right) - (h^3 + h^2K(h)), \tag{2.9}$$

and to use the notation

$$C(h) = h^3 + h^2K(h). \tag{2.10}$$

Then the depth-averaged velocity Q is given by

$$Q = (f(h) - DC(h)h_x + C(h)h_{xxx})/h, \tag{2.11}$$

and the PDE (2.7) is

$$h_t + f(h)_x = D(C(h)h_x)_x - (C(h)h_{xxx})_x. \tag{2.12}$$

In cases 1 and 2, a similar rescaling leads to equations similar to (2.12) but with different flux functions $f(h)$ (see [31]). Specifically:

1. When gravity dominates, the thin film will be coating by flowing down the solid surface. Reversing x so that increasing x is in the direction of flow, we obtain (2.12) with

$$f(h) = h^3 + h^2K(h). \tag{2.13}$$

2. When Marangoni forces dominate, the gravity term in the flux drops out, and we are left with

$$f(h) = h^2 + hK(h). \tag{2.14}$$

2.2. Traveling waves. We seek traveling wave solutions of (2.12). These take the form $h(x,t) = \overline{h}(x - st)$, where $\overline{h}$ is a function of the single traveling wave variable $\xi = x - st$, and s is the wave speed. Substituting into the PDE, and integrating once, we obtain (dropping the bars)

$$E - sh + f(h) = DC(h)h' - C(h)h''', \tag{2.15}$$

in which $'$ denotes differentiation with respect to ξ, and E is the constant of integration.

Now we are interested particularly in solutions that have a contact line, at which $h = 0$. Since the ODE is autonomous, we can assume the contact line is at $\xi = 0$. Upstream (i.e., $\xi \longrightarrow -\infty$), we assume the traveling wave approaches a constant height, with at least the first three derivatives approaching zero. Thus, we have boundary conditions

$$h(-\infty) = h_-, \qquad h'(-\infty) = h''(-\infty) = h'''(-\infty) = 0, \qquad h(0) = 0. \tag{2.16}$$

Letting $\xi \longrightarrow -\infty$, we find $E = sh_- - f(h_-)$.

At the other end, letting $\xi \longrightarrow 0-$, we have $Q \longrightarrow s$. (I.e., the average speed approaches the speed of the traveling wave at the contact line.) Using (2.11), we can rewrite (2.15) as

$$-s + Q(h) = E/h.$$

Letting $h \longrightarrow 0$ leads to the conclusion $E = sh_- - f(h_-) = 0$. Thus, the wave speed *s is determined by the upstream height h_-*:

$$s = \frac{f(h_-)}{h_-}. \tag{2.17}$$

In conclusion, the traveling wave satisfies the ODE

$$C(h)h''' = sh - f(h) + DC(h)h', \tag{2.18}$$

with boundary conditions

$$h(-\infty) = h_-, \quad h(0) = 0. \tag{2.19}$$

Finally, we discuss the leading order terms at the contact line in (2.18). To do so, we need to specify the constitutive function $k(h)$ in the Navier slip condition (2.3c). The form of this function is not decided upon [17, 18], but the idea is that slip should be confined to a small neighborhood of the contact line, where h is very small. Thus, $k(h)$ should be chosen so that it is nearly zero unless h is very small. Typically, $k(h)$ is chosen to be a power of h; in order to satisfy the above requirement, this power should be negative. Thus we take

$$k(h) = \eta h^{n-2}, \tag{2.20}$$

with $n < 2$ and $\eta > 0$. In [17], the choice is $n = 1$. This slip model was derived by Neogi and Miller for flow over a porous surface [33]. Other choices are possible, including $n = 2$ which can model polymer flow [12]. Note that for $n > 2$, $k(h)$ grows away from the contact line. With the choice (2.20), the function $K(h)$ defined in (2.8) becomes

$$K(h) = \beta h^{n-2}, \tag{2.21}$$

with $\beta = 3\eta H^{n-3}$. In particular, there are two parameters in this relation, namely β and n.

Now consider the leading order terms as $h \longrightarrow 0$, i.e., near the contact line. We have $C(h) = h^3 + h^2K(h) \sim \beta h^n$. Asymptotics for $f(h)$ depend on which case we are considering. In case 1, in which gravity dominates, $f(h) = h^3 + h^2K(h) \sim \beta h^n$. In case 2 and case 3, the terms from the Marangoni force are higher order. Specifically in case 3, $f(h) = (h^2 + \frac{2}{3}hK(h)) - (h^3 + h^2K(h)) \sim \frac{2}{3}\beta h^{n-1}$.

Retaining leading order terms in the ODE, we get, in case 1, the equation

$$\beta h^{n-1}h''' = s + \beta h^{n-1} + D\beta h^{n-1}h'. \tag{2.22}$$

Thus, for case 1, the choice $n = 1$ leads to a constant coefficient equation. Notably, there is then no singularity at $h = 0$.

In cases 2 and 3, we obtain the singular equation

$$hh''' = -\frac{2}{3} + Dhh'. \tag{2.23}$$

At the contact line, we expect h' to be finite (although it would be reasonable to consider solutions with a vertical tangent at the contact line). Moreover, the parameter D will be considered small or zero. Thus, we take $Dhh' \sim 0$ to leading order. Note that in arriving at (2.23), we have depended on two important assumptions: (1) There is a surface driving force $\tau > 0$, and (2) $n < 2$ in (2.20).

Rescaling ξ in (2.23), and dropping the last term, we are led to consider the initial value problem

$$yy''' = 1, \quad y(x) > 0 \quad \text{for} \quad x > 0, \quad y(0) = 0, \tag{2.24}$$

in which $x = -(\frac{2}{3})^{1/3}\xi$ and $y(x) = h(\xi)$.

3. Solutions near the contact line. In this section we explore properties of the initial value problem (2.24). In subsection 3.1 we establish an asymptotic series solution that has two free parameters, and in subsection 3.2 we reduce the third order equation to a planar vector field, whose phase portrait proves the existence of a two-parameter family of solutions.

3.1. Asymptotics. In this subsection, we elaborate on the proposed family (1.2) of solutions of (2.24) and show how the terms of an asymptotic series can be calculated systematically. To this end, consider the series[2]

$$y(x) = ax + \frac{1}{2a}x^2 \log x + bx^2 + \Sigma(x), \tag{3.1}$$

where $\Sigma(x)$ is expressed as a series with coefficients to be determined:

$$\Sigma(x) = \sum_{k=3}^{\infty} \sum_{j=2}^{k} d_{k\,j} x^k (\log\, x)^{k-j}. \tag{3.2}$$

Note that the series is organized as a sum of terms of increasing order (as $x \to 0+$). We will show that the coefficients $d_{k\,j}$ may be calculated in the same order: $d_{32}, d_{33}, d_{42}, d_{43}, d_{44}, d_{52}, \ldots$. In what follows, it will be helpful to adopt the convention that $d_{km} = 0$ whenever $m \le 1$.

Consider a single term $z_{kj}(x) = x^k (\log\, x)^{k-j}$ with $k \ge 3, 2 \le j \le k$. Then

$$z_{kj}''' = x^{k-3} \left\{ A_k (\log\, x)^{k-j} + B_{kj} (\log\, x)^{k-j-1} + C_{kj} (\log\, x)^{k-j-2} + A_{k-j} (\log\, x)^{k-j-3} \right\},$$

where the coefficients A, B, C are nonnegative; most importantly $A_k > 0$ for $k \ge 3$. They are given by the formulae

$$A_k = \begin{cases} k(k-1)(k-2) & \text{if} \quad k \ge 3, \\ 0 & \text{if} \quad k \le 2, \end{cases} \qquad B_{kj} = \begin{cases} (3k^2 - 6k + 2)(k-j) & \text{if} \quad k > j, \\ 0 & \text{if} \quad k \le j, \end{cases}$$

$$C_{kj} = \begin{cases} 3(k-1)(k-j)(k-j-1) & \text{if} \quad k \ge j+2, \\ 0 & \text{if} \quad k \le j+1. \end{cases}$$

[2]Hocking [19] considered a leading order expansion of this form for a correction to the trailing edge of Huppert's solution [24] for flow down an inclined plane.

Thus

$$\begin{aligned} y''' &= \frac{1}{ax} + \sum_{k=3}^{\infty}\sum_{j=2}^{k} d_{kj} z'''_{kj} \\ &= \frac{1}{ax} + x^{-3}\sum_{k=3}^{\infty}\sum_{j=2}^{k} \alpha_{kj} x^k (\log\, x)^{k-j}, \end{aligned} \tag{3.3}$$

where the coefficients α_{kj} are linear combinations of the d_{kj}'s:

$$\alpha_{kj} = A_k d_{kj} + B_{kj-1} d_{kj-1} + C_{kj-2} d_{kj-2} + A_{k-(j-3)} d_{kj-3}. \tag{3.4}$$

(Recall $d_{km} = 0$ if $m \le 1$.)

Now substitute (3.1), (3.3) into (1.1):

$$\begin{aligned} &\left[ax + \frac{1}{2a}x^2 \log x + bx^2 + \sum_{k=3}^{\infty}\sum_{j=2}^{k} d_{kj} x^k (\log\, x)^{k-j}\right] \\ &\qquad \times \left[\frac{1}{ax} + \sum_{k=3}^{\infty}\sum_{j=2}^{k} \alpha_{kj} x^{k-3} (\log\, x)^{k-j}\right] = 1. \end{aligned} \tag{3.5}$$

Equating terms, we get a family of equations for the coefficients d_{kj} (recall the coefficients α_{kj} depend linearly on the d's):

$$\begin{aligned} x\log x: &\qquad \frac{1}{2a^2} + a\alpha_{32} = 0, \\ x: &\qquad \frac{b}{a} + a\alpha_{33} = 0, \end{aligned} \tag{3.6}$$

$x^{k-2}(\log x)^{k-j}$:

$$a\alpha_{kj} + \tfrac{1}{2a}\alpha_{k-1j} + b\alpha_{k-1j-1} + \tfrac{1}{a} d_{k-1j-1} + \sum_{m+p=k+1}\ \sum_{n+q=j+1} d_{mn}\alpha_{pq} = 0. \tag{3.7}$$

In the final sum, the additional constraints on the indices are implied from (3.5):

$$m \ge 3, \quad p \ge 3, \quad 2 \le n \le m, \quad 2 \le q \le p.$$

In particular, these imply

$$m \le k-2 \quad \text{and} \quad p \le k-2. \tag{3.8}$$

Also note that there is no contribution from the double sum if $k = 4$, or when $j = 2$.

Now from (3.4), we observe that the equation for the coefficient of $x^{k-2}(\log x)^{k-j}$ has a term $aA_k d_{kj}$, and the other terms involve d_{km} with $m \le j-1$ (these terms come from α_{kj}) and d_{mq} with $m \le k-1, 2 \le q \le m$. Consequently, the equations can be solved successively for the coefficients d_{kj} in their natural order associated with terms of increasing order in the asymptotic expansion. Thus, the asymptotic series can be continued to all orders and defines a two-parameter family of formal solutions of the ODE (1.1).

3.2. Reduction to a planar vector field. In this subsection, we reduce the ODE to a planar vector field that we analyze directly. While this is not a new technique (see, for example, [9, 40], where similar reductions are performed on other third order ODEs related to similarity solutions of thin film equations), the real interest lies in using the planar vector field to identify a family of solutions of the ODE with the two-parameter family of asymptotic solutions in the previous subsection.

Consider the ODE (1.1):

$$yy''' = 1. \tag{3.9}$$

Since this equation is autonomous and has a natural scaling invariance (scaling x by a^2 and y by a^3 leaves the equation unchanged), we can reduce the equation to a second order equation that is also autonomous. This is achieved by writing $w = y'$ and letting the independent variable be y. Thus, (3.9) becomes the second order equation

$$yw\frac{d}{dy}w\frac{dw}{dy} = 1. \tag{3.10}$$

Now $y\frac{d}{dy}$ is the logarithmic derivative, so we redefine the independent variable as $\eta = \log y$, leading to

$$w\frac{d}{d\eta}w\frac{dw}{d\eta} = w^2\frac{dw}{d\eta} + e^\eta. \tag{3.11}$$

Now we write this equation as a first order system and rescale to make it autonomous. First let $v = w\frac{dw}{d\eta}$. Then

$$w\frac{dv}{d\eta} = e^\eta + vw, \quad w\frac{dw}{d\eta} = v. \tag{3.12}$$

Now we remove the singularity at $w = 0$ (or rather send it to infinity) by letting $u = 1/w$. Then

$$\frac{du}{d\eta} = -u^3v, \quad \frac{dv}{d\eta} = e^\eta u + v. \tag{3.13}$$

Finally we scale the variables to make the system autonomous:

$$U = e^{\eta/3}u, \qquad V = e^{-2\eta/3}v,$$

leading to the system

$$U' = \frac{1}{3}U - U^3V, \quad V' = U + \frac{1}{3}V. \tag{3.14}$$

In terms of the original variables, x, y, we have $U = y^{1/3}/y'$, $V = y^{1/3}y''$. The asymptotic form (1.2) of the solutions approaching $y = 0$ yields

$$U \sim a^{-2/3}x^{1/3} \longrightarrow 0+, \qquad V \sim a^{-2/3}x^{1/3}\log x \longrightarrow 0- \tag{3.15}$$

as $x \longrightarrow 0+$. Note that $V/U \sim \log x \longrightarrow -\infty$ as $x \longrightarrow 0+$. Thus we are interested in solutions of (3.14) approaching the origin as the independent variable $\eta = \log y$ approaches $-\infty$, with $U(\eta) > 0, V(\eta) < 0$.

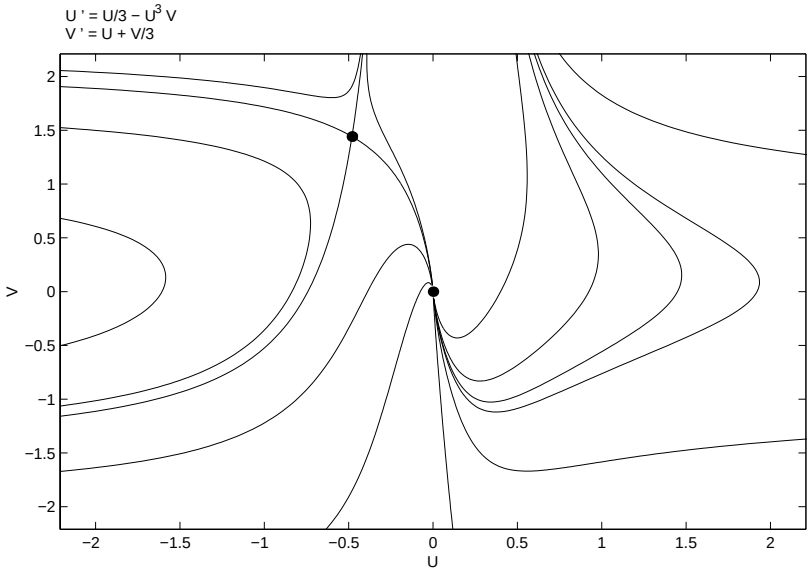


FIG. 3.1. *Phase portrait for system* (3.14).

The origin is an equilibrium with a double eigenvalue $1/3$, but a single eigenvector $(U, V) = (0, 1)$. The phase portrait is shown in Figure 3.1. Note that the saddle point at $V = -3U, U = -3^{-2/3}$ in the second quadrant is not relevant to us. The stable and unstable manifolds correspond to solutions with $y \longrightarrow \infty$ or $y \longrightarrow 0-$. It is straightforward to prove that the trajectories in the fourth quadrant have a unique minimum as they cross $V = -3U$ and cross the U axis as shown in the figure. Indeed, $dV/dU = 3$ on the U axis and is zero on the line $V = -3U$. Moreover,

$$\frac{dV}{dU} = \frac{U + \frac{1}{3}V}{\frac{1}{3}U - U^3V} < 3$$

and is positive for $U > 0,\ V < 0$. Thus, trajectories terminating at any point $U_1 > 0$ on the U axis are monotonically increasing from a point $(U_0, -3U_0)$ with $0 < U_0 < U_1$ and decrease monotonically to the left of the line $V = -3U$. However, the V axis is invariant for the vector field (3.14), so the trajectories are forced into the origin as η decreases. On every ray $V = -AU$ with $A > 3$, we have

$$\frac{dV}{dU} = \frac{U + \frac{1}{3}V}{\frac{1}{3}U - U^3V} > -A.$$

Consequently, trajectories cross every such ray as they approach the origin, proving that $dV/dU \longrightarrow -\infty$ as $\eta \longrightarrow -\infty$. In fact, neglecting the U^3V term as trajectories approach the origin, we find $V \sim 3U \log(U/U_0)$ as $U \longrightarrow 0+$.

This analysis of the fourth quadrant of the vector field proves that there is a one-parameter family of trajectories parameterized by $U_0 > 0$. The trajectories are also invariant under translation of η by a constant, since system (3.14) is autonomous. But $\eta = \log y$, so this corresponds to multiplying y by a constant, accompanied by the corresponding scaling of x, according to the natural scale invariance of (1.1). This is the second parameter that is apparent in the asymptotic series. We have thus proved the following proposition.

PROPOSITION 3.1. *There is a two-parameter family of solutions of the initial value problem* (1.1).

4. Numerical results. In this section, we show results of numerical simulations of the full traveling wave ODE in the spirit of [6]. In particular, we illustrate how solutions corresponding to contact lines relate to the parameters a, b in the asymptotic expansion derived in section 3.1, but we also explore the two simpler cases 1 and 2, in which gravity or the Marangoni forces dominate, respectively.

It is convenient to write the third order equation as a first order system, in which we replace h by u:

$$\begin{aligned} u' &= v, \\ v' &= w, \\ w' &= g(u, v, s), \end{aligned} \tag{4.1}$$

where $g(u, v, s) = \frac{su - f(u)}{C(u)} + Dv$.

The main parameter s is the traveling wave speed. We take $D = 0$, corresponding to fluid flowing down a vertical wall. The functions f, C were given in section 2:
Case I (gravity dominates): $f(u) = u^3 + u^2 K(u)$.
Case II (Marangoni force dominates): $f(u) = u^2 + uK(u)$.
Case III (gravity and Marangoni force are comparable): $f(u) = u^2 - u^3 + K(u)(\frac{2}{3}u - u^2)$.
Recall also the formulae for C and K:

$$C(u) = u^3 + K(u)u^2, \quad K(u) = \beta u^{n-2}. \tag{4.2}$$

In the function K, there are additional parameters $\beta > 0$ and $n < 2$; generally, we take $n = 1$ and $\beta = 0.01$, but we shall also consider the effect of varying β.

Equilibria of system (4.1) are $(u, v, w) = (\overline{u}, 0, 0)$, with $g(\overline{u}, 0, s) = 0$. I.e.,

$$f(\overline{u}) = s\overline{u}. \tag{4.3}$$

We study all the stable and unstable manifolds of equilibria, their boundaries, intersections, and behavior at $u = 0$ in order to gain some understanding of the overall phase portrait.

Computational algorithm. Trajectories for (4.1) are computed using the implicit Adams method in the LSODE package. To compute trajectories along a stable manifold starting near an equilibrium, we integrate backward in time. As for computing trajectories forward in time along an unstable manifold, this process is stable until the manifold comes near another equilibrium.

It is convenient and instructive (see [6]) to use a two-dimensional Poincaré section $\Sigma_{u=const}$ to represent these invariant manifolds. The Poincaré section with u constant has the property that trajectories cross it transversally, unless $v = 0$. In particular, the invariant manifolds intersect Σ_u in points or curves, depending on whether the manifold is one- or two-dimensional. In the Poincaré section, we shall easily visualize when a two-dimensional invariant manifold for one equilibrium is bounded by a one-dimensional invariant manifold for a different equilibrium.

Transverse intersections of two-dimensional manifolds correspond to structurally stable heteroclinic orbits between equilibria. In the Poincaré section, these appear as transverse intersections of curves.

We are interested in solutions $(u, v, w)(\xi)$ that reach $u = 0$ at a finite value of ξ. Specifically, we will say a trajectory $(u(\xi), v(\xi), w(\xi)), \xi < \xi_0$ *touches down* at $\xi = \xi_0$ if $u(\xi) \longrightarrow 0+$ as $\xi \longrightarrow \xi_0-$ and $u(\xi) > 0$ for $\xi < \xi_0$. Similarly, we will say $(u(\xi), v(\xi), w(\xi))$ is *unbounded* if $u(\xi) \longrightarrow \infty$ as $\xi \longrightarrow \pm\infty$.

The numerical results were obtained for specific choices of parameters: $s = 2/9$, $D = 0$, $n = 1$. (The choice $n = 1$ was suggested in [17].) The choice of s is intended to be representative of the physical solutions of interest. In case I, different choices of β are considered, while in cases II and III, $\beta = 0.01$ is taken to be representative.

4.1. Case I: Gravity dominates. When gravity dominates, (4.1) has the cubic polynomial flux function

$$f(u) = u^3 + \beta u. \tag{4.4}$$

Note that (4.1) has the symmetry property that it is unchanged by changing the sign of u and ξ. (Then w changes sign, but v does not.)

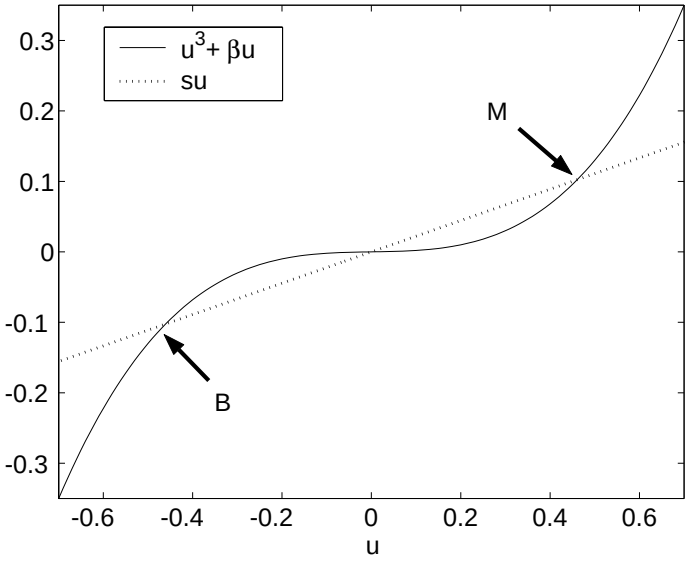


FIG. 4.1. $f(u)$ *and* su.

In Figure 4.1, we show (4.3) for $\beta = 0.01$. The equation has two solutions corresponding to equilibria of (4.1), namely $\overline{u} = \pm\frac{\sqrt{190}}{30}$. (Note that the solution $\overline{u} = 0$ is *not* an equilibrium.) The associated equilibria are

$$B = \left(-\frac{\sqrt{190}}{30}, 0, 0\right), \quad M = \left(\frac{\sqrt{190}}{30}, 0, 0\right).$$

Although B is not physical since it corresponds to negative u, it is nonetheless helpful to consider the associated invariant manifolds.

Linearizing around M, we find the system has two complex conjugate eigenvalues with positive real part (corresponding to a two-dimensional unstable manifold denoted $W^U(M)$), and one real, negative eigenvalue (corresponding to a one-dimensional stable manifold denoted $W^S(M)$). Correspondingly, the equilibrium B

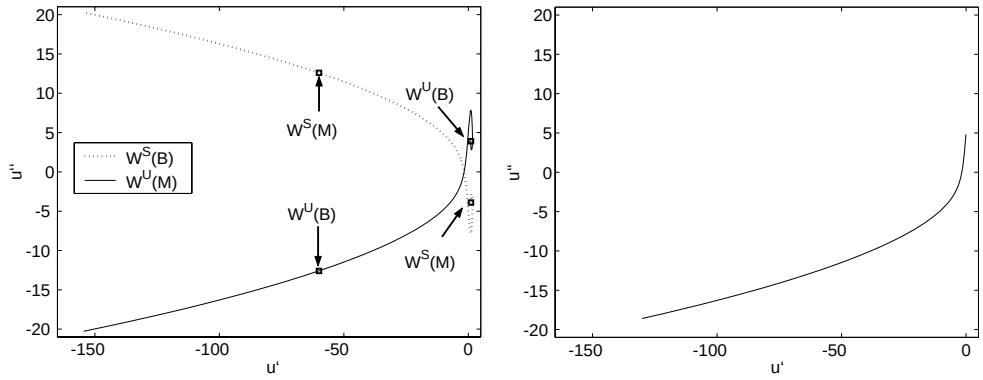


FIG. 4.2. *Case* I *with* $\beta = 0.01$. (a) Σ_0. (b) *Trajectories with a contact line. Shown are only the values at the first intersection with* Σ_0.

has a two-dimensional stable manifold $W^S(B)$ and a one-dimensional unstable manifold $W^U(B)$.

As described in [6, 30], the nature of the two-dimensional manifolds changes from node to focus as the parameter D varies away from zero. In discussing the phase portraits here (with $D = 0$), we consider the focus case only, for which the two associated eigenvalues are complex conjugates. This implies that, near M, solutions along the unstable manifold spiral out. Thus to compute the unstable manifold, we compute trajectories starting from a locus of points along a straight line through M in the tangent plane.

Global picture of phase space. We are interested in how trajectories touch down, so we study the Poincaré section Σ_0 at $u = 0$. By symmetry, the invariant manifolds of B and M are reflections of each other across the plane $u = 0$. The manifolds $W^U(B)$ and $W^S(M)$ are one-dimensional, so they intersect Σ_0 in isolated points, if at all. Each one-dimensional invariant manifold has two connected components, or branches, separated by the equilibrium. Referring to Figure 4.2(a), representing the Poincaré plane $u = 0$, for $\beta = 0.01$, we observe that one branch of $W^U(B)$ intersects Σ_0 twice: first at $(u, u', u'') = (0, 1.06, 3.9)$, then at $(0, -60.0, -12.6)$.[3] The other branch is unbounded at $u = -\infty$ and does not intersect Σ_0. Similarly, one branch of $W^S(M)$ intersects Σ_0 at $(u, u', u'') = (0, 1.06, -3.9)$, while the other branch is unbounded at $u = \infty$ and does not intersect Σ_0.

The two-dimensional manifolds $W^U(M)$ and $W^S(B)$ intersect Σ_0 in curves (see Figure 4.2(a)). $W^U(M)$ is bounded on both ends by the same branch of $W^U(B)$. Near this boundary $W^U(M)$ wraps around $W^U(B)$ an infinite number of times. The spiral of $W^U(M)$ around the first intersection of $W^U(B)$ near $(0, 1.06, 3.9)$ can be seen in Figure 4.2(a). By the time $W^U(B)$ intersects the second time near $(0, -60.0, -12.6)$, the spiral of $W^U(M)$ has become so elongated that it cannot be resolved at this scale.

The points P on Σ_0 of physical interest are those which represent trajectories from M that hit $u = 0$ for the first time at P. These correspond to traveling wave solutions which asymptote to $u = \frac{\sqrt{190}}{30}$ at $\xi = -\infty$ and have a contact line. Although trajectories continue into negative u values, solutions with $u < 0$ are no longer phys-

[3]Here and throughout this section, we give the numerical values of intersection points to one or two decimal places.

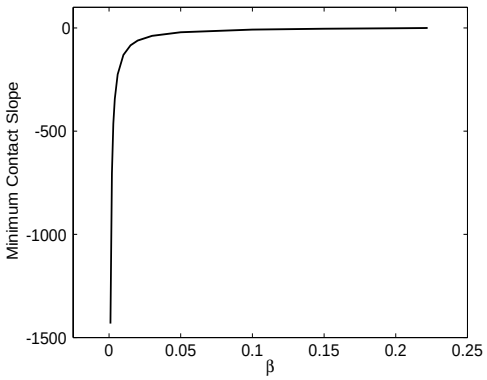


FIG. 4.3. *Minimum slope at $u = 0$ as a function of β, case* I.

ical. All trajectories in $W^U(M)$ hit $u = 0$. The curve of points corresponding to first intersections are shown in Figure 4.2(b). Note that one end of the curve terminates at $u' = 0$. This corresponds to a trajectory that turns around at $u = 0$. This same trajectory eventually winds around to intersect Σ_0 again, this time with a touchdown of slope $u' = -130.1$, corresponding to the other end of the curve shown in Figure 4.2(b). For trajectories inside this orbit, we obtain the finite range of touchdown slopes, spanning $-130.1 \le u' \le 0$. Such large slopes may of course take the model outside its range of validity. However, the slopes here are dimensionless, and may correspond in dimensional variables within the range of the model. This issue is examined in the context of thin film rupture in the paper of Zhang and Lister [42].

Dependence of the range of contact slopes on β. In Figure 4.3, we show how the minimum contact angle varies with the parameter β (from the Navier slip condition (4.2)) in the range $0 < \beta < 2/9$. Keeping $s = 2/9$, $D = 0$, and $n = 1$ fixed, the phase space of solutions is topologically equivalent for $0 < \beta < \frac{2}{9}$. As $\beta \to 0$, M and B approach $(\pm\frac{\sqrt{2}}{3}, 0, 0)$. As β increases from 0, M and B move closer together until all solutions of (4.3) vanish at $\beta = \frac{2}{9}$, where $f'(0) = s$. For all β in this range, the maximum contact slope is zero. The minimum contact slope decreases with decreasing β, as seen above. The plot shows computed values for the minimum value of u' at $u = 0$ for β ranging from 0.001 to 0.22.

4.2. Case II: Marangoni convection dominates. Now consider (4.1) when the Marangoni convection term dominates. In this subsection we consider only $\beta = 0.01$; in this case and in case III we compute solutions only in the physical range $u \ge 0$. The new flux function is quadratic:

$$f(u) = u^2 + \beta. \tag{4.5}$$

For different values of s, the ODE will have zero, one, or two equilibria. The case with $s = 2/9$ is shown in Figure 4.4. The two equilibria are

$$B = \left(\frac{10 - \sqrt{19}}{90}, 0, 0\right), \quad M = \left(\frac{10 + \sqrt{19}}{90}, 0, 0\right).$$

M has a two-dimensional unstable manifold $W^U(M)$ and a one-dimensional stable manifold $W^S(M)$. B has a two-dimensional stable manifold $W^S(B)$ and a one-dimensional unstable manifold $W^U(B)$.

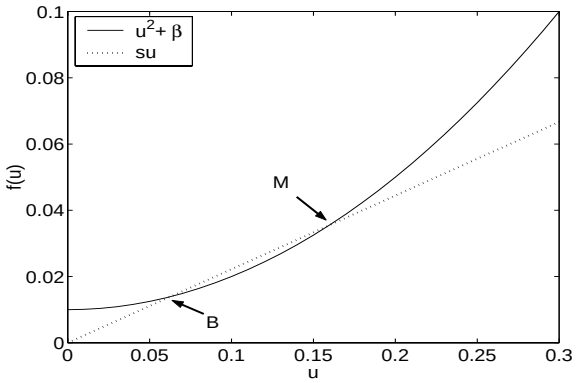


FIG. 4.4. *$f(u)$ and su, case* II*, with $\beta = 0.01$.*

The dynamics at the contact line are now fundamentally different than in the gravity dominated case. The flux function makes the differential equation (4.1) singular at $u = 0$. Note that if (u, v, w) touches down, then $w = u''$ will become unbounded in finite time, so an adaptive time step procedure such as the one we use is essential to capture detailed behavior near $u = 0$. The most important difference from the gravity dominated case is that now trajectories cannot be computed beyond $u = 0$. Thus, trajectories generically fall into two cases: those which touch down in finite time and those which escape to $u = \infty$. Between these two cases are heteroclinic orbits that approach equilibria as $\xi \longrightarrow \pm\infty$.

Global picture of phase space. Since u'' is unbounded at $u = 0$, it is not possible to study Σ_0. Instead, we choose the section $\Sigma_{0.1}$ between B and M. The information about boundaries and dimensions can be read off as before. It is necessary, however, to do further computations to find the range of contact slopes. Intersections of the two-dimensional manifolds again correspond to heteroclinic orbits between equilibria, representing traveling wave solutions with a precursor layer, as in [6].

Both branches of $W^S(M)$ are unbounded. One does not pass $\Sigma_{0.1}$. The other branch intersects $\Sigma_{0.1}$ at $(u, u', u'') = (0.1, 0.16, -0.43)$, narrowly avoids $u = 0$, and intersects again at $(u, u', u'') = (0.1, -2.00, 20.71)$ before heading to $u = \infty$. In contrast, both branches of $W^U(B)$ touch down. One branch does not intersect $\Sigma_{0.1}$. The other intersects first at $(0.1, 0.15, 0.52)$, then at $(0.1, -1.72, -2.02)$ before touching down. These intersection points are labeled in Figure 4.5.

The two-dimensional manifolds $W^U(M)$ and $W^S(B)$ intersect transversally in a single heteroclinic orbit from M to B. Hence the two branches of $W^U(B)$ are boundaries of $W^U(M)$. The curve $W^U(M)$ shown in the Poincaré section of Figure 4.5 represents trajectories that can intersect $\Sigma_0.1$ several times. All trajectories in $W^U(M)$ touch down, with a finite range of contact angles. For the specific parameters $\beta = 0.01$, $s = 2/9$, we find this range to be $-2.436 < u' < -0.464$. Note that, unlike in the gravity driven case, there are no trajectories with contact slope arbitrarily close to 0.

4.3. Case III: The full equation. We now consider the full equation with $\beta = 0.01$ and

$$f(u) = u^2 - u^3 + \beta\left(\frac{2}{3} - u\right). \tag{4.6}$$

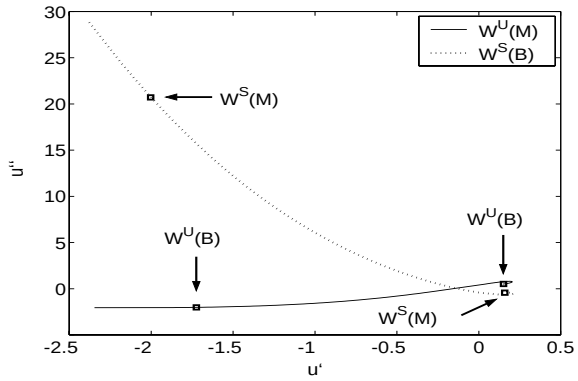


FIG. 4.5. $\Sigma_{0.1}$, *case* II.

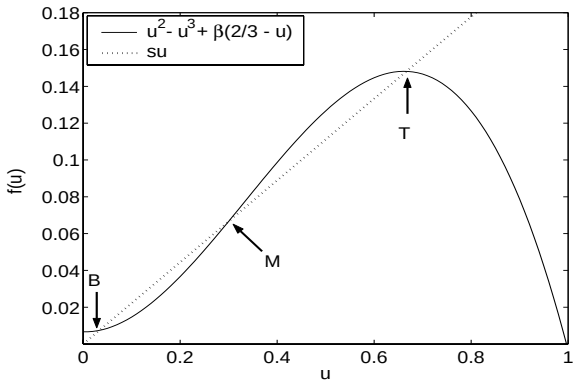


FIG. 4.6. $f(u)$ *and* su, *case* III, $\beta = 0.01$.

See [6] for a parallel discussion of the precursor model case without Navier slip. The flux is now nonconvex, which means the ODE may have one, two, or three equilibria, depending on the value of s. For $s = 2/9$, (4.3) has three solutions. We will find that phase space becomes correspondingly more complicated. The three equilibria (see Figure 4.6) are labeled

$$B = (1/30, 0, 0), \quad M = (3/10, 0, 0), \quad T = (2/3, 0, 0),$$

for bottom, middle, and top. M has a two-dimensional unstable manifold $W^U(M)$ and a one-dimensional stable manifold $W^S(M)$. The other equilibria B and T have two-dimensional stable manifolds $W^S(B), W^S(T)$ and one-dimensional unstable manifolds $W^U(B), W^U(T)$.

As in the case where Marangoni convection dominates gravity, trajectories in the invariant manifolds either will be heteroclinic orbits, unbounded, or will touch down. There is a Lyapunov function that prevents periodic or homoclinic orbits (cf. [6]). Solutions cannot be computed past $u = 0$ since the ODE becomes singular. In what follows, we describe results with respect to the Poincaré section $\Sigma_{0.2}$ between B and M at $u = 0.2$.

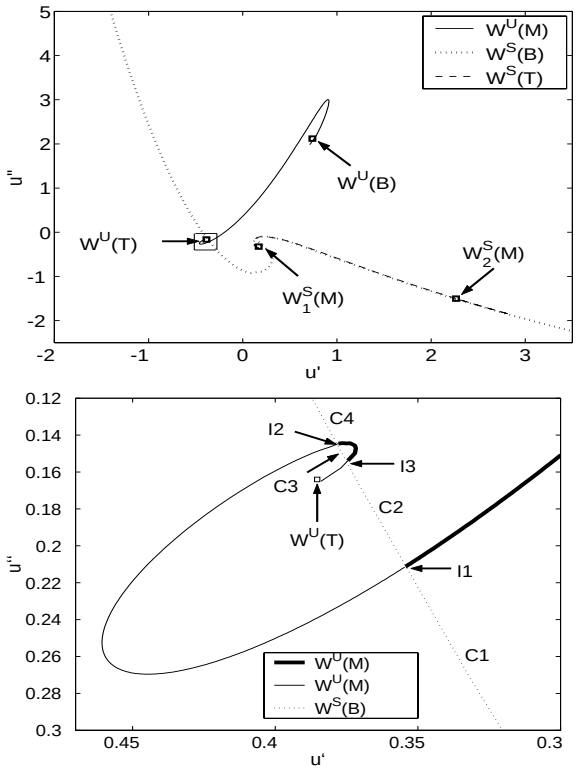


FIG. 4.7. *Case* III. (a) $\Sigma_{0.2}$. (b) *Blow-up of small box in* (a).

The one-dimensional manifolds. $W^U(B)$, $W^U(T)$, and $W^S(M)$ are the one-dimensional invariant manifolds. One branch of $W^U(B)$ touches down and does not pass through $u = 0.2$. The other branch becomes unbounded in u and passes $\Sigma_{0.2}$ at $(0.2, 0.739, 2.120)$. $W^U(T)$ exhibits similar behavior, with an unbounded branch that does not pass through $\Sigma_{0.2}$ and a branch that touches down and hits $\Sigma_{0.2}$ at $(0.2, -0.385, -0.164)$. Both branches of $W^S(M)$ touch down. One branch, $W_1^S(M)$, passes through $\Sigma_{0.2}$ at $(0.2, 0.169, -0.324)$. The other, $W_2^S(M)$, passes through at $(0.2, 2.264, -1.503)$.

The unstable manifold of M. The manifold $W^U(M)$ crosses $W^S(B)$ in three orbits, shown as intersection points in Figure 4.7(b). Note also that $W^U(M)$ wraps infinitely many times around both $W^U(B)$ and $W^U(T)$. These spirals indicate connections from M to both T and B. Connections to B are already evident. Connections to T would appear as an intersection of $W^U(M)$ and $W^S(T)$ in any Poincaré section placed between M and T.

The structure of the intersection of $W^U(M)$ with $W^S(B)$ is shown in more detail in Figure 4.7(b). $W^U(M)$ intersects $W^S(B)$ three times. Each of these intersections I1, I2, and I3 corresponds to a heteroclinic orbit from M to B. Furthermore, $W^S(B)$ divides the qualitative behavior of trajectories on $W^U(M)$. All trajectories on the same side as $W^U(T)$, i.e., between I1 and I2 and between I3 and $W^U(T)$, touch down. All trajectories on the other side, i.e., between $W^U(B)$ and I1 and between I2 and I3, are unbounded.

The stable manifold of T. $W^S(T)$ also has two boundaries, which are the two branches of $W^S(M)$. The spiral at $W_1^S(M)$ can be seen clearly, while the spiral at $W_2^S(M)$ is four orders of magnitude longer than it is wide, and requires corresponding resolution to be seen.

The stable manifold of B. Each trajectory on the stable manifold of B (which hits $\Sigma_{0.2}$) comes out from $B = (1/30, 0, 0)$, passes through $u = 0.2$, turns around, and hits $u = 0.2$ again before touching down. Thus each trajectory appears as two points on $\Sigma_{0.2}$. $W^S(B)$ is split by $W^U(M)$ into four distinctly behaving sheets which we label C1, C2, C3, and C4.

The sheet C1 approaches the heteroclinic orbit I1 at one end and wraps around $W_1^S(M)$ at the other. The sheet wraps back on itself around $(0.2, 0, -1)$. Each trajectory has one point between I1 and $(0.2, 0, -1)$ (where u is increasing) and one point between $(0.2, 0, -1)$ and $W_1^S(M)$ (where u is decreasing). Trajectories which pass very close to I1 going out also pass very close to $W^S(M)$ coming back. In fact, C1 spirals around $W_1^S(M)$ infinitely many times.

Outgoing trajectories in C2 are bounded by I1 and I3. The trajectories which pass very close to I1 spiral around $W_2^S(M)$ on their second pass, and the trajectories which pass very close to I3 spiral around $W_1^S(M)$ when they return. Trajectories between these two boundaries are very close to $W^S(T)$ on their return.

Outgoing trajectories on C3 are bounded by I3 and I2. Trajectories that pass near both of these orbits on the way out spiral tightly around $W_2^S(M)$ on their way back. Trajectories between the two boundaries form a thin loop that follows $W^S(T)$ very closely, and stretches down to $(0.2, 2.84, -1.86)$.

Finally, outgoing trajectories on C4 are bounded on one side by I2. The trajectories which pass near to I2 spiral tightly around $W_1^S(M)$ on their return journey. Trajectories which pass further away from C1 return in a line that follows $W^S(T)$ and eventually stretches beyond it.

Note that both $W_1^S(M)$ and $W_2^S(M)$ have quadruple spirals, that is, four sheets wrapping around them infinitely many times. $W_1^S(M)$ has one boundary each of $W^S(T)$, C1, C2, and C4, whereas $W_2^S(M)$ has one boundary each from $W^S(T)$ and C2, and two from C3.

4.4. Connection to the asymptotics. In the computations for case III, for a given wave speed s, we find a one-parameter family of trajectories that touch down. Here we relate this one-parameter family to the two-parameter family of local touch down solutions given by the asymptotics of section 2. Specifically for $s = 2/9$, we find a finite range of values for the parameter a in the asymptotic solution.

With the scaling $\xi - \xi_0 = -(\frac{2}{3})^{1/3}x$ of section 2, the asymptotic expansion about a point $\xi = \xi_0$ of touchdown becomes

$$h(\xi) = -a\left(\frac{2}{3}\right)^{\frac{1}{3}}\eta + \frac{1}{2a}\left(\frac{2}{3}\right)^{\frac{2}{3}}\eta^2 \log\left(-\left(\frac{2}{3}\right)^{\frac{1}{3}}\eta\right) + b\left(\frac{2}{3}\right)^{\frac{2}{3}}\eta^2 + \text{h.o.t.}, \tag{4.7}$$

in which $\eta = \xi - \xi_0 < 0$. In particular, the contact slope is given by $v(\xi_0) = -a\left(\frac{2}{3}\right)^{1/3}$. We can read off the limiting value of $v(\xi)$ as u approaches zero and thus obtain the value of a. Corresponding values for b also come from the form (1.2). Specifically, we find

$$w = u'' = \frac{1}{a}\left(\frac{2}{3}\right)^{\frac{2}{3}} \log\left(-\left(\frac{2}{3}\right)^{\frac{1}{3}}\eta\right) + \frac{3}{2a}\left(\frac{2}{3}\right)^{\frac{2}{3}} + 2\left(\frac{2}{3}\right)^{\frac{2}{3}} b + \text{h.o.t.} \tag{4.8}$$

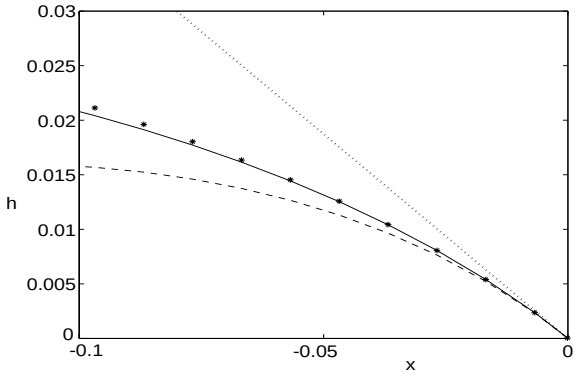


FIG. 4.8. *.....first order, - - -second order, *****third order, a=0.429, b=0.762.*

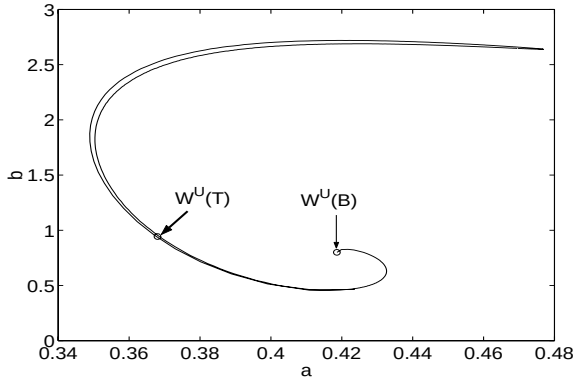


FIG. 4.9. *Asymptotic parameters a and b.*

Thus,

$$e^{\left(a\left(\frac{3}{2}\right)^{2/3}w(\xi)\right)} \approx -\left(\frac{2}{3}\right)^{1/3} e^{\left(\frac{3}{2}+2ab\right)}(\xi-\xi_0). \tag{4.9}$$

Consequently, $e^{\left(a\left(\frac{3}{2}\right)^{2/3}w(\xi)\right)}$ approaches a constant slope m as u approaches zero. Computing this value of m from the calculated trajectory, and the parameter a as above, the expression

$$m = -\left(\frac{2}{3}\right)^{1/3} e^{\left(\frac{3}{2}+2ab\right)} \tag{4.10}$$

determines b. In Figure 4.8, we show a sample comparison between the first three terms of the asymptotic expansion with the corresponding numerical solution of the full equation, with parameters a, b calculated as described above.

Range of *a* and *b*. Plotting the two asymptotic parameters a and b against each other, for all trajectories on $W^U(M)$ which touch down, gives a curve in the (a, b) plane. Consider the close up Poincaré section of the intersection between $W^U(M)$ and $W^S(B)$. In Figure 4.9, we show this curve for points on $W^U(M)$ between I1 and

I2 corresponding to trajectories that touch down in finite time. Points adjacent to I_1 and I_2 correspond to identical values of a and b, in fact, the same values as for the branch of $W^U(B)$ that touches down. The curve in Figure 4.9 varies smoothly but very nearly doubles back on itself. Trajectories near the heteroclinic orbits approach B and then shoot off along $W^U(B)$.

The points along $W^U(M)$ from I3 to $W^U(T)$ give another one-parameter family of solutions. Their a, b values range continuously from the a, b value of the branch of $W^U(B)$ that touches down (since I2 is another heteroclinic orbit) to the a, b value of the branch of $W^U(T)$. This second curve closely follows the first one.

5. Discussion. In this section, we discuss the significance of the results for the traveling wave problem and the moving contact line. The most striking conclusion from our numerical results is that for each of the three problems considered, there is a limited range of contact angles. This particular observation has not been noted in previous studies of the traveling wave problem with slip [31, 38] for the case of gravity driven films. For the Marangoni cases II and III the effect is even more pronounced; the range does not include a zero contact angle, in contrast to well-known results for (2.12) without convection (i.e., $f = 0$) for both traveling waves [9] and weak solutions of the full PDE [8, 3]. For example, in case I, Figure 4.2(b), for each contact slope u' in the range $(-130.1, 0]$, there is a unique traveling wave solution that touches down with that slope. In contrast, in case III, not only are there no traveling waves that touch down with zero slope, but for each slope in the range there can be more than one traveling wave with that touchdown slope, as demonstrated in Figure 4.9.

In experiments, the contact line speed is often observed to be related to the dynamic contact angle or the slope of the film as the thickness approaches zero. For the model, this would result in a boundary condition relating the contact angle to the speed of the wave. Since the upstream thickness is related to the speed s via (2.17), for traveling waves, such a boundary condition becomes a relationship between upstream thickness and contact angle. In case I, the boundary condition would select a unique traveling wave, provided the contact slope is in the admissible range. However, in case III, such a law does not in general select a unique traveling wave. A similar nonuniqueness was found for the precursor model [6] in the case of gravity and opposing Marangoni stress.

There is also the issue of whether the contact angle should be related to the slope of the free surface at zero height, or to an observed slope, that might be taken to be the maximum slope, generally slightly away from the contact line itself. There is much discussion of this issue in the literature (see [25, 26] and the references therein). The Navier slip condition is an attempt to incorporate in a continuum model the physical effects at the molecular scale. For this reason, the asymptotic results apply strictly at the contact line. As in [25], it would be possible to carry out matched asymptotics to relate the asymptotic solution at zero height to solutions of the full problem away from a small neighborhood of the contact line. In this paper, we have instead compared numerical solutions with the asymptotic solution; the discussion above is based on this comparison.

Case III is particularly interesting because of the nonconvex flux and the connection to compressive and undercompressive waves discussed in [6]. The notion of compressive and undercompressive carries over to this paper in the following way: we call trajectories from M that touch down compressive traveling waves. If the trajectory from T touches down, we refer to it as an undercompressive wave. Unlike the precursor model in [6], we do not have characteristics ahead of the wave, hence these

names are used to distinguish between cases where characteristics from the bulk film go into the contact line (compressive) or come out of the contact line (undercompressive). In the former case, information from the bulk can influence the contact line. In the latter case, information from the contact line is carried into the bulk.

Note that there is a range of speeds for which we have three equilibria B, M, and T. We conjecture that for speeds in an interval within this range, one branch of the unstable manifold from T touches down to $u = 0$. Each such trajectory will have a touchdown angle. It would be interesting to know if a given boundary condition at the contact line selects a unique undercompressive wave.

It would be interesting to consider the traveling waves that touch down in the context of the dynamic free boundary problem for the PDE (2.12). We expect that the structure of case III is as rich as that for the precursor model [6, 4]. The contact line raises a new issue: that of how to treat the free boundary for the full PDE. The full nonlinear dynamics have not been explored, with the exception of [35], in which a special case of the PDE without convection was studied.

Acknowledgment. We gratefully acknowledge helpful discussions with David Schaeffer.

REFERENCES

[1] W. D. Bascom, R. L. Cottington, and C. R. Singleterry, *Dynamic surface phenomena in the spontaneous spreading of oils on solids*, in Contact Angle, Wettability and Adhesion, F. M. Fowkes, ed., American Chemical Society, Washington, DC, 1964, pp. 355–379.

[2] E. Beretta, J. Hulshof, and L. A. Peletier, *On an ODE from forced coating flow*, J. Differential Equations, 130 (1996), pp. 247–265.

[3] E. Beretta, M. Berstch, and R. Dal Passo, *Nonnegative solutions of a fourth order nonlinear degenerate parabolic equation*, Arch. Ration. Mech. Anal., 129 (1995), pp. 175–200.

[4] A. Bertozzi, A. Münch, M. Shearer, and K. Zumbrun, *Stability of compressive and undercompressive thin film travelling waves*, European J. Appl. Math., 12 (2001), pp. 253–291.

[5] A. L. Bertozzi, A. Münch, X. Fanton, and A. M. Cazabat, *Contact line stability and 'undercompressive shocks' in driven thin film flow*, Phys. Rev. Lett., 81 (1998), pp. 5169–5172.

[6] A. L. Bertozzi, A. Münch, and M. Shearer, *Undercompressive shocks in thin film flows*, Phys. D, 134 (1999), pp. 431–464.

[7] A. L. Bertozzi and M. Shearer, *Existence of undercompressive traveling waves in thin film equations*, SIAM J. Math. Anal., 32 (2000), pp. 194–213.

[8] A.L. Bertozzi and M. Pugh, *The lubrication approximation for thin viscous films: Regularity and long time behavior of weak solutions*, Comm. Pure Appl. Math., 49 (1996), pp. 85–123.

[9] S. Boatto, L. Kadanoff, and P. Olla, *Travelling wave solutions to thin film equations*, Phys. Rev. E, 48 (1993), p. 4423.

[10] J. B. Brzoska, F. Brochard-Wyart, and F. Rondelez, *Exponential growth of fingering instabilities of spreading films under horizontal thermal gradients*, Europhys. Lett., 19 (1992), pp. 98–102.

[11] A. M. Cazabat, F. Heslot, S. M. Troian, and P. Carles, *Finger instability of this spreading films driven by temperature gradients*, Nature, 346 (1990), pp. 824–826.

[12] P. G. de Gennes, *Wetting: Statics and dynamics*, Rev. Mod. Phys., 57 (1985), p. 827.

[13] E. B. Dussan V and S. Davis, *On the motion of a fluid–fluid interface along a solid surface*, J. Fluid Mech., 65 (1974), pp. 71–95.

[14] P. Ehrhard and S. H. Davis, *Non-isothermal spreading of liquid drops on horizontal plates*, J. Fluid. Mech., 229 (1991), pp. 365–388.

[15] S. Goldstein, ed., *Modern Developments in Fluid Dynamics*, Vol. 2, Dover, New York, 1965.

[16] A. A. Golovin, B. Y. Rubinstein, and L. M. Pismen, *Effect of van der waals interaction on the fingering instability of thermally driven thin wetting films*, Langmuir, 17 (2001), pp. 3930–3936.

[17] H. P. Greenspan, *On the motion of a small viscous droplet that wets a surface*, J. Fluid Mech., 84 (1978), pp. 125–143.

[18] P. J. Haley and M. J. Miksis, *The effect of the contact line on droplet spreading*, J. Fluid Mech., 223 (1991), pp. 57–81.
[19] L. Hocking, *Spreading and instability of a viscous fluid sheet*, J. Fluid Mech., 221 (1990), pp. 373–392.
[20] L. M. Hocking, *A moving fluid interface on a rough surface*, J. Fluid Mech., 76 (1976), pp. 801–817.
[21] L. M. Hocking, *A moving fluid interface. Part* 2. *The removal of the force singularity by a slip flow*, J. Fluid Mech., 79 (1977), pp. 209–229.
[22] L. M. Hocking, *The spreading of a thin drop by gravity and capillarity*, Quant. J. Mech. Appl. Math., 36 (1983), pp. 55-69.
[23] L. M. Hocking, *Rival contact-angle models and the spreading of drops*, J. Fluid. Mech., 239 (1992), pp. 671–681.
[24] H. Huppert, *Flow and instability of a viscous current down a slope*, Nature, 300 (1982), pp. 427–429.
[25] S. Kalliadasis and H.-C. Chang, *Apparent dynamic contact angle of an advancing gas-liquid meniscus*, Phys. Fluids, 6 (1994), pp. 12–23.
[26] S. Kalliadasis and H.-C. Chang, *Dynamics of liquid spreading on solid surfaces*, Ind. Eng. Chem. Res., 35 (1996), pp. 2860–2874.
[27] D. E. Kataoka and S. M. Troian, *A theoretical study of instabilities at the advancing front of thermally driven coating films*, J. Coll. Int. Sci., 192 (1997), pp. 350–362.
[28] D. E. Kataoka and S. M. Troian, *Stabilizing the advancing front of thermally driven climbing films*, J. Coll. Int. Sci., 203 (1998), pp. 335–344.
[29] V. Ludviksson and E. N. Lightfoot, *The dynamics of thin liquid films in the presence of surface-tension gradients*, Am. Inst. Chem. Engrs. J., 17 (1971), pp. 1166–1173.
[30] A. Münch, *Shock transitions in Marangoni-gravity driven thin film flow*, Nonlinearity, 13 (2000), pp. 731–746.
[31] A. Münch and B. Wagner, *Numerical and asymptotic results on the linear stability of a thin film spreading down a slope of small inclination*, European J. Appl. Math., 10 (1999), pp. 297–318.
[32] C. Navier, *Memoire sur les lois du mouvement des fluids*, Memoires de l'Academie Royale des Sciences de l'Institut de France, 6 (1823), pp. 389–440.
[33] P. Neogi and C. A. Miller, *Spreading kinetics of a drop on a rough solid surface*, J. Coll. Int. Sci., 92 (1983), pp. 338–349.
[34] A. Oron, S. H. Davis, and S. G. Bankoff, *Long-scale evolution of thin liquid films*, Rev. Mod. Phys., 69 (1997), pp. 931–980.
[35] F. Otto, *Lubrication approximation with prescribed nonzero contact angle*, Comm. Partial Differential Equations, 23 (1998), pp. 2077–2164.
[36] M. Schneemilch and A. M. Cazabat, *Shock separation in wetting films driven by thermal gradients*, Langmuir, 16 (2000), pp. 9850–9856.
[37] M. Schneemilch and A. M. Cazabat, *Wetting films in thermal gradients*, Langmuir, 16 (2000), pp. 8796–8801.
[38] M. A. Spaid and G. M. Homsy, *Stability of Newtonian and viscoelastic dynamic contact angles*, Phys. Fluids, 8 (1996), pp. 460–478.
[39] E. O. Tuck and L. W. Schwartz, *A numerical and asymptotic study of some third-order ordinary differential equations relevant to draining and coating flows*, SIAM Rev., 32 (1990), pp. 453–469.
[40] O. V. Voinov, *Inclination angles of the boundary in moving liquid layers*, Zh. Prikl. Mekh. Tekh. Fiz., 2 (1977), pp. 92–99.
[41] M. B. Williams and S. H. Davis, *Nonlinear theory of film rupture*, J. Coll. Int. Sci., 90 (1982), pp. 220–228.
[42] W. W. Zhang and J. R. Lister, *Similarity solutions for van der Waals rupture of a thin film on a solid substrate*, Phys. Fluids, 11 (1999), pp. 2454–2462.

SIAM J. APPL. MATH.
Vol. 63, No. 3, pp. 745–776

AVERAGING OF DISPERSION-MANAGED SOLITONS: EXISTENCE AND STABILITY*

DMITRY E. PELINOVSKY† AND VADIM ZHARNITSKY‡

Abstract. We consider existence and stability of dispersion-managed solitons in the two approximations of the periodic nonlinear Schrödinger (NLS) equation: (i) a dynamical system for a Gaussian pulse and (ii) an average integral NLS equation. We apply normal form transformations for finite-dimensional and infinite-dimensional Hamiltonian systems with periodic coefficients. First-order corrections to the leading-order averaged Hamiltonian are derived explicitly for both approximations. Bifurcations of soliton solutions and their stability are studied by analysis of critical points of the first-order averaged Hamiltonians. The validity of the averaging procedure is verified and the presence of ground states corresponding to dispersion-managed solitons in the averaged Hamiltonian is established.

Key words. existence and stability of pulses, optical solitons, dispersion management, averaging theory, normal form transformations, errors and convergence of asymptotic series, periodic NLS equation, integral NLS equation, Gaussian approximation

AMS subject classifications. 35Q55, 78M30, 78M35

PII. S0036139902400477

1. Introduction.

1.1. Motivations. Ultrafast high–bit-rate optical communication networks are enhanced by the dispersion management technology when two optical fibers of opposite dispersion are periodically concatenated into a line [1]. If the communication network has low path-averaged dispersion and high local dispersion, the data signals are optimally transmitted from the input to output ends through a periodic sequence of compression and expansion cycles. The long-haul dispersion management is technologically combined with standard loss management when a periodic chain of amplifiers compensates distributive fiber losses.

Many recent experimental groups reported revolutionary performance of dispersion-managed (DM) pulses in optical communication networks [2, 3]. Two regimes were studied in detail: DM solitons and chirped return-to-zero pulses. DM solitons are time bits transmitted stationary on the average through the long-haul communication network [2]. The chirped return-to-zero pulses are weakly broadened on the average due to transmission, and some post-transmission compression may be required at the output of the network [3].

This paper addresses the stationary DM solitons and resolves yet open problems of existence and stability of stationary DM solitons described by a periodic nonlinear Schrödinger (NLS) equation. Theoretical studies of DM solitons are based on one of the three averaging methods: (i) variational Gaussian approximation, (ii) asymptotic reduction to an integral NLS equation, and (iii) numerical split-step averaging algorithm (see the latest reviews [4, 5, 6]).

*Received by the editors January 3, 2002; accepted for publication (in revised form) July 27, 2002; published electronically January 17, 2003.
http://www.siam.org/journals/siap/63-3/40047.html
†Department of Mathematics, McMaster University, Hamilton, Ontario, Canada, L8S 4K1 (dmpeli@math.mcmaster.ca). This author's research was supported by NSERC grant 5-36694.
‡Mathematical Sciences Research, Bell Laboratories, Lucent Technologies, Murray Hill, NJ 07974 (vz@research.bell-labs.com).

The variational Gaussian approximation truncates the periodic NLS equation at a finite-dimensional Hamiltonian system with periodic coefficients. The truncation is performed by integrating the Lagrangian density of the NLS equation over the Gaussian pulse and varying the resulting function with respect to parameters of the Gaussian pulse [7, 8]. Periodic orbits of the nonautonomous Hamiltonian system correspond to stationary DM solitons [9, 10]. In an optimal design of the dispersion map, the evolution length for dispersion variations is much shorter than the lengths of average dispersion and fiber nonlinearity. Within this limit, the nonautonomous Hamiltonian system can be averaged over the map period. The averaging procedure results, at the leading-order approximation, in a planar dynamical system [11, 12]. Existence and stability of DM solitons can be studied by analyzing trajectories on a phase plane near the critical points of the Hamiltonian system [13]. One of the drawbacks of the Gaussian approximation is the lack of information about the error of the averaging procedure.

The asymptotic reduction to an integral NLS equation is also based on averaging of the periodic NLS equation over a short period of the dispersion map [14, 15]. The method is, however, much more general, since the kernel for the averaging transformation is the most general Fourier solution of the linear periodic Schrödinger equation, which includes the Gaussian pulse as a particular case. Stationary DM solitons are approximated by the time pulse solutions of the nonlinear eigenvalue problem associated with the integral NLS equation. The DM soliton solutions have constantly rotating complex phase along the fiber [15, 16]. Only when the integral NLS equation is approximated at the Gaussian pulse [17], the resulting dynamical system reproduces the same planar Hamiltonian system as in [13]. In the asymptotic reduction method, the integral NLS equation can be viewed as the leading-order term in a set of canonical transformations applied to the periodic NLS equation [18].

At last, the numerical split-step averaging algorithm is applied to separate the pulse resolution in time and the almost periodic evolution of the pulse along the fiber by averaging the output of the split-step method over many time periods [5, 6]. A single pulse with a preserved value of energy was found to converge to a stationary DM soliton unless various resonances and temporal instabilities resulted in an unpredictable loss of convergence of the numerical algorithm [6, 19].

We are motivated by a number of averaging methods applied to the periodic NLS equation and by contradictory results on existence and stability of DM solitons found within these methods. In order to justify and clarify these methods, we develop a systematic asymptotic procedure for averaging of the periodic NLS equation, based on normal form transformations. We extend the perturbation expansions to the next order, where the first-order corrections to the leading-order equations are derived. The validity of averaging methods and the errors (accuracy) of the leading-order and first-order approximations are proved rigorously for a two-step dispersion map. Branches of stationary DM solitons and their stability are analyzed within the averaged equations.

This paper is structured as follows. In section 1.2 we describe the physical model, parameters, and normalizations. In section 1.3 we discuss two approximations of DM solitons and summarize the previously known results, together with our main propositions. In sections 2.1–2.4 we study the Gaussian approximation, in combination with the leading and first orders of the averaging method. We find explicitly analytical curves for existence and stability of the DM solitons in this lower-dimensional approximation. In sections 3.1–3.3 we analyze the full PDE problem and prove convergence of the leading and first orders of the averaging method. The existence of

ground states is proved in the averaged equation but the analytical curves are implicit in this higher-dimensional approximation. Section 4 describes open problems of analysis beyond the first-order averaging theory.

1.2. Model and parameterizations. The NLS equation for optical pulses in dispersion-compensated fibers is

$$i\frac{\partial U}{\partial Z} - \frac{1}{2}\beta_2(Z)\frac{\partial^2 U}{\partial T^2} + \gamma_2(Z)|U|^2 U = 0, \tag{1.1}$$

where $U(Z,T)$ is the electric field envelope of the carrier wave at the operating wavelength λ_0, while $\beta_2(Z)$ and $\gamma_2(Z)$ are the fiber dispersion and nonlinearity [1]:

$$\beta_2 = -\frac{\lambda_0^2}{2\pi c}D(Z), \quad \gamma_2 = \gamma \exp\left[\int_0^Z g(Z')dZ' - \alpha Z\right], \quad \gamma = \frac{2\pi n_2}{\lambda_0 A_{\rm eff}}\ (>0). \tag{1.2}$$

All units in (1.1)–(1.2) have dimensional form, such that $D(Z)$ is the dispersion coefficient measured in ps/(nm $\times$ km), c is the speed of light in km/sec, $|u|^2$ is the light intensity in mW, n_2 is the nonlinear refractive index in $(\mu\text{m})^2$/mW, $A_{\rm eff}$ is the effective fiber area in $(\mu\text{m})^2$, α is the distributive loss coefficient in km^{-1}, and $g(Z)$ is the periodic amplification. For example, if $A_{\rm eff} = 50(\mu\text{m})^2$, $c = 3\ 10^5$km/sec, $\lambda_0 = 1.5\mu$m, $n_2 = 2.5\ 10^{-11}(\mu\text{m})^2$/mW, and $D = 0.12$ps/(nm $\times$ km), then $\beta_2 \approx -0.1\text{ps}^2$/km and $\gamma \approx 2\ 10^{-3}(\text{mW} \times \text{km})^{-1}$, which are reasonable values for these coefficients.

The dispersion map $D(Z)$ consists of two piecewise-constant fibers of lengths L_1 and L_2 in km, such that $L_1 + L_2 = L_{\rm DM}$, which have dispersion values D_1 and D_2. The total number of fiber segments is $N_{\rm DM}$. The amplification map $g(Z)$ is periodic with period $L_{\rm AM}$, where the ratio $L_{\rm DM}/L_{\rm AM}$ is integer. A typical loss compensation due to erbium-doped fiber amplifiers is

$$g = \alpha L_{\rm AM} \sum_{n=1}^{N_{\rm AM}} \delta(Z - nL_{AM}), \tag{1.3}$$

where $N_{\rm AM}$ is the number of amplifiers over the transmission line: $Z \in [0, N_{\rm DM}L_{\rm DM}]$ and the amplifiers compensate the losses exactly. As a result, the fiber nonlinearity $\gamma_2(Z)$ is a periodic function with period $L_{\rm AM}$. We will use throughout the paper the lossless approximation when $\gamma_2(Z) = \gamma$ is constant. The lossless approximation occurs in the limit $L_{\rm AM} \ll L_{\rm DM}$, when $\lim_{L_{\rm AM}\to 0}\int_0^Z g(Z')dZ' = \alpha Z$. This approximation is sufficiently accurate for modeling fibers with distributed (e.g., Raman) amplification or fibers with several amplifiers at the dispersion compensation period L_{DM} [1]. In other cases, our results are still expected to hold qualitatively.

We can rescale variables (Z,T,U) by introducing characteristic pulse power P_0 in mW, characteristic pulse width T_0 in ps, and characteristic (nonlinear) length $L_{\rm NL} = (\gamma P_0)^{-1}$ in km:

$$Z = L_{\rm NL}z, \quad T = T_0 t, \quad U = \sqrt{P_0}u. \tag{1.4}$$

The periodic NLS equation (1.1) in new variables reduces to the dimensionless form [1],

$$i\frac{\partial u}{\partial z} + \frac{m}{2\epsilon}d\left(\frac{z}{\epsilon}\right)\frac{\partial^2 u}{\partial t^2} + \frac{1}{2}d_0\frac{\partial^2 u}{\partial t^2} + |u|^2 u = 0, \tag{1.5}$$

with the dimensionless parameters

$$m = \frac{\lambda_0^2 L_1 L_2 (D_1 - D_2)}{4\pi c L_{\rm DM} T_0^2}, \quad d_0 = \frac{\lambda_0^2 (D_1 L_1 + D_2 L_2)}{2\pi c \epsilon T_0^2}, \quad \epsilon = \gamma L_{\rm DM} P_0. \tag{1.6}$$

The normalized periodic function $d(\zeta)$ has the unit period for $\zeta = z/\epsilon$ and zero average: $d(\zeta+1) = d(\zeta)$ and $\int_0^1 d(\zeta) d\zeta = 0$. It is defined explicitly as

$$\begin{aligned} d &= \frac{2}{l} \quad \text{for} \quad \zeta \in [0, l), \\ d &= \frac{2}{l-1} \quad \text{for} \quad \zeta \in [l, 1), \end{aligned} \tag{1.7}$$

where $0 < l < 1$ is the ratio of the first fiber leg to the total map period, i.e., $l = L_1/L_{\rm DM}$. We assume that the first leg is for the focusing fiber, i.e., $D_1 > 0$, and the second leg is for the defocusing fiber, i.e., $D_2 < 0$. As a result, the parameter m is positive, $m > 0$. The general problem (1.5) has four parameters:

- m (> 0)—the strength of the local (varying) dispersion,
- d_0—the strength of the average dispersion,
- ϵ (> 0)—the period of the dispersion map,
- l $(0 < l < 1)$—the relative length of the focusing fiber leg to the total map period.

Parameters m and ϵ can be normalized to unity by applying the transformation to the periodic NLS equation (1.5):

$$\zeta = \frac{z}{\epsilon}, \qquad \tau = \frac{t}{\sqrt{m}}, \qquad w(\zeta, \tau) = \sqrt{\epsilon} u(z, t), \tag{1.8}$$

where $w(\zeta, \tau)$ solves the standardized periodic NLS equation:

$$i \frac{\partial w}{\partial \zeta} + \frac{1}{2} d(\zeta) \frac{\partial^2 w}{\partial \tau^2} + \frac{1}{2} D_0 \frac{\partial^2 w}{\partial \tau^2} + |w|^2 w = 0, \qquad D_0 = \frac{\epsilon d_0}{m}. \tag{1.9}$$

Thus, the periodic NLS equation (1.9) depends only on two parameters: l (through $d(\zeta)$) and D_0.

In this paper, we study a formal asymptotic limit $\epsilon \to 0$ of solutions $u(z,t)$ of the periodic NLS equation in the form (1.5). This asymptotic limit corresponds to the limit of small solutions $w(\zeta, \tau)$ (in a $L^2(\mathbb{R})$ norm) of the periodic NLS equation in the form (1.9).

1.3. DM solitons and main results on existence and stability. DM solitons can be defined as special solutions of the periodic NLS equation (1.5) in two conventional approximations: (i) Gaussian pulse [7, 8] and (ii) an averaged integral NLS equation [14, 15].

DEFINITION 1.1. *A DM soliton is an approximate quasi-periodic solution of the NLS equation* (1.5) *in the form of the Gaussian pulse with variable coefficients:*

$$u(z,t) = \sqrt{c}\, \exp\left(-\frac{t^2}{2(a+ib)} + i\phi \right), \tag{1.10}$$

where $a(z+\epsilon) = a(z)$, $b(z+\epsilon) = b(z)$, $\phi(z+\epsilon) = \phi(z) + \epsilon\mu$, *and*

$$c = \frac{e a^{1/2}}{\sqrt{2}(a+ib)}, \qquad e = \sqrt{\frac{2}{\pi}} \int_{-\infty}^{\infty} |u|^2(z,t) dt. \tag{1.11}$$

The three varying parameters $a(z)$, $b(z)$, and $\phi(z)$ are the pulse width, chirp, and the gauge rotation phase, respectively. The constant parameters μ and e are the phase propagation constant and the pulse energy, respectively.

The variational equations are derived by minimizing the Lagrangian density of the periodic NLS equation (1.5) at the Gaussian pulse (1.10) (see, e.g., [22]). It is then found that the varying parameters $a(z)$ and $b(z)$ satisfy the nonautonomous dynamical system:

$$\frac{da}{dz} = \frac{ea^{5/2}b}{(a^2+b^2)^{3/2}}, \tag{1.12}$$

$$\frac{db}{dz} = \frac{m}{\epsilon} d\left(\frac{z}{\epsilon}\right) + d_0 - \frac{ea^{3/2}(a^2-b^2)}{2(a^2+b^2)^{3/2}}. \tag{1.13}$$

The phase parameter $\phi(z)$ is coupled with $a(z)$ and $b(z)$ by the nonhomogeneous equation:

$$\frac{d\phi}{dz} = \frac{ea^{1/2}(3a^2+5b^2)}{8(a^2+b^2)^{3/2}}. \tag{1.14}$$

The dynamical system (1.12)–(1.13) has been studied numerically under different parameterizations (see reviews [4, 5, 6]). The system was found to be Hamiltonian [9], where the phase plane was used for matching trajectories of two autonomous systems derived for the piecewise-constant function $d(z)$. Existence of periodic solutions of (1.12)–(1.13) was recently proved by Kunze [10]. The leading-order averaging system was derived from (1.12)–(1.13) by Turitsyn et al. [11, 12]. A single branch of periodic solutions of the system was found for $d_0 \geq 0$, while two branches coexist for $d_{\min} < d_0 < 0$ at any given e [13].

DEFINITION 1.2. *DM soliton is a stationary pulse solution of the averaged integral NLS equation:*

$$\mu\hat{W}(\omega) = -\frac{1}{2}d_0\omega^2\hat{W}(\omega) + \iint_{-\infty}^{\infty} \frac{\sin\left[m(\omega-\omega_1)(\omega-\omega_2)\right]}{m(\omega-\omega_1)(\omega-\omega_2)}\hat{W}(\omega_1)\hat{W}(\omega_2)\hat{W}(\omega_1+\omega_2-\omega)d\omega_1 d\omega_2, \tag{1.15}$$

where $\hat{W}(\omega) \in H^s(\mathbb{R})$ with $s \geq 1$ and $d_0 > 0$.

The integral NLS equation (1.15) is derived from the periodic NLS equation (1.5) in the limit $\epsilon \to 0$ by using the asymptotic averaging method explained in section 3. The integral NLS equation (1.15) follows from (3.15) for stationary pulse solutions: $\hat{V}(z,\omega) = \hat{W}(\omega)e^{i\mu z}$, where $\hat{W}(\omega)$ is real function.

Existence of stationary pulse solutions of (1.15) for $d_0 > 0$ and $\mu > 0$ was proved by Zharnitsky et al. [18]. Recently Kunze proved existence of ground state solutions $\hat{W}(\omega) \in L^2(\mathbb{R})$ for $d_0 = 0$ and $\mu > 0$ [20], which was a considerably more difficult problem due to the absence of the gradient term in the Hamiltonian. Numerical results suggest nonexistence of ground state solutions for $d_0 < 0$ due to resonance of stationary pulses with linear spectrum of the averaged integral NLS equation [13, 21]. Iterations of a numerical method for finding stationary pulse solutions of (1.15) diverge for both branches of the Gaussian pulse solutions, which exist for (1.12)–(1.13) with $d_{\min} < d_0 < 0$ (see details in [21]). No rigorous results on nonexistence of ground states of (1.15) for $d_0 < 0$ are yet available.

Definitions 1.1 and 1.2 above are commuting in the sense that (i) the system (1.12)–(1.13) can be averaged in the limit $\epsilon \to 0$ [13] and (ii) the variational Gaussian approximation can be applied to the integral NLS equation (1.15) [17]. Both the reductions result in the same set of equations for an averaged Gaussian pulse. In order to analyze the parameter dependence of DM solitons, we consider the following two equivalent parameterizations.

Suppose there exist periodic solutions of (1.12)–(1.13) or stationary pulse solutions of (1.15). The DM solitons are parameterized as $e = f_\mu(\mu; d_0, l, m, \epsilon)$, where $e = f_\mu(\mu)$ is a continuous (possibly multibranched) function of μ. Indeed, solutions $\hat{W}(\omega)$ of (1.15) smoothly depend on parameter μ in the domain of their existence, where $\hat{W}(\omega) \in H^s(\mathbb{R})$ with $s \geq 1$. Then, the function $e = f_\mu(\mu)$ is defined by (1.11) as a continuous function of μ. If there are several solutions of (1.15) for the same value of μ, the function $e = f_\mu(\mu)$ has several branches for a fixed value of μ. Alternatively, solutions $(a(z), b(z))$ of (1.12)–(1.13) smoothly depend on e in the domain of their existence. Then, μ is defined by $\mu = (\phi(z+\epsilon) - \phi(z))/\epsilon = f_\mu^{-1}(e)$. The function $e = f_\mu(\mu)$ is invertible for each branch of solution, where $f'_\mu(\mu) \neq 0$.

For an alternative parameterization, we define the effective pulse width as

$$\tau^2(z) = \frac{2}{e}\sqrt{\frac{2}{\pi}} \int_{-\infty}^{\infty} t^2 |u|^2(z,t) dt \tag{1.16}$$

and the minimal pulse width as

$$\tau^2_{\min} = \min_{z \in [0,\epsilon]} \tau^2(z). \tag{1.17}$$

The DM solitons are parameterized as $e = f_s(s; d_0, l, m, \epsilon)$, where $s = 1/\tau^2_{\min}$ is the square inverse of the minimal pulse width. The function $e = f_s(s)$ is a continuous (possibly multibranched) function of s. For each branch of stationary pulse solutions of (1.15), there exists a continuous map $s = h_\mu(\mu)$ defined by (1.16). Then, the function $e = f_s(s)$ is parameterized by μ. Also, for each branch of periodic solutions of (1.12)–(1.13), there exists a continuous function $s = f_s^{-1}(e)$ defined by

$$\tau^2_{\min} = \min_{z \in [0,\epsilon]} \frac{(a^2 + b^2)}{a} = \min_{z \in [0,\epsilon]} a(z). \tag{1.18}$$

Here we have used (1.10), (1.11), and (1.17) for the first equality and (1.12) for the second equality. The function $e = f_s(s)$ is inverted for each branch of solution, where $f'_s(s) \neq 0$.

LEMMA 1.3. *DM solitons are parameterized by three parameters: $E = \epsilon e/\sqrt{m}$ (energy), $M = \epsilon\mu$ (propagation constant), and $S = ms$ (map strength).*

Proof. The statement is proved by applying transformation (1.8) to the integral quantities (1.11) and (1.16) and to the phase $\phi(z)$ of the Gaussian pulse (1.10). ∎

The DM solitons can be analyzed in the Gaussian approximation for several alternative representations in variables E, M, and S: (i) on the plane (D_0, E) for different values of S; (ii) on the plane (S, E) for different values of S_0; (iii) on the plane (S, M) for different values of D_0; and (iv) on the plane (M, E) for different values of D_0. The four equivalent representations are shown on Figure 1.1(a)–(d), where we reproduce our main results on computations of the first-order averaging theory for the Gaussian approximation. The parameter l is fixed at $l = 0.1$. The solid curves show the result of the first-order averaging theory for $\epsilon > 0$. The dotted curves show the result of the leading-order averaging theory in the limit $\epsilon = 0$.

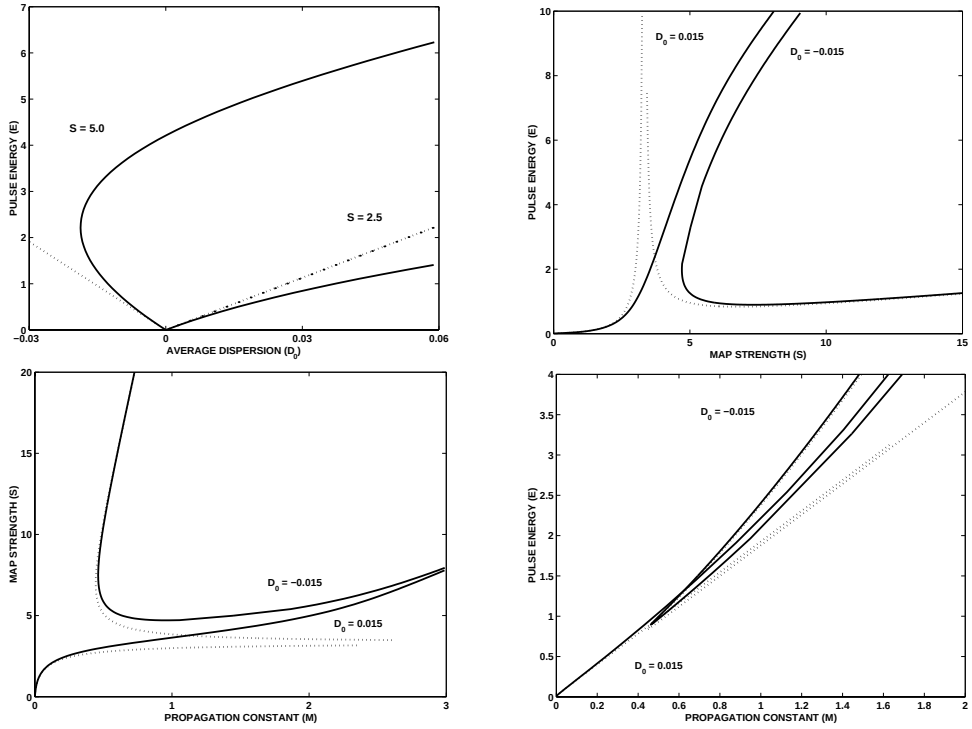


FIG. 1.1. *Parameterizations of DM solitons in the first-order averaging theory for Gaussian approximation:* (a) *plane* (D_0, E) *for* $S = 2.5$ *and* $S = 5$*;* (b) *plane* (S, E) *for* $D_0 = 0.015$ *and* $D_0 = -0.015$*;* (c) *plane* (S, M) *for* $D_0 = 0.015$ *and* $D_0 = -0.015$*; and* (d) *plane* (M, E) *for* $D_0 = 0.015$ *and* $D_0 = -0.015$*. The dotted curves display the leading-order averaging theory.*

There exists only one branch of periodic solutions of (1.12)–(1.13) for $0 < S \leq S_{\rm thr}$, where $S_{\rm thr} \approx 3.32$. This branch extends for $D_0 \geq 0$ (see Figure 1.1(a)). When $S > S_{\rm thr}$, the dependence of E versus D_0 becomes two-folded: two branches of periodic solutions exist for $D_{\rm min} < D_0 < 0$ and one branch exists for $D_0 \geq 0$. When $S \to \infty$, all branches of periodic solutions diverge to infinitely large values of E (see Figure 1.1(b)).

The role of planes (D_0, E), (S, E), and (S, M) is different from that of the plane (M, E) in the leading-order averaging theory. Indeed, the leading-order averaged system (dotted curves on Figures 1.1(a)–(d)) describes only the lower branch of periodic solutions for $D_0 < 0$ (see Figure 1.1(a)). The functions $E = f_S(S)$ and $M = h_S(S)$ are single-branched for any S (see Figures 1.1(b), (c)). However, the function $E = f_M(M)$ has two branches at the leading-order approximation (see Figure 1.1(d)), i.e., two solutions for E correspond to the same value of M, and vice verse.

For $D_0 \geq 0$, there is only one branch of periodic solutions. This branch is bounded by the threshold value $S < S_{\rm thr}$ in the leading-order approximation (see Figures 1.1(a)–(c)), while the function $E = f_M(M)$ is unbounded on the plane (M, E) (see Figure 1.1(d)). Thus, again the planes (D_0, E), (S, E), and (S, M) give bad leading-order approximations of periodic solutions compared to the plane (M, E).

The discrepancy between the four alternative parameterizations of periodic solutions of (1.12)–(1.13) disappears in the first-order averaging approximation, shown on Figures 1.1(a)–(d) by solid curves. Indeed, all curves are unbounded for $D_0 \geq 0$ and

all curves become two-valued functions of S, M, and E for any $D_0 < 0$. The upper branch of periodic solutions on the planes (D_0, E), (S, E), and (S, M) is captured in the first-order averaging theory (see Figure 1.1(a)–(c)). On the other hand, the single branch for $D_0 \geq 0$ and the two branches for $D_0 < 0$ are weakly affected on the plane (M, E) in the first-order approximation, compared to the leading-order theory (see Figure 1.1(d)). Thus, the plane (M, E) is the most appropriate tool for analysis of periodic solutions both in the leading-order and first-order averaging theory.

We can now formulate algebraically the main results of the paper for existence and stability of DM solitons. The results are written in terms of parameters ϵ, m, d_0, e, μ, and s, while the transformation to parameters D_0, E, M, and S is prescribed by Lemma 1.3. The results are proved for the Gaussian pulse approximation by explicit computations (section 2) and for the averaged integral NLS equation by standard PDE analysis (section 3).

PROPOSITION 1.4. *Parameters of DM solitons have the following expansions in powers of ϵ:*

$$d_0 = em^{1/2} g_d^{(0)}(my) + \epsilon e^2 (1-2l) g_d^{(1)}(my) + \mathrm{O}(\epsilon^2), \tag{1.19}$$

$$\mu = \frac{e}{m^{1/2}} g_\mu^{(0)}(my) + \epsilon \frac{e^2}{m} (1-2l) g_\mu^{(1)}(my) + \mathrm{O}(\epsilon^2), \tag{1.20}$$

$$s = y + \epsilon \frac{e}{m^{3/2}} g_s^{(1)}(my, l) + \mathrm{O}(\epsilon^2), \tag{1.21}$$

where y is a parameter and the functions $g_d^{(0,1)}$, $g_\mu^{(0,1)}$, and $g_s^{(1)}$ are continuous for $y > 0$.

Proof. Here we prove the result only in the limit $\epsilon \to 0$, when the leading-order averaging theory results in the integral NLS equation (1.15). The kernel of (1.15) is independent of l. Consider the scaling transformation $\hat{W}(\omega) = \lambda \hat{W}'(\omega)$. This transformation leaves (1.15) invariant if parameters μ, d_0, e, s, and m in (1.11), (1.15), and (1.16) transform as follows:

$$\mu = \lambda^2 \mu', \quad d_0 = \lambda^2 d_0', \quad e = \lambda^2 s', \quad s = s', \quad m = m'. \tag{1.22}$$

It is clear from (1.22) that the ratios d_0/e and μ/e are invariant under this transformation and are, therefore, functions of s and m. The particular form used in (1.19)–(1.20) has been chosen to match with explicit computations of functions $g_d^{(0)}(my)$ and $g_\mu^{(0)}(my)$. ∎

COROLLARY 1.5. *In the limit $\epsilon \to 0$, the functions $e = f_s(s)$, $\mu = h_s(s)$, and $e = f_\mu(\mu)$ have the form*

$$e = \frac{d_0}{\sqrt{m} g_d^{(0)}(ms)}, \quad \mu = \frac{d_0}{m} \frac{g_\mu^{(0)}(ms)}{g_d^{(0)}(ms)}, \quad \frac{\mu\sqrt{m}}{e} = g_\mu^{(0)} (g_d^{(0)})^{-1} \left(\frac{d_0}{e\sqrt{m}} \right). \tag{1.23}$$

COROLLARY 1.6. *In the limit $\epsilon \to 0$ and $d_0 = 0$, the function $e = f_\mu(\mu)$ has the form*

$$e = \alpha \mu, \qquad \alpha = \frac{\sqrt{m}}{g_\mu^{(0)}(ms_*)},$$

where $s = s_*$ is the root of the equation $g_d^{(0)}(ms_*) = 0$.

PROPOSITION 1.7. *In the limit $\epsilon \to 0$ and $m \to 0$, the functions $e = f_\mu(\mu)$ and $s = f_\mu(\mu)$ have the form*

$$e^2 = d_0 \hat{f}_\mu(\mu), \qquad s = \frac{1}{d_0}\hat{h}_\mu(\mu). \tag{1.24}$$

Proof. In the limit $m \to 0$, the integral NLS equation (1.15) becomes the Fourier form of the NLS equation. Consider the scaling transformation for the NLS equation: $\hat{W}(\omega) = \lambda \hat{W}'(\omega')$ and $\omega = \lambda^{-1}\omega'$. This transformation leaves the NLS equation invariant if parameters μ, d_0, e, and s in (1.11), (1.15), and (1.16) transform as follows:

$$\mu = \mu', \qquad d_0 = \lambda^2 d_0', \qquad e = \lambda e', \qquad s = \lambda^{-2} s'. \tag{1.25}$$

It is clear from (1.25) that the quantities e^2/d_0 and sd_0 are invariant under this transformation and are, therefore, functions of μ. ∎

PROPOSITION 1.8. *A single branch of DM solitons exists and is linearly stable for $d_0 \geq 0$ in both Gaussian and integral NLS approximations. Two branches of DM solitons exist for $d_0 < 0$ in the Gaussian approximation. For a fixed μ, the branch with larger e is linearly unstable and the branch with smaller e is linearly stable.*

Linearized stability of DM solitons in the Gaussian approximation (1.12)–(1.13) was studied by Pelinovsky in the limit $\epsilon \to 0$ [13]. We extend this analysis in the first-order averaging theory in section 2 of this paper. Zharnitsky et al. [18] proved that the DM solitons are ground states of the averaged integral NLS equation (1.15) for $d_0 > 0$. The ground states realize a stable minimum of the Hamiltonian functional. We extend this result for the first-order averaged Hamiltonian in section 3. Open problems for nonexistence of ground states for $d_0 < 0$ and nonexistence of quasi-periodic solutions of the periodic NLS equation (1.5) are discussed in section 4.

2. Variational Gaussian approximation. Here we analyze the dynamical system (1.12)–(1.13) derived in the variational Gaussian approximation. We construct the Hamiltonian structure for the system and develop a systematic averaging procedure based on the theory of canonical transformations in section 2.1. The first-order corrections to the leading-order averaged theory are computed in section 2.2. Existence and stability of critical points of the first-order averaged Hamiltonian are studied in sections 2.3 and 2.4.

The dynamical system (1.12)–(1.13) can be written as a Hamiltonian system in canonical variables (ξ, η):

$$\xi = b - m d_{-1}\left(\frac{z}{\epsilon}\right), \qquad \eta = \frac{1}{a}, \tag{2.1}$$

where $d_{-1}(\zeta)$ is the antiderivative of $d(\zeta)$ for $\zeta = z/\epsilon$ with unit period and mean zero:

$$d_{-1}(\zeta) = \int_0^\zeta d(\zeta')d\zeta' - \int_0^1 d\zeta \int_0^\zeta d(\zeta')d\zeta'. \tag{2.2}$$

For the piecewise-constant approximation $d(\zeta)$ in (1.7), the mean-zero antiderivative $d_{-1}(\zeta)$ is

$$\begin{aligned} d_{-1} &= \frac{2\zeta}{l} - 1 \quad \text{for} \quad \zeta \in [0, l), \\ d_{-1} &= \frac{2(\zeta - 1)}{(l-1)} - 1 \quad \text{for} \quad \zeta \in [l, 1). \end{aligned} \tag{2.3}$$

The system (1.12)–(1.13) in canonical variables (ξ, η) has a classical Hamiltonian structure:

$$\frac{d\xi}{dz} = \frac{\partial H}{\partial \eta}, \qquad \frac{d\eta}{dz} = -\frac{\partial H}{\partial \xi}, \tag{2.4}$$

where the Hamiltonian $H = H(\xi, \eta, z/\epsilon)$ is

$$H = d_0 \eta - e \left(\frac{\eta}{1 + \eta^2 (\xi + m d_{-1}(z/\epsilon))^2} \right)^{1/2}. \tag{2.5}$$

The decoupled equation (1.14) for the phase parameter $\phi(z)$ can be expressed through $H(\xi, \eta, z/\epsilon)$ as

$$\frac{d\phi}{dz} = \frac{1}{4} \left(d_0 \eta + \eta \frac{\partial H}{\partial \eta} - 2H \right). \tag{2.6}$$

There exists a canonical transformation from the Hamiltonian structure (2.4)–(2.5) to the one studied in [9]. The canonical transformation (2.1) and the Hamiltonian (2.5) were first reported by Turitsyn et al. [12]. The Hamiltonian structure (2.4)–(2.5) is more convenient for developing a systematic averaging procedure based on series of canonical transformations in powers of ϵ. We will study solutions of the system (2.4)–(2.5) in the domain $\mathcal{D}_+$:

$$\mathcal{D}_+ = \{ (\xi, \eta) : \ \xi \in \mathbb{R}, \ \eta > 0 \}. \tag{2.7}$$

LEMMA 2.1. *Suppose the initial point (ξ_0, η_0) belongs to $\mathcal{D}_+$. Then, a solution $(\xi(z), \eta(z))$ stays in $\mathcal{D}_+$ for any finite z: $0 \le z \le z_0 < \infty$.*

Proof. Integrating (1.12) in the canonical variables (2.1), one can find

$$\frac{1}{\eta^{1/2}} = \frac{1}{\eta_0^{1/2}} + \frac{e}{2} \int_0^z \frac{\eta (\xi + m d_{-1}) dz}{(1 + \eta^2 (\xi + m d_{-1})^2)^{3/2}}.$$

Since the integrand is never singular, the triangular inequality implies for $0 \le z \le z_0$ that

$$\left| \frac{1}{\eta^{1/2}} \right| \le \frac{1}{\eta_0^{1/2}} + \frac{eM}{2} z_0,$$

where

$$M = \max_{0 \le z \le z_0} \frac{\eta |\xi + m d_{-1}|}{(1 + \eta^2 (\xi + m d_{-1})^2)^{3/2}}.$$

Therefore, the point (η, ξ) never crosses the left boundary of $\mathcal{D}_+$ at $\eta = 0$. Direct integration of (1.12)–(1.13) with the variables (2.1) and similar estimates of the resulting integrals show that $|\xi|$ and η remain bounded in the domain $\mathcal{D}_+$ for any finite z: $0 \le z \le z_0$. ∎

2.1. Averaging of the periodic Hamiltonian system (2.4)–(2.5). The periodic Hamiltonian system (2.4)–(2.5) is averaged according to the formalism of normal form transformations [23]. We denote $\zeta = z/\epsilon$ such that $H = H(\xi, \eta, \zeta)$. In the domain $\mathcal{D}_+$ defined by (2.7), there exists a near-identity generating function:

$$F(\xi, y, \zeta) = \xi y + \sum_{n=1}^{N+1} \epsilon^n F_n(\xi, y, \zeta) + \mathrm{O}(\epsilon^{N+2}), \tag{2.8}$$

where the correction terms $F_n(x,y,\zeta)$ for $1 \le n \le (N+1)$ are periodic mean-zero functions of ζ:

$$F_n(x,y,\zeta+1) = F_n(x,y,\zeta), \qquad \int_0^1 F_n(x,y,\zeta)d\zeta = 0. \tag{2.9}$$

The generating function $F(\xi,y,\zeta)$ defines the near-identical canonical transformation

$$x = \frac{\partial F}{\partial y}(\xi,y,\zeta), \qquad \eta = \frac{\partial F}{\partial \xi}(\xi,y,\zeta) \tag{2.10}$$

and takes the Hamiltonian $H(\xi,\eta,\zeta)$ to the form

$$\begin{aligned} H_{\rm new}(x,y,\zeta) &= H(\xi(x,y,\zeta),\eta(x,y,\zeta),\zeta) + \frac{1}{\epsilon}\frac{\partial F}{\partial \zeta}(\xi(x,y,\zeta),y,\zeta) \\ &= H_N(x,y) + {\rm O}(\epsilon^{N+1}), \end{aligned} \tag{2.11}$$

where $H_N(x,y)$ is the Nth-order averaged Hamiltonian:

$$H_N(x,y) = \sum_{n=0}^{N} \epsilon^n h_n(x,y). \tag{2.12}$$

When the remainder term of order of ${\rm O}(\epsilon^{N+1})$ is neglected, the new canonical variables (x,y) solve the averaged Hamiltonian dynamical system:

$$\frac{dx}{dz} = \frac{\partial H_N}{\partial y}, \qquad \frac{dy}{dz} = -\frac{\partial H_N}{\partial x}. \tag{2.13}$$

The difference between solutions of the full system (2.4) and the averaged system (2.13) is controlled with the accuracy of ${\rm O}(\epsilon^{N+1})$ on the interval $0 \le z \le z_0$. Convergence and bounds of the normal-form transformations in Hamiltonian systems with fast dependence on time was proved by Neishtadt [24].

The canonical transformation (2.10)–(2.11) follows from the invariance of the Lagrangian of the system (2.4) [23]:

$$\mathcal{L} = \eta\frac{d\xi}{dz} - H(\xi,\eta,\zeta) = -x\frac{dy}{dz} - H_{\rm new}(x,y,\zeta) + \frac{dF}{dz}(\xi,y,\zeta).$$

In the domain $\mathcal{D}_+$, the Hamiltonian $H(\xi,\eta,\zeta)$ is a C^∞ function of ξ and η. Then, the generating functions $F_n(\xi,y,\zeta)$ are C^∞ functions of ξ and y. Provided the asymptotic series (2.8) converges uniformly, there exists ϵ_0 such that for $0 \le \epsilon \le \epsilon_0$

$$\frac{\partial^2 F}{\partial y \partial \xi} = 1 + \sum_{n=1}^{N+1} \epsilon^n \frac{\partial^2 F}{\partial y \partial \xi} + {\rm O}(\epsilon^{N+2}) > 0.$$

According to the inverse function theorem, the near-identity transformations (2.10) define classical perturbation series for $\xi(x,y,\zeta)$ and $\eta(x,y,\zeta)$ in powers of ϵ (see Chapter 2.2(a) in [23]). Here, ζ is the fast "time" for periodic oscillations of $H(\xi,\eta,z/\epsilon)$ and z is the slow "time" for averaged dynamics of the new canonical variables (x,y).

For $N = 0$, the leading-order averaged dynamical system is

$$\frac{dx}{dz} = \frac{\partial h_0}{\partial y}, \qquad \frac{dy}{dz} = -\frac{\partial h_0}{\partial x}, \tag{2.14}$$

where $h_0(x,y)$ is the leading-order averaged Hamiltonian $H_0(x,y)$:

$$H_0(x,y) = h_0(x,y) = \int_0^1 H(x,y,\zeta)d\zeta. \tag{2.15}$$

The leading-order averaged Hamiltonian can be computed explicitly from (2.3), (2.5), and (2.15) as

$$h_0(x,y) = d_0 y - \frac{e}{2my^{1/2}} \log\left[f_0(x,y)\right], \tag{2.16}$$

where

$$f_0(x,y) = \frac{y(x+m) + (1+y^2(x+m)^2)^{1/2}}{y(x-m) + (1+y^2(x-m)^2)^{1/2}}. \tag{2.17}$$

The leading-order Hamiltonian $h_0(x,y)$ does not depend on parameter l. However, the first-order correction term $h_1(x,y)$ does depend on l in the first-order averaged Hamiltonian $H_1(x,y)$.

2.2. First-order averaged Hamiltonian $H_1(x,y)$. The first-order averaged Hamiltonian can be easily derived from the formalism of the normal form transformations. It follows from (2.8), (2.10), and (2.11) that the first-order correction term $F_1(x,y,\zeta)$ is the periodic mean-zero function of ζ:

$$F_1(x,y,\zeta) = \left\{h_0(x,y) - H(x,y,\zeta)\right\}_{-1}, \tag{2.18}$$

where $\{H(x,y,\zeta)\}_{-1}$ is the mean-zero antiderivative of $H(x,y,\zeta)$ in ζ; see (2.2). Expanding the near-identity canonical transformations (2.8) and (2.10) in powers of ϵ, we define the perturbation series for $\xi(x,y,\zeta)$ and $\eta(x,y,\zeta)$:

$$\xi = x + \epsilon\xi_1(x,y,\zeta) + \mathrm{O}(\epsilon^2), \qquad \xi_1 = -\frac{\partial F_1}{\partial y}(x,y,\zeta), \tag{2.19}$$

$$\eta = y + \epsilon\eta_1(x,y,\zeta) + \mathrm{O}(\epsilon^2), \qquad \eta_1 = \frac{\partial F_1}{\partial x}(x,y,\zeta). \tag{2.20}$$

Similarly, the first-order correction term $h_1(x,y)$ is found in the form

$$h_1(x,y) = \int_0^1 \left(-\frac{\partial H}{\partial x}\frac{\partial F_1}{\partial y} + \frac{\partial H}{\partial y}\frac{\partial F_1}{\partial x} - \frac{\partial^2 F_1}{\partial\zeta\partial x}\frac{\partial F_1}{\partial y} \right)(x,y,\zeta)d\zeta \tag{2.21}$$

$$= \int_0^1 \left(\frac{\partial H}{\partial y}\frac{\partial F_1}{\partial x} \right)(x,y,\zeta)d\zeta, \tag{2.22}$$

where we have used (2.9) and (2.18) for the second equality in (2.22). The first-order averaged dynamical system (2.13) then takes the form

$$\frac{dx}{dz} = \frac{\partial h_0}{\partial y} + \epsilon\frac{\partial h_1}{\partial y}, \qquad \frac{dy}{dz} = -\frac{\partial h_0}{\partial x} - \epsilon\frac{\partial h_1}{\partial x}. \tag{2.23}$$

A remarkable result is that the first-order correction term $h_1(x,y)$ vanishes in the case of symmetric maps, when $l = 1/2$.

LEMMA 2.2. *When the dispersion map is symmetric, i.e., $l = 1/2$, then $h_1(x,y) = 0$ and $F_1(x,y,0) = F_1(x,y,l) = 0$.*

Proof. When $d(\zeta) = 4$ for $z \in [0, 1/2)$ and $d(\zeta) = -4$ for $z \in [-1/2, 0)$ (see (1.7)), the mean-zero antiderivative function $d_{-1}(\zeta)$ is even in ζ, i.e., $d_{-1}(-\zeta) = d_{-1}(\zeta)$ (see (2.3)). As a result, the Hamiltonian $H(x, y, \zeta)$ in (2.5) can be represented by the Fourier cosine-series:

$$H(x, y, \zeta) = h_0(x, y) + \sum_{n=1}^{\infty} c_n(x, y) \cos(2\pi n\zeta), \tag{2.24}$$

where $c_n(x, y)$ are some Fourier coefficients. As a result, the first-order correction term $F_1(x, y, \zeta)$ given by (2.18) is computed as the Fourier sine-series:

$$F_1(x, y, \zeta) = -\sum_{n=1}^{\infty} \frac{1}{2\pi n} c_n(x, y) \sin(2\pi n\zeta). \tag{2.25}$$

It is clear that $F_1(x, y, 0) = F_1(x, y, 1/2) = 0$. The first-order correction $h_1(x, y)$ given by (2.22) is the average of the product of the Fourier cosine- and sine-series, which is zero. □

In general case, when $l \neq 1/2$, the first-order averaging theory is equivalent to the following result. If (x, y) solve the averaged equations (2.23) and (ξ, η) solve the full equations (2.4)–(2.5) with close initial values—$|\xi_0 - x_0 - \epsilon\xi_1(x_0, y_0, 0)| \leq c_x\epsilon^2$ and $|\eta_0 - y_0 - \epsilon\eta_1(x_0, y_0, 0)| \leq c_y\epsilon^2$, where c_x and c_y are some constants, then the solutions (x, y) and (ξ, η) remain within the linear accuracy in ϵ at the distances $0 \leq z \leq z_0$:

$$\sup_{z\in[0,z_0]} |\xi(z) - x(z) - \epsilon\xi_1(x, y, \zeta)| \leq C_x\epsilon^2, \qquad \sup_{z\in[0,z_0]} |\eta(z) - y(z) - \epsilon\eta_1(x, y, \zeta)| \leq C_y\epsilon^2, \tag{2.26}$$

where C_x and C_y are some constants. The standard proof of this statement is based on convergence of the perturbation series (2.19)–(2.20) [23].

When the dispersion map is symmetric with equal legs, i.e., $l = 1/2$, the corrections $\xi_1(x, y, \zeta)$ and $\eta_1(x, y, \zeta)$ vanish at the points $\zeta = 0$ and $\zeta = \frac{1}{2}$. As a result, the distance between solutions (x, y) and (ξ, η) remains within the quadratic accuracy in ϵ at the ends of the dispersion map, i.e., at $z = k\epsilon$ and $z = (k - \frac{1}{2})\epsilon$, where $k \in \mathbb{Z}_+$:

$$\sup_{z\in[0,z_0]} |\xi(z = k\epsilon) - x(z = k\epsilon)| \leq C_x\epsilon^2, \qquad \sup_{z\in[0,z_0]} |\eta(z = k\epsilon) - y(z = k\epsilon)| \leq C_y\epsilon^2. \tag{2.27}$$

This result is related to the Strang's work [25] on symmetrization of the split-step methods for solving PDEs. The quadratic convergence occurs only at the ends of the dispersion map, while it is linear in the interior of the dispersion map.

Remark 2.1. The symmetric dispersion map with $l = 1/2$ can be translated for any ζ_0 such that $d(\zeta + \zeta_0) = -d(\zeta_0 - \zeta)$. The first-order correction $h_1(x, y)$ vanishes for all such symmetric maps. In particular, the symmetric dispersion map used in numerical modeling of the NLS equation by the split-step method is $d(\zeta) = 4$ for $\zeta \in [0, 1/4) \cup [3/4, 1)$ and $d(\zeta) = -4$ for $\zeta \in [1/4, 3/4)$. This map is equivalent to our symmetric map with $l = 1/2$ by the translation with $\zeta_0 = 1/4$.

The first-order correction term $h_1(x, y)$ can be found explicitly by direct compu-

tations from (2.3), (2.5), (2.16), (2.18), and (2.22). The explicit formula is

$$\begin{aligned} h_1(x,y) &= \frac{e^2(1-2l)}{8m^2}\left[\frac{(x+m)}{1+y^2(x+m)^2}-\frac{(x-m)}{1+y^2(x-m)^2}\right] \\ &+ \frac{e^2(1-2l)}{16m^3y^2}\left[\log\left[f_0(x,y)\right]+\frac{2y(x-m)}{(1+y^2(x-m)^2)^{1/2}}-\frac{2y(x+m)}{(1+y^2(x+m)^2)^{1/2}}\right] \\ &\times\left[\log\left[f_0(x,y)\right]+\frac{xy}{(1+y^2(x-m)^2)^{1/2}}-\frac{xy}{(1+y^2(x+m)^2)^{1/2}}\right] \\ &+ \frac{e^2(1-2l)}{16m^3y^2}\left[\frac{3+y^2(x+m)^2}{(1+y^2(x+m)^2)^{1/2}}-\frac{3+y^2(x-m)^2}{(1+y^2(x-m)^2)^{1/2}}\right] \\ &\times\left[\frac{1}{(1+y^2(x+m)^2)^{1/2}}-\frac{1}{(1+y^2(x-m)^2)^{1/2}}\right], \end{aligned} \tag{2.28}$$

where $f_0(x,y)$ is defined by (2.17). We confirm from (2.28) that $h_1(x,y)=0$ for $l=1/2$. The first-order averaged Hamiltonian $H_1(x,y)$ is analyzed next for existence and stability of critical points. The critical points of the averaged Hamiltonian correspond to the Gaussian approximation of the DM solitons.

2.3. Existence of critical points of the first-order averaged Hamiltonian. The first-order averaged Hamiltonian is $H_1(x,y)=h_0(x,y)+\epsilon h_1(x,y)$, where $h_0(x,y)$ and $h_1(x,y)$ are given explicitly in (2.16) and (2.28).

LEMMA 2.3. *The points $(0,y_*)$ are critical points of the first-order averaged Hamiltonian $H_1(x,y)$ if y_* is an extremum of the function $H_1(0,y)$:*

$$\begin{aligned} H_1(0,y) &= d_0 y - \frac{e}{2my^{1/2}}\log\left[\hat{f}_0(my)\right] \\ &+ \epsilon\frac{e^2(1-2l)}{16m^3y^2}\left[\log^2\left[\hat{f}_0(my)\right]-\frac{4my}{(1+m^2y^2)^{1/2}}\log\left[\hat{f}_0(my)\right]+\frac{4m^2y^2}{1+m^2y^2}\right], \end{aligned} \tag{2.29}$$

where $\hat{f}_0(my)=f_0(0,y)$, i.e.,

$$\hat{f}_0(my)=\left[\frac{(1+m^2y^2)^{1/2}+my}{(1+m^2y^2)^{1/2}-my}\right].$$

Proof. The variation of $h_0(x,y)$ in x leads to the only solution $x=0$. The same solution gives also an extremum of $h_1(x,y)$ in x. The variation of $H_1(0,y)$ in y defines the critical point $y=y_*$. □

Proof of Proposition 1.4. The first equation (1.19) follows from the condition $H_1'(0,y)=0$. The continuous functions $g_d^{(0,1)}$ are computed explicitly:

$$g_d^{(0)} = -\frac{1}{4m^{3/2}y^{3/2}}\left[\log\left[\hat{f}_0(my)\right]-\frac{4my}{(1+m^2y^2)^{1/2}}\right], \tag{2.30}$$

$$\begin{aligned} g_d^{(1)} &= \frac{1}{8m^3y^3}\left[\log^2\left[\hat{f}_0(my)\right]\right. \\ &\left.- \frac{2my(2+3m^2y^2)}{(1+m^2y^2)^{3/2}}\log\left[\hat{f}_0(my)\right]+\frac{4m^2y^2(1+2m^2y^2)}{(1+m^2y^2)^2}\right]. \end{aligned} \tag{2.31}$$

The second equation (1.20) follows from the nonhomogeneous equation (2.6) reduced for the perturbation expansion (2.19)–(2.20):

$$\frac{d\phi}{dz} = \frac{1}{4}\left[d_0 y + y\frac{dx}{dz} + y\frac{\partial \xi_1}{\partial \zeta} - 2H(x,y,\zeta) + \epsilon\left(d_0\eta_1 + \eta_1\frac{\partial H}{\partial y}(x,y,\zeta)\right.\right. \\ \left.\left. + \; y\left(\frac{\partial \xi_1}{\partial x}\frac{dx}{dz} + \frac{\partial \xi_1}{\partial y}\frac{dy}{dz}\right) + y\frac{\partial \xi_2}{\partial \zeta} - 2\left(\frac{\partial H}{\partial x}\xi_1 + \frac{\partial H}{\partial y}\eta_1\right)(x,y,\zeta)\right) + \mathrm{O}(\epsilon^2)\right]. \tag{2.32}$$

Integrating (2.33) over $\zeta \in [0,1]$ at the critical point $(0, y_*)$, we define μ as

$$\mu = \frac{1}{\epsilon}\left(\phi(z+\epsilon) - \phi(z)\right) = \frac{1}{4}\left[d_0 y_* - 2h_0(0,y_*) - 3\epsilon h_1(0,y_*) + \mathrm{O}(\epsilon^2)\right], \tag{2.33}$$

where we have utilized (2.15) and (2.22). The continuous functions $g_d^{(0,1)}$ in (1.20) are computed explicitly:

$$g_\mu^{(0)} = \frac{1}{16 m^{1/2} y^{1/2}}\left[5\log\left[\hat{f}_0(my)\right] - \frac{4my}{(1+m^2y^2)^{1/2}}\right], \tag{2.34}$$

$$g_\mu^{(1)} = -\frac{1}{64 m^2 y^2}\left[5\log^2\left[\hat{f}_0(my)\right]\right. \\ \left. - \; \frac{2my(8+9m^2y^2)}{(1+m^2y^2)^{3/2}}\log\left[\hat{f}_0(my)\right] + \frac{4m^2y^2(4+5m^2y^2)}{(1+m^2y^2)^2}\right]. \tag{2.35}$$

The third equation (1.21) follows from (1.18):

$$s = \frac{1}{\tau_{\min}^2} = \max_{0\le z\le \epsilon}\eta(z) = y + \epsilon \max_{0\le\zeta\le 1}\eta_1(0,y_*,\zeta) + \mathrm{O}(\epsilon^2). \tag{2.36}$$

If (a,b) is a nonconstant periodic solution of z, then $a(z)$ has at least two extremal points on the interval $z \in [0,\epsilon)$. The extremal values for $a(z)$ occur for $z = z_*$, where $b(z_*) = 0$; see (1.12). It follows from (2.3) and (2.18) that $d_{-1}(\zeta_*) = 0$ and $F_1(0,y_*,\zeta_*) = 0$ for $\zeta_* = l/2$ and $\zeta_* = (1+l)/2$. As a result, $b(z_*) = \mathrm{O}(\epsilon^2)$ and $\xi(\zeta_*) = \mathrm{O}(\epsilon^2)$, see (2.1) and (2.19), i.e., $a(z)$ and $\eta(z)$ have extrema for $z_* = \epsilon\zeta_*$. Computing the derivative of $F_1(x,y,\zeta)$ in x for $(0,y_*,\zeta)$, we find that the maximal value of $\eta_1(0,y_*,\zeta_*)$ occurs at $\zeta_* = l/2$ and the continuous functions $g_s^{(1)}$ in (1.21) are computed explicitly:

$$g_s^{(1)} = \frac{m^{1/2}y^{1/2}}{2}\left(l + \frac{l-1}{(1+m^2y^2)^{1/2}}\right) + \frac{(1-2l)}{4m^{1/2}y^{1/2}}\log\left[\hat{f}_0(my)\right]. \tag{2.37}$$

□

Figures 1.1(a)–(d) are constructed with the help of explicit formulas (1.19)–(1.21), (2.30)–(2.31), (2.34)–(2.35), and (2.37). The dotted curves show the limit $\epsilon = 0$, the solid curves show the results of the first-order averaging theory for $\epsilon > 0$, with $l = 0.1$ fixed. The first-order averaging theory corresponds well to numerical analysis of the full equations (1.12)–(1.13); see [4, 7, 8].

2.4. Stability of critical points of the first-order averaged Hamiltonian. Linear stability of the critical points $(0, y_*)$ in the first-order averaged system (2.23) is defined by concavity of the quadratic form:

$$H_1(x,y) - H_1(0,y_*) = \frac{1}{2}\frac{\partial^2 H_1}{\partial x^2}\bigg|_{(0,y_*)}x^2 + \frac{1}{2}\frac{\partial^2 H_1}{\partial y^2}\bigg|_{(0,y_*)}(y-y_*)^2, \tag{2.38}$$

since the derivative of $H_1(x,y)$ in x is zero for any $(0,y)$. The critical point $(0,y_*)$ is linearly stable if it corresponds to an extremum of the quadratic form (2.38); the stable critical points are centers on the phase plane (x,y). The critical point is linearly unstable if it corresponds to a saddle point of the quadratic form (2.38). The unstable critical points appear as saddle points on the phase plane (x,y). It is easy to analyze the linear stability of the critical point $(0,y_*)$ with the help of the function $e = f_s(s)$ shown on Figure 1.1(b).

LEMMA 2.4. *The critical point $(0,y_*)$ of the first-order averaged Hamiltonian $H_1(x,y)$ is stable for $d_0 \geq 0$. For $d_0 < 0$, define $s_{\rm thr}$ and $s_{\rm stab}$ as the turning and minimal points of the curve $e = f_s(s)$, i.e., $f_s'(s_{\rm thr}) = \infty$ and $f_s'(s_{\rm stab}) = 0$. The critical point $(0,y_*)$ is stable for $d_0 < 0$ in the following cases:* (i) *for the upper branch of the curve $e = f_s(s)$, when $s \geq s_{\rm thr}$ and* (ii) *for the lower branch of the curve $e = f_s(s)$, when $s_{\rm thr} < s < s_{\rm stab}$.*

Proof. It follows from (2.16) that

$$\left.\frac{\partial^2 h_0}{\partial x^2}\right|_{x=0} = \frac{e y^{5/2}}{(1+m^2y^2)^{3/2}} > 0$$

for any $y > 0$. Therefore, there exists ϵ_0 such that the curvature of $H_1(x,y)$ is positive in x for $0 \leq \epsilon \leq \epsilon_0$. Then, the stability criterion is $H_1''(0,y_*) > 0$, where

$$H_1''(0,y) = -em^{3/2} g_d^{(0)\prime}(my) - \epsilon e^2 m(1-2l) g_d^{(1)\prime}(my) = f_s'(s) W(s),$$

where

$$W(s) = \left[g_d^{(0)}(my) + 2\epsilon e(1-2l) g_d^{(1)}(my) + \mathrm{O}(\epsilon^2) \right] \left[1 + \epsilon \frac{e}{m^{1/2}} g_s^{(1)\prime}(my,l) + \mathrm{O}(\epsilon^2) \right].$$

For $d_0 > 0$, it follows from Figure 1.1(b) and (1.19) that $f_s'(s) > 0$ and $g_d^{(0)}(my) > 0$ for any $y > 0$. As a result, there exists ϵ_0 such that the curvature of $H_1(0,y)$ is positive in y for $0 \leq \epsilon \leq \epsilon_0$. Therefore, the critical point $(0,y_*)$ is stable for $d_0 \geq 0$ and $0 \leq \epsilon \leq \epsilon_0$.

For $d_0 < 0$, the upper and lower branches of the function $e = f_s(s)$ are separated by the point $s = s_{\rm thr}$, where $f_s'(s_{\rm thr}) = \infty$ (see Figure 1.1(b)). Since $H_1''(0,y)$ may not be singular in the domain $y > 0$, the function $W(s)$ changes sign at $s = s_{\rm thr}$. It follows from (1.19) that $g_d^{(0)}(my) < 0$ for $d_0 < 0$. Therefore, it is clear that $W(s) > 0$ for the upper branch of $e = f_s(s)$ and $W(s) < 0$ for the lower branch of $e = f_s(s)$ on Figure 1.1(b). On the other hand, $f_s'(s) > 0$ for the upper branch of $e = f_s(s)$ and $f_s'(s) < 0$ for the lower branch of $e = f_s(s)$ between $s_{\rm thr} < s < s_{\rm stab}$, where $f_s'(s_{\rm stab}) = 0$; see Figure 1.1(b). As a result, the curvature of $H_1(0,y)$ in y is positive for the two cases (i) and (ii). For the lower branch of $e = f_s(s)$ at $s > s_{\rm stab}$, the curvature of $H_1(0,y)$ in y is negative, since $f_s'(s) > 0$ and $W(s) < 0$. As a result, the critical point $(0,y_*)$ is linearly unstable for the lower branch of $e = f_s(s)$ at $s > s_{\rm stab}$. □

Lemma 2.4 corresponds to Proposition 1.8 for Gaussian pulses in the first-order averaging theory. At the plane (μ,e), the point $s = s_{\rm stab}$ is the point of minimal e, i.e., it is a branching point of the function $e = f_\mu(\mu)$. As a result, for $d_0 < 0$, the upper branch of the function $e = f_\mu(\mu)$ is linearly unstable, while the lower branch of the function $e = f_\mu(\mu)$ is linearly stable [13].

We compute the phase plane $H_1(x,y) = \text{const}$ of the first-order averaged Hamiltonian from (2.16) and (2.28). The phase plane is shown on Figure 2.1(a)–(b) in standardized variables $X = x/m$ and $Y = my$ for $D_0 = 0.015$ and $D_0 = -0.015$,

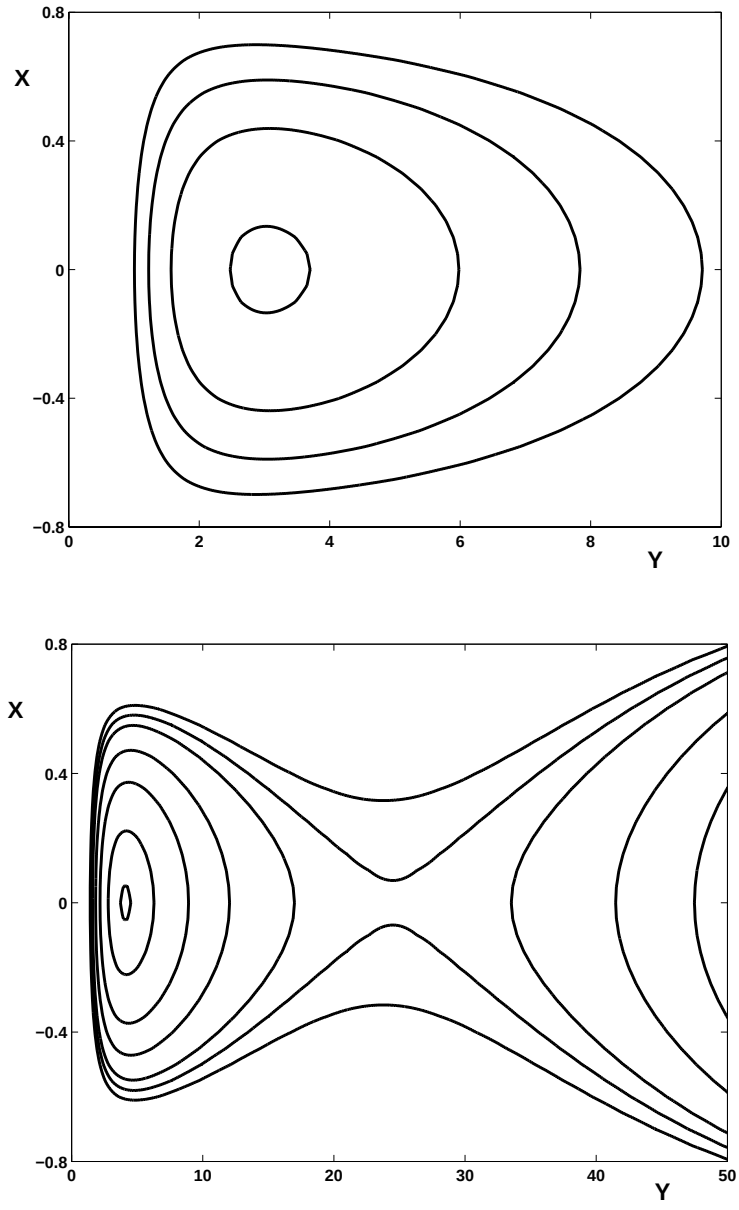


FIG. 2.1. *The contour levels of the first-order averaged Hamiltonian* $H_1(X, Y)$ *for* $D_0 = 0.015$ (a) *and* $D_0 = -0.015$ (b). *The other parameters are* $E = 2$ *and* $l = 0.1$.

with $l = 0.1$. In the initial-value problem, the energy parameter is constant, taken as $E = 2$.

For $D_0 \geq 0$ there is only one critical point, which is a center (see Figure 2.1(a)). The trajectories of the dynamical system (2.23) are all closed around the stable center point. This dynamics corresponds to small oscillations of the DM soliton, perturbed by an initial condition.

For $D_0 < 0$, two critical points coexist for the same value of E. The critical point with a larger value of Y_* is unstable (saddle point), while that with smaller value of Y_* is stable (center) (see Figure 2.1(b)). The critical point with a larger value of Y_* corresponds to a shorter DM soliton. If the shorter soliton is shortened by an initial perturbation, i.e., $Y(0) > Y_*$, it is destroyed, since the trajectory (X, Y) is unbounded on the phase plane of the first-order averaged system. We speculate that the shorter soliton transforms into chirped quasi-linear waves but this process is beyond the variational Gaussian approximation. Since solutions of (1.12)–(1.13) are bounded in the domain $\mathcal{D}_+$ for any finite z, the transformation happens over infinite propagation distances z.

In the other case, when the shorter DM soliton is broadened by the initial perturbation, i.e., $Y(0) < Y_*$, the trajectory (X, Y) is trapped inside the separatrix loop of the center point. In this case, the pulse undertakes large-amplitude oscillations around the stable longer DM soliton, similarly to the case $D_0 \geq 0$. Instability of short DM solitons along the lower branch of the (E, S) curve is confirmed in numerical computations [7].

3. Reduction to an averaged integral NLS equation. Here we analyze and extend the integral NLS equation (1.15), derived by means of averaging of the periodic NLS equation (1.5). We develop a formal method of canonical transformations for PDEs in section 3.1. In the leading and first orders in powers of ϵ, we prove convergence of the periodic NLS equation (1.5) to an averaged integral NLS equation in section 3.2. Existence of ground states of the first-order averaged Hamiltonian is proved for the case $d_0 > 0$ in section 3.3.

The periodic NLS equation (1.5) has the standard Hamiltonian structure:

$$i\frac{\partial u}{\partial z} = \frac{\delta H}{\delta \bar{u}}, \qquad -i\frac{\partial \bar{u}}{\partial z} = \frac{\delta H}{\delta u}, \tag{3.1}$$

where the Hamiltonian $H = H(u, \bar{u}, z/\epsilon)$ is

$$H = \frac{1}{2}\int_{-\infty}^{\infty} \left[\frac{m}{\epsilon} d\left(\frac{z}{\epsilon}\right) \left|\frac{\partial u}{\partial t}\right|^2 + d_0 \left|\frac{\partial u}{\partial t}\right|^2 - |u|^4 \right] dt. \tag{3.2}$$

LEMMA 3.1. *Define a fundamental solution of the linear periodic equation*

$$i\frac{\partial u}{\partial z} + \frac{m}{2\epsilon} d\left(\frac{z}{\epsilon}\right) \frac{\partial^2 u}{\partial t^2} = 0 \tag{3.3}$$

in the operator form:

$$u(z,t) = T\left(\frac{z}{\epsilon}\right) u(0,t). \tag{3.4}$$

The operator $T(\zeta)$ for $\zeta = z/\epsilon$ is a unitary operator with unit period:

$$T^{-1}(\zeta) = \overline{T}(\zeta), \qquad T(\zeta + 1) = T(\zeta), \tag{3.5}$$

where $\bar{T}$ is complex conjugate.

Proof. In the Fourier space of t, the operator $T(\zeta)$ is a multiplication operator:

$$\hat{u}(z,\omega) = T_\omega(\zeta)\hat{u}(0,\omega), \qquad T_\omega(\zeta) = e^{-\frac{im}{2} d_{-1}(\zeta)\omega^2}, \tag{3.6}$$

where $d_{-1}(\zeta)$ is given by (2.2)–(2.3) and $\hat{u}(\zeta,\omega)$ is the Fourier transform of $u(\zeta,t)$ in t. The two properties (3.5) follow from the Fourier form (3.6). □

Using a linear canonical transformation

$$u(z,t) = T(\zeta)v(z,t), \qquad \bar{u}(z,t) = T^{-1}(\zeta)\bar{v}(z,t), \qquad \zeta = \frac{z}{\epsilon}, \tag{3.7}$$

we eliminate the fast periodic term from (1.5) and rewrite the Hamiltonian system (3.1) in new canonical variables $(v,\bar{v})$,

$$i\frac{\partial v}{\partial z} = \frac{\delta H}{\delta \bar{v}}, \qquad -i\frac{\partial \bar{v}}{\partial z} = \frac{\delta H}{\delta v}, \tag{3.8}$$

with the new Hamiltonian $H = H(v,\bar{v},\zeta)$:

$$H = \frac{1}{2}\int_{-\infty}^{\infty}\left[d_0\left|\frac{\partial v}{\partial t}\right|^2 - |T(\zeta)v|^4\right]dt. \tag{3.9}$$

The periodic NLS equation in new variables $(v,\bar{v})$ can be written in the operator form:

$$i\frac{\partial v}{\partial z} + \frac{1}{2}d_0\frac{\partial^2 v}{\partial t^2} + T^{-1}(\zeta)\left(|T(\zeta)v|^2\,T(\zeta)v\right) = 0. \tag{3.10}$$

In the Fourier space of t, the operator equation (3.10) takes the form of a periodic integral NLS equation:

$$i\frac{\partial \hat{v}}{\partial z}(\omega) - \frac{1}{2}d_0\omega^2\hat{v}(\omega) + \int\!\!\int_{-\infty}^{\infty} K_\omega(\zeta)\hat{v}(\omega_1)\hat{v}(\omega_2)\bar{\hat{v}}(\omega_1+\omega_2-\omega)d\omega_1 d\omega_2 = 0, \tag{3.11}$$

where $K_\omega(\zeta)$ is defined by

$$K_\omega(\zeta) = e^{-\frac{im}{2}d_{-1}(\zeta)(\omega_1^2+\omega_2^2-(\omega_1+\omega_2-\omega)^2-\omega^2)} = e^{imd_{-1}(\zeta)(\omega-\omega_1)(\omega-\omega_2)}. \tag{3.12}$$

The asymptotic reduction of (3.11) to an integral NLS equation is based on averaging of the Hamiltonian (3.9) in ζ [18]. A direct averaging method produces the following leading-order averaged Hamiltonian $H_0(V,\bar{V})$:

$$H_0(V,\bar{V}) = h_0(V,\bar{V}) = \int_0^1 H(V,\bar{V},\zeta)d\zeta = \frac{1}{2}\int_{-\infty}^{\infty}\left[d_0\left|\frac{\partial V}{\partial t}\right|^2 - \int_0^1 |T(\zeta)V|^4 d\zeta\right]dt. \tag{3.13}$$

The leading-order averaged Hamiltonian $H_0(V,\bar{V})$ generates the averaged integral NLS equation in the operator form:

$$i\frac{\partial V}{\partial z} + \frac{1}{2}d_0\frac{\partial^2 V}{\partial t^2} + \int_0^1 T^{-1}(\zeta)\left(|T(\zeta)v|^2\,T(\zeta)v\right)d\zeta = 0. \tag{3.14}$$

In the Fourier space of t, the integral NLS equation (3.14) takes an explicit form:

$$i\frac{\partial \hat{V}}{\partial z}(\omega) - \frac{1}{2}d_0\omega^2\hat{V}(\omega) + \int\!\!\int_{-\infty}^{\infty}\langle K_\omega\rangle\hat{V}(\omega_1)\hat{V}(\omega_2)\bar{\hat{V}}(\omega_1+\omega_2-\omega)d\omega_1 d\omega_2 = 0, \tag{3.15}$$

where $\langle K_\omega \rangle$ is the average of $K_\omega(\zeta)$ over $\zeta \in [0,1]$. When the antiderivative function $d_{-1}(\zeta)$ is given by (2.3), the kernel $\langle K_\omega \rangle$ is computed explicitly as

$$\langle K_\omega \rangle = \frac{\sin m(\omega-\omega_1)(\omega-\omega_2)}{m(\omega-\omega_1)(\omega-\omega_2)}. \tag{3.16}$$

The integral equation (3.15) with the kernel (3.16) becomes (1.15) for stationary pulse solutions: $\hat{V}(z,\omega) = \hat{W}(\omega)e^{i\mu z}$.

The asymptotic reduction of the periodic NLS equation (1.5) to the averaged integral NLS equation (3.15) was first derived in [14, 15]. Higher-order corrections to the averaged integral NLS equation were considered in [26, 27] with the help of formal Lie transformations. We develop a method of formal canonical transformations for the Hamiltonian $H(v,\bar{v},\zeta)$ and, in addition, we prove convergence of the averaging procedure at the leading and first orders in powers of ϵ.

3.1. Averaging of the periodic integral NLS equation (3.11)–(3.12). The periodic integral NLS equation (3.11)–(3.12) can be averaged with the help of the normal form transformations, formally generalized for infinite-dimensional Hamiltonian systems. In this generalization, the generating functional $F(v,\bar{V},\zeta)$ replaces the generating function $F(\xi,y,\zeta)$ (see (2.8)):

$$F(v,\bar{V},\zeta) = \int_{-\infty}^{\infty} dt \left[v\bar{V} + \sum_{n=1}^{N+1} \epsilon^n F_n(v,\bar{V},\zeta) + \mathrm{O}(\epsilon^{N+2}) \right], \tag{3.17}$$

where the correction terms $F_n(V,\bar{V},\zeta)$ for $1 \le n \le (N+1)$ are periodic mean-zero functions of ζ:

$$F_n(V,\bar{V},\zeta+1) = F_n(V,\bar{V},\zeta), \qquad \int_0^1 F_n(V,\bar{V},\zeta)d\zeta = 0. \tag{3.18}$$

The generating functional $F(v,\bar{V},\zeta)$ defines the near-identical canonical transformation

$$\bar{v} = \frac{\delta F}{\delta v}, \qquad V = \frac{\delta F}{\delta \bar{V}}, \tag{3.19}$$

and takes the Hamiltonian $H(v,\bar{v},\zeta)$ to the form

$$\begin{aligned} H_{\rm new}(V,\bar{V},\zeta) &= H(v(V,\bar{V},\zeta),\bar{v}(V,\bar{V},\zeta),\zeta) + \frac{i}{\epsilon}\frac{\partial F}{\partial \zeta}(v(V,\bar{V},\zeta),\bar{V},\zeta) \\ &= H_N(V,\bar{V}) + \mathrm{O}(\epsilon^{N+1}), \end{aligned} \tag{3.20}$$

where $H_N(V,\bar{V})$ is the Nth-order averaged Hamiltonian

$$H_N(V,\bar{V}) = \sum_{n=0}^{N} \epsilon^n h_n(V,\bar{V}). \tag{3.21}$$

When the remainder term of order of $\mathrm{O}(\epsilon^{N+1})$ is neglected, the new canonical variables $(V,\bar{V})$ solve the averaged Hamiltonian dynamical system:

$$i\frac{\partial V}{\partial z} = \frac{\delta H_N}{\delta \bar{V}}, \qquad -i\frac{\partial \bar{V}}{\partial z} = \frac{\delta H_N}{\delta V}, \tag{3.22}$$

The difference between solutions of the full system (3.8) and the averaged system (3.22) is controlled with the accuracy of $\mathrm{O}(\epsilon^{N+1})$ on the interval $0 \le z \le z_0$.

The Lagrangian functional for the system (3.8) is transformed as follows:

(3.23)
$$L = i\int_{-\infty}^{\infty} \bar{v}\frac{\partial v}{\partial z}dt - H(v,\bar{v},\zeta) = -i\int_{-\infty}^{\infty} V\frac{\partial \bar{V}}{\partial z}dt - H_{\text{new}}(V,\bar{V},\zeta) + i\frac{dF}{dz}(v,\bar{V},\zeta),$$

where

$$\frac{dF}{dz} = \frac{1}{\epsilon}\frac{\partial F}{\partial \zeta} + \int_{-\infty}^{\infty} dt\left(\frac{\partial v}{\partial z}\frac{\delta F}{\delta v} + \frac{\partial \bar{V}}{\partial z}\frac{\delta F}{\delta \bar{V}}\right).$$

If $F(V,\bar{V},\zeta)$ generates $\bar{v}$ and V according to (3.19), then the Hamiltonian $H(v,\bar{v},\zeta)$ transforms according to (3.20). The method of normal form transformations in (3.17)–(3.23) is a formal algorithmic procedure. Still we are able to prove convergence of the first-order averaged theory in a suitable function space; see section 3.2.

The difference between solutions of the averaged integral NLS equation (3.15) and the periodic integral NLS equation (3.11) is defined with the help of the first-order correction $F_1(V,\bar{V},\zeta)$ in (3.17). The first-order correction can be found from (3.9), (3.13), and (3.20):

$$F_1(V,\bar{V},\zeta) = \frac{-i}{2}\left\{|T(\zeta)V|^4 - \int_0^1 |T(\zeta)V|^4\, d\zeta\right\}_{-1}, \tag{3.24}$$

where $\{*\}_{-1}$ stands for the mean-zero antiderivative in ζ defined by (2.2). Expanding the near-canonical transformations (3.17) and (3.19) in powers of ϵ, we define the perturbation series for $v(V,\bar{V},\zeta)$:

$$v = V + \epsilon i\Phi(V,\bar{V},\zeta) + O(\epsilon^2), \qquad \bar{v} = \bar{V} - \epsilon i\overline{\Phi(V,\bar{V},\zeta)} + O(\epsilon^2), \tag{3.25}$$

where $\Phi(V,\bar{V},\zeta)$ is formally computed as

$$\Phi = \left\{T^{-1}(\zeta)\left(|T(\zeta)V|^2\, T(\zeta)V\right) - \int_0^1 T^{-1}(\zeta)\left(|T(\zeta)V|^2\, T(\zeta)V\right)d\zeta\right\}_{-1}. \tag{3.26}$$

In the Fourier form, $\Phi(V,\bar{V},\zeta)$ is expressed explicitly as $\hat{\Phi}_\omega(V,\bar{V},\zeta)$:

$$\hat{\Phi}_\omega = \iint_{-\infty}^{\infty} \{K_\omega(\zeta) - \langle K_\omega\rangle\}_{-1}\, \hat{V}(\omega_1)\hat{V}(\omega_2)\hat{\bar{V}}(\omega_1+\omega_2-\omega)d\omega_1 d\omega_2. \tag{3.27}$$

With the use of correction $\Phi(V,\bar{V},\zeta)$, the first-order correction term $h_1(V,\bar{V})$ of the new averaged Hamiltonian is found in the form

$$h_1(V,\bar{V}) = -i\int_{-\infty}^{\infty} dt\int_0^1 \left(|T(\zeta)V|^2\left(T^{-1}(\zeta)\bar{V}\right)T(\zeta)\Phi(V,\bar{V},\zeta) \ - \text{c.c.}\right)d\zeta, \tag{3.28}$$

where c.c. stands for complex conjugation and we have used the periodicity of $\Phi(V,\bar{V},\zeta)$ in ζ. The first-order correction $h_1(V,\bar{V})$ vanishes in the case of symmetric maps, when $l = 1/2$.

LEMMA 3.2. *When the dispersion map is symmetric, i.e., $l = 1/2$, then $h_1(V,\bar{V}) = 0$ and $\Phi(V,\bar{V},0) = \Phi(V,\bar{V},l) = 0$.*

Proof. If $l = 1/2$, then the mean-zero antiderivative function $d_{-1}(\zeta)$ is even: $d_{-1}(-\zeta) = d_{-1}(\zeta)$. The operator $\hat{T}_\omega(\zeta)$ and the kernel $K_\omega(\zeta)$ in (3.6) and (3.12) are even functions of ζ and can be expanded into the Fourier cosine-series, e.g.,

$$K_\omega(\zeta) = \langle K_\omega \rangle + \sum_{n=1}^{\infty} k_{\omega n} \cos(2\pi n \zeta). \tag{3.29}$$

It follows from (3.27) that the first-order correction $\hat{\Phi}_\omega(V, \bar{V}, \zeta)$ is expanded into the Fourier sine-series:

$$\hat{\Phi}_\omega(V, \bar{V}, \zeta) = \sum_{n=1}^{\infty} \phi_{\omega n}(V, \bar{V}) \sin(2\pi n \zeta). \tag{3.30}$$

The integrand of (3.28) contains only the product of the Fourier cosine- and sine-series, which has zero mean. □

The first-order averaged Hamiltonian is finally defined as $H_1(V, \bar{V}) = h_0(V, \bar{V}) + \epsilon h_1(V, \bar{V})$, where $h_0(V, \bar{V})$ and $h_1(V, \bar{V})$ are given by (3.13) and (3.28).

3.2. Averaging theorems for the first-order averaged integral NLS equation. Here we justify the first-order averaging theory for the periodic integral NLS equation (3.10). In order to shorten notation, we introduce the operator $Q(v, \zeta)$ for the cubic nonlinear term in (3.10):

$$Q(v, \zeta) = Q(v, v, v, \zeta), \qquad Q(u, v, w, \zeta) = T^{-1}(\zeta) \left(T(\zeta) u \, T(\zeta) v \, \overline{T(\zeta) w} \right). \tag{3.31}$$

In the operator form,

$$\Phi(V, \bar{V}, \zeta) = \left\{ Q(V, \zeta) - \langle Q \rangle(V) \right\}_{-1},$$

and the first-order averaged integral NLS equation can be written in the form

$$i \frac{\partial V}{\partial z} + \frac{1}{2} d_0 \frac{\partial^2 V}{\partial t^2} + \langle Q \rangle(V) + \epsilon \langle Q_1 \rangle(V) = 0, \tag{3.32}$$

where

$$\langle Q \rangle(V) = \int_0^1 Q(V, \zeta) d\zeta, \tag{3.33}$$

and

$$\langle Q_1 \rangle(V) = \frac{\delta h_1}{\delta \bar{V}} = i \int_0^1 \left[Q(V, V, \Phi, \zeta) - 2 Q(V, \Phi, V, \zeta) \right] d\zeta. \tag{3.34}$$

First, we list some properties of Q and Φ and formulate a local existence result for the first-order averaged integral NLS equation (3.32).

PROPOSITION 3.3. *Let u, v, w be in $H^s(\mathbb{R})$ $(s \geq 0)$; then the following inequalities hold:*

$$||Q(u, v, w, \zeta)||_{H^s} \leq C_s ||u||_{H^s} ||v||_{H^s} ||w||_{H^s}, \tag{3.35}$$

$$||Q(u, \zeta)||_{H^s} \leq C_s ||u||^3_{H^s}, \tag{3.36}$$

$$||\langle Q(u) \rangle||_{H^s} \leq C_s ||u||^3_{H^s}, \tag{3.37}$$

$$||\Phi(u, \bar{u}, \zeta)|| \leq C_s ||u||^3_{H^s}, \tag{3.38}$$

$$||Q(u, u, \Phi(u, \bar{u}, \zeta), \zeta)|| \leq C_s ||u||^5_{H^s}, \tag{3.39}$$

$$||\langle Q(u, u, \Phi(u, \bar{u}, *), *) \rangle|| \leq C_s ||u||^5_{H^s}. \tag{3.40}$$

Proof. The first inequality is proven using the well-known property of $H^s(\mathbb{R})$ with, e.g., $s \geq 1$,

$$||uv||_{H^s} \leq C_s||u||_{H^s}||v||_{H^s},$$

the isometric properties of $T(\zeta)$,

$$||T(\zeta)u||_{H^s} = ||u||_{H^s},$$

and

$$T(\zeta)Q(u,v,w,\zeta) = T(\zeta)uT(\zeta)v\overline{T(\zeta)w} \Rightarrow ||T(\zeta)Q(u,v,w,\zeta)||_{H^s}$$
$$\leq C||T(\zeta)u||_{H^s}||T(\zeta)v||_{H^s}||T(\zeta)w||_{H^s} \Rightarrow ||Q(u,v,w,\zeta)||_{H^s} \leq C||u||_{H^s}||v||_{H^s}||w||_{H^s}.$$

The remaining inequalities (3.36)–(3.40) are obtained by direct application of the first inequality (3.35). ∎

PROPOSITION 3.4. *Let $V(0) \in H^s(\mathbb{R})$ with $s \geq 1$. Then there exists $z_0 > 0$ such that the first-order averaged equation* (3.32) *has a unique solution $V(z) \in L^\infty([0,z_0], H^s(\mathbb{R}))$.*

Proof. The local existence for (3.32) with $\epsilon = 0$ has been proven in [28] by using the standard application of contraction mapping. In the general case, when $\epsilon \neq 0$, the proof of local existence is similar. First, we rewrite (3.32) in the integral form:

$$V(z) = T_0(z)V(0) + \int_0^Z T_0(z-z')\left(\langle Q\rangle(V(z')) + \epsilon\langle Q_1\rangle(V(z'))\right)dz',$$

where $T_0(z)$ is the operator associated with the fundamental solution of the linear Schrödinger equation:

$$i\frac{\partial V}{\partial z} + \frac{1}{2}d_0\frac{\partial^2 V}{\partial t^2} = 0.$$

Estimating the difference between two solutions, we obtain

$$||V_1(z) - V_2(z)||_{H^s} \leq z_0 C_s(||V_1(0)||_{H^s}, ||V_2(0)||_{H^s}, \epsilon)||V_1(0) - V_2(0)||_{H^s},$$

which is a contraction if z_0 is sufficiently small (uniformly in ϵ). Using the standard energy estimates, we also obtain

$$\frac{\partial}{\partial z}||V(z)||^2_{H^s} \leq C_s(||V(z)||^2_{H^s}, \epsilon)||V(z)||^2_{H^s},$$

where C_s is a smooth function in both variables, thus implying uniqueness. ∎

PROPOSITION 3.5. *Let $V(0) \in H^1(\mathbb{R})$ and $d_0 \neq 0$. Then there exists a global solution $V \in L^\infty([0,\infty), H^1(\mathbb{R}))$ with initial data $V(0)$.*

Proof. For the proof we use first-order averaged Hamiltonian $H_1(V,\bar{V})$, conserved in z. It is shown in section 3.3 that the Hamiltonian $H_1(V,\bar{V})$ is bounded uniformly in $\epsilon \in [0,\epsilon_0]$, provided $||V||_{L^2}$ is fixed. Therefore, since the Hamiltonian is conserved in z, the gradient term must be bounded:

$$\int_{-\infty}^{\infty} |\partial_t V(z)|^2 dt \leq C(||V(0)||_{L^2}, ||\partial_t V(0)||_{L^2}, d_0),$$

which implies that $||V(z)||_{H^1}$ is uniformly bounded, thus proving global existence of solutions. ∎

Remark 3.1. If $d_0 = 0$, then a global solution $V(z)$ still exists in $H^1(\mathbb{R})$, although it is not uniformly bounded.

Before proving convergence of the first-order averaging theory, we reproduce the leading-order averaging theory from [28].

THEOREM 3.6 (see [28]). *Let $V(z) \in L^\infty([0,z_0], H^s(\mathbb{R}))$, where $s \geq 2$, be a solution of the averaged NLS equation* (3.32) *with $\epsilon = 0$ and $v(z)$ be a solution of the full equation* (3.10) *such that $||v(0) - V(0)||_{H^{s-2}} \leq C\epsilon$. Then, for sufficiently small positive $\epsilon < \epsilon_0$ we have $v(z) \in L^\infty([0,z_0], H^{s-2}(\mathbb{R}))$ and the solutions stay close at the distances $0 \leq z \leq z_0$:*

$$\sup_{z\in[0,z_0]} ||v(z) - V(z)||_{H^{s-2}} \leq C\epsilon. \tag{3.41}$$

We prove the analogous theorem for the first-order averaged integral NLS equation (3.32).

THEOREM 3.7. *Let $V(z) \in L^\infty([0,z_0], H^s(\mathbb{R}))$, where $s \geq 4$, be a solution of the first-order averaged integral NLS equation* (3.32) *and $v(z)$ be a solution of the full equation* (3.10) *such that $||v(0) - V(0) - i\epsilon\Phi(V(0), \bar{V}(0), 0)||_{H^{s-4}} \leq C\epsilon$. Then, for sufficiently small positive $\epsilon < \epsilon_0$ we have $v(z) \in L^\infty([0,z_0], H^{s-4}(\mathbb{R}))$ and the solutions are ϵ-close on $0 \leq z \leq z_0$:*

$$\sup_{z\in[0,z_0]} ||v(z) - V(z) - i\epsilon\Phi(V, \bar{V}, \zeta)||_{H^{s-4}} \leq C\epsilon^2. \tag{3.42}$$

Proof. We start with the averaged integral NLS equation (3.32) and use near-identical transformations to transform it to the periodic integral NLS equation (3.10). In the last step we compare the solutions of the transformed and the reduced equations by using Gronwall's inequality. This approach has a technical advantage over the "direct" approach, which starts from the original equation (3.10) and transforms it to the averaged equation (3.32). Indeed, for the periodic equation (3.10), there is no a priori ϵ-independent estimate on the existence interval.

Let us make a transformation $V = v_1 - w_1$ in (3.32), where v_1 is a new variable and w_1 is a small correction. We formally obtain

$$\begin{aligned} &i\frac{\partial v_1}{\partial z} + \frac{1}{2}d_0\frac{\partial^2 v_1}{\partial t^2} + Q(v_1,\zeta) \\ &\qquad = i\frac{\partial w_1}{\partial z} + \frac{1}{2}d_0\frac{\partial^2 w_1}{\partial t^2} + Q(v_1,\zeta) - \langle Q\rangle(V) + \epsilon\langle Q_1\rangle(V). \end{aligned} \tag{3.43}$$

Choosing $w_1 = i\epsilon\Phi(V, \bar{V}, \zeta)$, we obtain

$$i\frac{\partial v_1}{\partial z} + \frac{1}{2}d_0\frac{\partial^2 v_1}{\partial t^2} + Q(v_1,\zeta) = R_1(V,\zeta), \tag{3.44}$$

where

$$\begin{aligned} R_1(V,\zeta) = &-\epsilon\frac{\partial}{\partial z}\Phi(V,\bar{V},\zeta) + \epsilon i\frac{1}{2}d_0\frac{\partial^2}{\partial t^2}\Phi(V,\bar{V},\zeta) \\ &+ Q(v_1,\zeta) - Q(V,\zeta) + \epsilon\langle Q_1\rangle(V). \end{aligned} \tag{3.45}$$

We expand the right-hand side of (3.46) as

$$\begin{aligned} Q(v_1,\zeta) - Q(V,\zeta) &= Q(V + i\epsilon\Phi,\zeta) - Q(V,\zeta) \\ &= -i\epsilon Q(V,V,\Phi,\zeta) + 2i\epsilon Q(V,\Phi,V,\zeta) - \epsilon^2 Q(\Phi,\Phi,V,\zeta) \\ &\quad + 2\epsilon^2 Q(\Phi,v_1,\Phi,\zeta) + i\epsilon^3 Q(\Phi,\zeta). \end{aligned}$$

If $\langle Q_1 \rangle(V)$ is defined by (3.34), then (3.46) transforms to the periodic NLS equation with a mean-zero error mismatch of order $\mathrm{O}(\epsilon)$:

$$\begin{aligned} R_1 = &-\epsilon \frac{\partial}{\partial z}\Phi(V,\bar{V},\zeta) + \epsilon i \frac{1}{2} d_0 \frac{\partial^2}{\partial t^2}\Phi(V,\bar{V},\zeta) \\ &- i\epsilon\{Q(V,V,\Phi,\zeta)\} + 2i\epsilon\{Q(V,\Phi,V,\zeta)\} \\ &- \epsilon^2 Q(\Phi,\Phi,V,\zeta) + 2\epsilon^2 Q(\Phi,v_1,\Phi,\zeta) + i\epsilon^3 Q(\Phi,\zeta), \end{aligned}$$

where $\{Q\}$ stands for the mean-zero periodic part of Q in ζ. By using properties of Q and Φ from Proposition 3.3 and taking into account the loss of two derivatives, we find the estimate $||R_1(V,\zeta)||_{H^{s-2}} \leq C\epsilon$.

Now, we carry out another transformation $v_1 = v_2 - w_2$, where $w_2 = -i\epsilon \{R_1(V,\zeta)\}_{-1}$. The mean-zero antiderivative of $R_1(V,\zeta)$ in ζ satisfies the estimate $||w_2||_{H^{s-2}} \leq C\epsilon^2$. After rearranging the terms we recover the equation

$$i\frac{\partial v_2}{\partial z} + \frac{1}{2} d_0 \frac{\partial^2 v_2}{\partial t^2} + Q(v_2,\zeta) = R_2(V,\zeta), \tag{3.46}$$

where $R_2(V,\zeta)$ has a long expression in powers of ϵ^2 and higher. With the help of Proposition 3.3, we can estimate all terms of R_2 as $||R_2(V,\zeta)||_{H^{s-4}} \leq C\epsilon^2$. Comparing solutions of (3.46) and (3.10), we obtain an equation for the difference $f := v_2 - v$:

$$i\frac{\partial f}{\partial z} + \frac{1}{2} d_0 \frac{\partial^2 f}{\partial t^2} + Q(v_2,\zeta) - Q(v,\zeta) = R_2(V,\zeta). \tag{3.47}$$

The difference in the left-hand side of (3.47) can be estimated as

$$\begin{aligned} ||Q(v_2,\zeta) - Q(v,\zeta)||_{H^{s-2}} &= ||Q(v_2,\zeta) - Q(f - v_2,\zeta)||_{H^{s-2}} \\ &\leq C_s(||f||_{H^{s-2}}, ||v_2||_{H^{s-2}})||f||_{H^{s-2}}. \end{aligned}$$

The growth of f can be estimated by using the standard energy estimates. We differentiate the equation for f $1, 2, \ldots, n$ times, multiply each of them with $\partial_k \overline{f}$ $(k = 1, 2, \ldots, n)$, subtract complex conjugates, and finally take the sum to obtain

$$\frac{\partial}{\partial z}||f||^2_{H^n} \leq C_s(||f||_{H^n}, ||v_2||_{H^n})||f||^2_{H^n} + C||R_2||_{H^n}||f||_{H^n}.$$

In the last inequality, we can take n: $0 \leq n \leq s-4$ (thus, we have to assume $s \geq 4$) and using Gronwall's inequality, we obtain

$$||f(z)||_{H^{s-4}} \leq C_1(e^{C_2 z}\epsilon^2 + ||f(0)||_{H^{s-4}}),$$

which proves (3.42). □

COROLLARY 3.8. *Suppose the dispersion map $d(\zeta)$ is symmetric with equal legs, i.e., $l = 1/2$. If the solutions $V(z)$ and $v(z)$ are close in the sense of $||v(0) - V(0)||_{H^{s-4}} \leq C\epsilon^2$, then for sufficiently small positive $\epsilon < \epsilon_0$ the solutions remain within the quadratic accuracy at the distances $0 \leq z \leq z_0$ at the points $z = k\epsilon$ and $z = (k - \frac{1}{2})\epsilon$, where $k \in \mathbb{Z}_+$:*

$$\sup_{z \in [0,z_0]} ||v(z = k\epsilon) - V(z = k\epsilon)||_{H^{s-4}} \leq C\epsilon^2. \tag{3.48}$$

The quadratic convergence is based on the fact that $h_1(V,\bar{V}) = 0$ for $l = 1/2$ and $\Phi(V,\bar{V},0) = \Phi(V,\bar{V},l) = 0$; see Lemma 3.2. As a result, we have an improved (quadratic) convergence between solutions of the periodic integral NLS equation (3.11) and the integral NLS equation (3.15). It is only the linear convergence between solutions of the full and averaged equations valid at any point z of the dispersion map in the general case $l \neq 1/2$.

3.3. Existence and stability of ground states of the first-order averaged Hamiltonian. The first-order averaged Hamiltonian functional $H_1(V,\bar{V})$ is a constant of motion in the averaged system; therefore its extrema are expected to be stable solutions. Unfortunately, Hamiltonians in such problems are not bounded from either above or below. The way out is to consider a constrained variational problem, since there exists another conserved quantity e defined by (1.11). We show that the obtained Hamiltonian possesses a constrained minimum for the case $d_0 > 0$. The constrained minimum implies stability of a stationary pulse in this case.

Let us consider the following minimization problem:

$$P_E = \inf\left\{ H_1(V,\bar{V}), V \in H^1(\mathbb{R}), \int_{-\infty}^{+\infty} |V|^2 dt = E \right\}. \tag{3.49}$$

First, we show that the Hamiltonian is bounded from below, $P_E > -\infty$, which is a necessary condition for the presence of a smooth minimizer. Note that the Hamiltonian is unbounded from above for $d_0 > 0$ because of the gradient term in (3.13).

PROPOSITION 3.9. *The Hamiltonian functional $H_1(V,\bar{V})$ is uniformly bounded from below if $d_0 \geq 0$ and E is fixed.*

Proof. Since the gradient term is positive, we need only to establish the boundedness of the other two terms. The leading-order term $h_0(V,\bar{V})$ can be bounded by applying Hölder and Strichartz estimates [28]:

$$\int_0^1 \int_{-\infty}^{+\infty} |T(\zeta)V|^4 dt d\zeta = \int_0^1 \int_{-\infty}^{+\infty} |T(\zeta)V||T(\zeta)V|^3 dt d\zeta$$

$$\leq \left(\int_0^1 \int_{-\infty}^{+\infty} |T(\zeta)V|^2 dt d\zeta \right)^{\frac{1}{2}} \left(\int_0^1 \int_{-\infty}^{+\infty} |T(\zeta)V|^6 dt d\zeta \right)^{\frac{1}{2}} \leq E^{\frac{1}{2}} C_S E^{\frac{3}{2}} = C_S E^2,$$

where we have used the isometry of $T(\zeta)$ in $L^2(\mathbb{R})$ as well as the Strichartz inequality:

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} |T(\zeta)V|^6 dt dz \leq C_s^2 E^3.$$

Now we estimate the first-order term $h_1(V,\bar{V})$ as

$$\begin{aligned}
&\left| \int_0^1 \int_{-\infty}^{+\infty} \overline{T(\zeta)V}^2 \, T(\zeta)V \, T(\zeta)\Phi(V,\bar{V},\zeta) \, dt d\zeta \right| \\
&\leq \left(\int_0^1 \int_{-\infty}^{+\infty} |T(\zeta)V|^6 dt d\zeta \right)^{\frac{1}{2}} \left(\int_0^1 \int_{-\infty}^{+\infty} |T(\zeta)\Phi(V,\bar{V},\zeta)|^2 dt d\zeta \right)^{\frac{1}{2}} \\
&\leq C_S E^{\frac{3}{2}} \left(\int_0^1 \int_{-\infty}^{+\infty} |\Phi(V,\bar{V},\zeta)|^2 dt d\zeta \right)^{\frac{1}{2}}.
\end{aligned}$$

The integral of $|\Phi(V,\bar{V},\zeta)|^2$ in ζ can be estimated from the definition (3.26), rewritten as

$$\Phi(V,\bar{V},\zeta) = \int_0^{\zeta} \Psi(\zeta_1,t) d\zeta_1 - \int_0^1 \int_0^{\zeta_2} \Psi(\zeta_1,t) d\zeta_1 d\zeta_2 - \left(\zeta - \frac{1}{2} \right) \int_0^1 \Psi(\zeta_1,t) d\zeta_1,$$

where we used the notation

$$\Psi(\zeta,t) = T^{-1}(\zeta) \left(|T(\zeta)V|^2 T(\zeta)V \right).$$

The product $|\Phi|^2$ contains 10 terms, which can be estimated in a straightforward way using Strichartz estimate. We give an example of how to carry out one of these estimates:

$$\begin{aligned}
&\left|\int_0^1 d\zeta \int_{-\infty}^{\infty} dt \left[\int_0^{\zeta} \Psi(\zeta_1,t)d\zeta_1 \int_0^1 \int_0^{\zeta_3} \overline{\Psi(\zeta_2,t)}d\zeta_2 d\zeta_3\right]\right| \\
&\le \left|\int_0^1 d\zeta \int_{-\infty}^{\infty} dt \left[\int_0^1 |\Psi(\zeta_1,t)|d\zeta_1 \int_0^1 \int_0^1 |\Psi(\zeta_2,t)|d\zeta_2 d\zeta_3\right]\right| \\
&= \left|\int_{-\infty}^{\infty} dt \left[\int_0^1 |\Psi(\zeta_1,t)|d\zeta_1 \int_0^1 |\Psi(\zeta_2,t)|d\zeta_2\right]\right| \\
&= \int_0^1 d\zeta_2 \int_0^1 d\zeta_1 \int_{-\infty}^{\infty} dt|\Psi(\zeta_1,t)||\Psi(\zeta_2,t)| \\
&\le \left(\int_0^1 d\zeta_1 \int_0^1 d\zeta_2 \int_{-\infty}^{\infty} dt|\Psi(\zeta_1,t)|^2\right)^{\frac{1}{2}} \left(\int_0^1 d\zeta_1 \int_0^1 d\zeta_2 \int_{-\infty}^{\infty} dt|\Psi(\zeta_2,t)|^2\right)^{\frac{1}{2}} \\
&\le \int_0^1 \int_{\infty}^{\infty} |\Psi(\zeta,t)|^2 dtd\zeta.
\end{aligned}$$

The last integral is estimated using the definition of $\Psi(\zeta,t)$ and the Strichartz estimate:

$$\begin{aligned}
\int_0^1 \int_{-\infty}^{\infty} |\Psi(\zeta,t)|^2 dtd\zeta &= \int_0^1 \int_{-\infty}^{\infty} \left|T^{-1}(\zeta)\left(T(\zeta)V^2\overline{T(\zeta)V}\right)\right|^2 dtd\zeta \\
&= \int_0^1 \int_{-\infty}^{\infty} |T(\zeta)V|^6 dtd\zeta \le C_S^2 E^3. \qquad (3.50)
\end{aligned}$$

Therefore the term $h_1(V,\bar{V})$ in the Hamiltonian $H_1(V,\bar{V})$ is bounded by $C_S E^{3/2} C_S E^{3/2} = C_S^2 E^3$. □

The next step is to verify the subadditivity condition which is necessary for the construction of a converging minimizing sequence [31]. The subadditivity property holds in the case $\epsilon = 0$ (see [28]). Here we show that it also holds for sufficiently small ϵ.

LEMMA 3.10. *For any $E > 0$ there exist $\epsilon_0 > 0$ (which may depend on E) such that for any $0 < \epsilon < \epsilon_0$ any minimizing sequence V_n possesses a subsequence V_{n_k} satisfying the subadditivity property*

$$P_{E_1+E2} < P_{E_1} + P_{E_2} \text{ provided } E = E_1 + E_2. \qquad (3.51)$$

Proof. The proof is a simple application of a scaling argument, followed by some estimates using smallness of ϵ. Consider a one-parameter family $V^\lambda = \sqrt{\lambda}V$ with $\lambda \in (0,1)$; then

$$E^\lambda = \int_{-\infty}^{\infty} |V^\lambda|^2 dt = \lambda E.$$

Introducing the notation for the Hamiltonian,

$$H_1(V,\bar{V}) = H^{(2)}(V,\bar{V}) - H^{(4)}(V,\bar{V}) + \epsilon H^{(6)}(V,\bar{V}),$$

where $H^{(2,4,6)}(V,\bar{V})$ represent quadratic gradient term, positive quartic term, and the sixth order perturbation term, respectively. The Hamiltonian then scales as follows:

$$H_1^\lambda = \lambda H^{(2)} - \lambda^2 H^{(4)} + \lambda^3 \epsilon H^{(6)}$$

and then

$$H_1^\lambda - \lambda H_1 = (\lambda - \lambda^2) H^{(4)} + (\lambda - \lambda^3)\epsilon H^{(6)} = \lambda(1-\lambda)(H^{(4)} + (1+\lambda)\epsilon H^{(6)}).$$

Note that for $\epsilon = 0$, $H_1^\lambda > \lambda H_1$, which implies the subadditivity $P_{\lambda E} > \lambda P_E$. The latter results in (3.51) for the same E. For sufficiently small ϵ, the condition (3.51) is expected to hold since $H^{(6)}$ is uniformly bounded. Indeed, if we fix $E > 0$, then for $\epsilon = 0$ the infimum is negative, $P_E^0 < 0$, as shown in [28]. For positive ϵ, the infimum cannot change by more than $\epsilon C_S^2 E^3$; therefore $P_E^\epsilon \le P_E^0 + \epsilon C_S^2 E^3$ remains negative.

By definition, for any minimizing sequence we have $H(V_n) \to P_E^\epsilon$ and therefore for sufficiently large $n \ge N$ the quartic term $H^{(4)}$ has to be bounded from below:

$$H^{(4)} \ge |P_E^\epsilon| - \epsilon C_S^2 E^3 - \delta(N) \Rightarrow H^{(4)} \ge |P_E^0| - 2\epsilon C_S^2 E^3 - \delta(N).$$

Then we prove the estimate:

$$\begin{aligned} H^{(4)} + (1+\lambda)\epsilon H^{(6)} &\ge |P_E^0| - 2\epsilon C_S^2 E^3 - \delta(N) - 2\epsilon C_S^2 E^3 \\ &= |P_E^0| - 4\epsilon C_S^2 E^3 - \delta(N) \ge |P_E^0| - 5\epsilon C_S^2 E^3, \end{aligned}$$

where the last step in the inequalities was done by taking N sufficiently large. Therefore, by requiring that

$$5\epsilon C_S^2 E^3 < \frac{1}{2}|P_E^0|$$

we achieve the subadditivity condition for the minimizing sequence. □

We will also use lemma on localization from [28]. The lemma says that finite energy cannot propagate too far in the linear Schrödinger equation if the initial data are sufficiently smooth.

LEMMA 3.11 (see [28]). *Let $V \in H^1(\mathbb{R})$, $T(\zeta)$ be a free Schrödinger propagator and let*

$$\epsilon(\zeta) = \sup_{\xi \in \mathbb{R}} \int_{\xi-1}^{\xi+1} |T(\zeta)V|^2 dt. \tag{3.52}$$

Then the following estimate holds:

$$\epsilon(\zeta) \le 2\epsilon(0) + \sqrt{\epsilon^2(0) + 2C\epsilon(0)\zeta}. \tag{3.53}$$

Now, we are ready to establish the convergence of a minimizing sequence. The two results above make the convergence proof straightforward and very similar to the one with $\epsilon = 0$; see [28]. Therefore, we sketch only the proof of the main result, providing details only when they are different from the case $\epsilon = 0$.

PROPOSITION 3.12. *If $d_0 > 0$ and $0 < \epsilon < \frac{|P_E^0|}{10\, C_S^2 E^3}$, then there exists a minimizer $W \in H^1(\mathbb{R}) \cap C^\infty(\mathbb{R})$ of the constrained minimization problem* (3.49).

Proof. First we observe that any minimizing sequence $V_n \in H^1(\mathbb{R})$ must possess a bounded derivative

$$\int_{-\infty}^{\infty} \left| \frac{\partial V_n}{\partial t} \right|^2 dt < C, \tag{3.54}$$

for otherwise $\{H_1(V_n, \bar{V}_n)\}_{n=0}^{\infty}$ would have an unbounded subsequence (since $H^{(2)}(V_n, \bar{V}_n)$ would dominate over $H^{(4)}(V_n, \bar{V}_n)$ and $H^{(6)}(V_n, \bar{V}_n)$, which are uniformly bounded). Then there exists a weakly converging subsequence in $L^2(\mathbb{R})$. In order to assure the strong convergence $V_n \to W$ with W satisfying the constraint (3.49), we have to show that the sequence is tight (the energy does not escape to infinity[1]). We assume that $n > N$ with N sufficiently large so that the subadditivity condition would hold. Now, we use the concentration-compactness principle [29], which says that there exists a subsequence V_{n_k}, denoted by V_k, for which one of the following statements is true:

1. (convergence) For some sequence $\{t_k\}_{k=0}^{\infty}$ the translated sequence converges to some limit $V_k(t - t_k) \to W$ (as $k \to \infty$) satisfying the constraint (3.49).
2. (vanishing) The following identity is true:
$$\sup_{y \in \mathbb{R}} \int_{y-1}^{y+1} |V_k|^2 dt \to 0 \quad \text{as} \quad k \to \infty.$$
3. (splitting) There exist $E_1, E_2 > 0$ ($E = E_1 + E_2$) such that for any $\epsilon > 0$ one can find two sequences v_k, w_k and $K > 0$ so that for any $k > K$ we have
$$\int_{-\infty}^{\infty} |V_k - (W_k + U_k)|^2 dt < \epsilon,$$
where
$$\int_{-\infty}^{\infty} |W_k|^2 dt = E_1, \qquad \int_{-\infty}^{\infty} |U_k|^2 dt = E_2$$
such that
$$\operatorname{dist}(\operatorname{supp}(W_k), \operatorname{supp}(U_k)) \to \infty.$$

Our goal is to rule out the second and the third possibilities in order to prove convergence of a minimizing sequence. It has been shown in [28] for $\epsilon = 0$ that vanishing implies that $H^{(4)}(V_k, \bar{V}_k) \to 0$. This is in contradiction with the sequence being minimizing as the infimum is negative and $H_1(V_k, \bar{V}_k) \to 0$. The proof that $H^{(4)}(V_k, \bar{V}_k) \to 0$ is based on Cazenave's estimate [30],

$$\int_{-\infty}^{\infty} |V|^4 dt \le C \, ||V||_{H^1}^2 \, \sup_{y \in \mathbb{R}} \int_{y-1}^{y+1} |V|^2 dt, \tag{3.55}$$

and on the lemma on localization (3.52)–(3.53). Combing these estimates in a similar way, we prove that $H^{(6)}(V_k) \to 0$.

We also show that the splitting may not occur. By contradiction, we assume that splitting occurs and show that the sequence is not minimizing by using the subadditivity condition (3.51). The proof is identical to the one in [28] and therefore is omitted here.

Since both the vanishing and splitting scenarios do not occur, the concentration-compactness principle implies that the sequence $V_k \to W$ strongly in $L^2(\mathbb{R})$ [29]. Using the standard argument (see section 3.1 in [28]), we show that $V_k \to W$ strongly in $H^1(\mathbb{R})$. The minimizer weakly satisfies the Euler–Lagrange equation

$$-\mu W + \frac{1}{2} d_0 W''(t) + \langle Q \rangle(W) + \epsilon \langle Q_1 \rangle(W) = 0. \tag{3.56}$$

[1]There is no problem with the local loss of compactness since on any finite interval $I \subset \mathbb{R}$ the space $H^1(I)$ is compactly embedded in $L^2(I)$.

Using the bootstrapping procedure, we show that the solution is smooth. If $W \in H^1(\mathbb{R})$, then $\langle Q \rangle(W) + \epsilon \langle Q_1 \rangle(W) \in H^1(\mathbb{R})$. Due to the presence of the term $d_0 W''(t)$ in (3.56), the solution is extended to function space $W \in H^3(\mathbb{R})$. Continuing this, we obtain that $W \in H^s(\mathbb{R})$ for any $s \geq 1$ and $W \in C^\infty(\mathbb{R})$. □

The minimizer $W(t)$ obtained in Proposition 3.12 defines a stationary pulse solution $V(z,t) = W(t)e^{i\mu z}$ of the first-order averaged integral NLS equation, where $\mu = f_\mu^{-1}(e)$ and $e = f_\mu(\mu)$ is a continuous function. Thus, the existence and stability of a single branch of DM solitons is proved for $d_0 > 0$ in the first-order averaged integral NLS equation (3.32). This completes the proof of Proposition 1.8 for the integral NLS approximation. If the stationary pulse solutions $W(t)$ are computed numerically, all other parameters of DM solitons can be computed for the case $d_0 > 0$, in direct correspondence with Proposition 1.4. Otherwise, the analytical dependencies in (1.19)–(1.21) remain implicitly defined by the averaged integral equation (3.56).

4. Conclusion. We have studied existence and stability of dispersion-managed (DM) solitons for the periodic NLS equation. We defined the DM solitons either as periodic solutions of a low-dimensional system for parameters of a Gaussian pulse or as stationary pulse solutions of the averaged integral NLS equation. In both cases, we have found and analyzed the first-order averaged Hamiltonian. Some open problems appear beyond this analysis and are worth mentioning here.

First, it is a conjecture that DM solitons do not exist as quasi-periodic solutions of the periodic NLS equation (1.5), contrary to the approximating Gaussian pulses. Recent work of Yang and Kath [19] discusses parametric resonances between localized pulses and linear Bloch waves associated with the varying dispersion $d(z)$. Asymptotic and numerical analysis confirmed that the quasi-periodic pulses produce nonlocalized radiation tails, which escape the localized region to infinity [19]. The radiation tail is exponentially small in the limit $\epsilon \to 0$, i.e., it appears beyond any asymptotic expansion in powers of ϵ. In our analysis, all the resonant terms are removed from the leading and first order of the asymptotic series. As a result, the quasi-periodic pulses exist in the averaged integral NLS equation (3.32), at least for $d_0 > 0$.

Second, the first-order constrained Hamiltonian $H_1(V, \bar{V})$ was shown to possess a constrained minimum only for $d_0 > 0$. With the use of the new work by Kunze [20], the constrained minimum can be shown to exist for $d_0 = 0$. However, it is impossible to prove whether or not a local extremum of the averaged Hamiltonian exists for $d_0 < 0$ even in the limit $\epsilon \to 0$. Indeed, the operator $\mu - \frac{1}{2} d_0 \partial_{tt}$ is not positive-definite for $\mu > 0$ and $d_0 < 0$, and a strong resonance occurs between spectra of a localized pulse and linear waves. As a result, the Hamiltonian functional $H_1(V, \bar{V})$ is unbounded from below even for the constrained problem (3.49).

Two branches of Gaussian pulse solutions exist for $d_0 < 0$: one is stable and the other one is unstable in the propagation in z. However, iterations of a numerical method quickly diverge for the branch of unstable Gaussian pulses [13] and slowly diverge for the branch of stable Gaussian pulses [21]. Rigorous analysis of existence or nonexistence of stationary solutions of the problem (3.56) with $d_0 < 0$ is not completed yet.

Finally, the higher-order averaged Hamiltonian can be found and analyzed for the case $d_0 > 0$ in a similar manner. However, the constrained minimization procedure fails already for the second-order Hamiltonian, which has a correction $H^{(8)}(V, \bar{V})$ that contains eight powers of V and $\bar{V}$. Because of such higher-order nonlinearity, the correction $H^{(8)}(V, \bar{V})$ is not be bounded from below by the Strichartz estimate (3.50). Therefore, higher-order averaged equations become less useful for analysis.

Acknowledgments. D.P. thanks G. Biondini, L. Chan, I. Gabitov, W. Kath, and T. Lakoba for their continuous encouragements to extend and clarify the paper [13]. Both authors are grateful to H. Kalisch and M. Weinstein for collaborations and valuable comments.

REFERENCES

[1] T.I. Lakoba and G.P. Agrawal, *Optimization of the average-dispersion range for long-haul dispersion-managed soliton systems,* J. Lightwave Tech., 18 (2000), p. 1504.
[2] M. Nakazawa, H. Kubota, K. Suzuki, E. Yamada, and A. Sahara, *Recent progress in soliton transmission technology,* Chaos, 10 (2000), p. 486.
[3] S. Turitsyn, M.P. Fedoruk, E.G. Shapiro, V.K. Mezentsev, and E.G. Turitsyna, *Novel approaches to numerical modeling of periodic dispersion-managed fiber communication systems,* IEEE J. Quantum Electr., 6 (2000), p. 263.
[4] S.K. Turitsyn and E.G. Shapiro, *Variational approach to the design of optical communication systems with dispersion management,* Opt. Fiber Tech., 4 (1998), p. 151.
[5] V. Cautaerts, A. Maruta, and Y. Kodama, *On the dispersion managed soliton,* Chaos, 10 (2000), p. 515.
[6] J.H. Nijhof, W. Forysiak, and N.J. Doran, *The averaging method for finding exactly periodic dispersion-managed solitons,* IEEE J. Quantum Electr., 6 (2000), p. 330.
[7] A. Berntson, N.J. Doran, W. Forysiak, and J.H.B. Nijhof, *Power dependence of dispersion-managed solitons for anomalous, zero, and normal path-average dispersion,* Opt. Lett., 23 (1998), p. 900.
[8] V.S. Grigoryan and C.R. Menyuk, *Dispersion-managed solitons at normal average dispersion,* Opt. Lett., 23 (1998), p. 609.
[9] J.N. Kutz, P. Holmes, S.G. Evangelides, and J.P. Gordon, *Hamiltonian dynamics of dispersion managed breathers,* J. Opt. Soc. Amer. B, 15 (1998), p. 87.
[10] M. Kunze, *Periodic solutions of a singular Lagrangian system related to dispersion-managed fiber communication devices,* Nonlinear Dynamics and Systems Theory, 1 (2001), p. 159.
[11] S.K. Turitsyn, A.B. Aceves, C.K.R.T. Jones, and V. Zharnitsky, *Average dynamics of the optical soliton in communication lines with dispersion management: Analytical results,* Phys. Rev. E, 58 (1998), p. R48.
[12] S.K. Turitsyn, A.B. Aceves, C.K.R.T. Jones, V. Zharnitsky, and V.K. Mezentsev, *Hamiltonian averaging in soliton-bearing systems with a periodically varying dispersion,* Phys. Rev. E, 59 (1999), p. 3843.
[13] D. Pelinovsky, *Instabilities of dispersion-managed solitons in the normal dispersion regime,* Phys. Rev. E, 62 (2000), p. 4283.
[14] I. Gabitov and S.K. Turitsyn, *Averaged pulse dynamics in a cascaded transmission system with passive dispersion compensation,* Opt. Lett., 21 (1996), p. 327.
[15] M.J. Ablowitz and G. Biondini, *Multiscale pulse dynamics in communication systems with strong dispersion management,* Opt. Lett., 23 (1998), p. 1668.
[16] C. Pare, V. Roy, F. Lesage, P. Mathieu, and P.A. Belanger, *Coupled-field description of zero-average dispersion management,* Phys. Rev. E, 60 (1999), p. 4836.
[17] G. Biondini and S. Chakravarty, *Nonlinear evolution of dispersion-managed return-to-zero pulses,* Opt. Lett., 26 (2001), p. 1761.
[18] V. Zharnitsky, E. Grenier, S. Turitsyn, C.K.R.T. Jones, and J.S. Hesthaven, *Ground states of dispersion-managed nonlinear Schrödinger equation,* Phys. Rev. E, 62 (2000), p. 7358.
[19] T. Yang and W.L. Kath, *Radiation loss of dispersion–managed solitons in optical fibers,* Phys. D, 149 (2001), p. 80.
[20] M. Kunze, *On a Variational Problem with Lack of Compactness Related to the Nonlinear Schrödinger Equation,* preprint, University of Essen, Essen, Germany, 2002.
[21] P.M. Lushnikov, *Dispersion–managed soliton in a strong dispersion map limit,* Opt. Lett., 26 (2001), p. 1535.
[22] T.I. Lakoba and D.J. Kaup, *Hermite–Gaussian expansion for pulse propagation in strongly dispersion–managed fibers,* Phys. Rev. E, 58 (1998), p. 6728.
[23] A.J. Lichtenberg and M.A. Lieberman, *Regular and Stochastic Motion,* Springer–Verlag, New York, 1983.
[24] A.I. Neishtadt, *The separation of motions in systems with rapidly rotating phase,* J. Appl. Math. Mech., 48 (1985), p. 133.

[25] G. Strang, *Accurate partial difference methods. Nonlinear problems,* Numer. Math., 6 (1964), p. 37.
[26] A. Hasegawa and Y. Kodama, *Solitons in Optical Communications,* Oxford University Press, New York, 1995.
[27] I. Gabitov, T. Schafer, and S.K. Turitsyn, Phys. Lett. A, 265 (2000), p. 274.
[28] V. Zharnitsky, E. Grenier, S. Turitsyn, and C.K.R.T. Jones, *Stabilizing effects of dispersion management,* Phys. D, 152/153 (2001), p. 794.
[29] P.-L. Lions, *The concentration-compactness principle in the calculus of variations. The limit case.* II, Rev. Mat. Iberoamericana, 1 (1985), p. 45.
[30] T. Cazenave, *An Introduction to Nonlinear Schrödinger Equations,* Textos de Métodos Matematicos 22, IMUFRJ, Rio de Janeiro, 1989.
[31] M. Struwe, *Variational Methods*, Ergeb. Math. Grenzgeb. 3, Springer-Verlag, Berlin, 1996.

SIAM J. APPL. MATH.
Vol. 63, No. 3, pp. 777–800

UNIFORM ASYMPTOTIC EXPANSION OF THE SQUARE-ROOT HELMHOLTZ OPERATOR AND THE ONE-WAY WAVE PROPAGATOR*

MAARTEN V. DE HOOP† AND A. K. GAUTESEN‡

Abstract. The Bremmer coupling series solution of the wave equation, in generally inhomogeneous media, requires the introduction of pseudodifferential operators. Such operators appear in the diagonalization process of the acoustic system's matrix of partial differential operators upon extracting a principal direction of (one-way) propagation. In this paper, in three dimensions, uniform asymptotic expansions of the Schwartz kernels of these operators are derived. Also, we derive a uniform asymptotic expansion of the one-way propagator appearing in the series. We focus on designing closed-form representations, valid in the high-frequency limit, taking into account *critical scattering-angle* phenomena. The latter phenomena are not dealt with in the standard calculus of pseudodifferential operators. Our expansion is not limited by propagation angle. In principle, the uniform asymptotic expansion of a kernel follows by matching its asymptotic behaviors away and near its diagonal.

Key words. wave field decomposition, Bremmer series, uniform asymptotics

AMS subject classifications. 35L05, 35C10, 47G30, 35C20, 34E20

PII. S0036139902402300

1. Introduction.

Directional wave field decomposition is a tool for analyzing and computing wave propagation in configurations with a special directionality, such as a waveguiding structure. Such a method consists of three main steps: (i) decomposing the field into two constituents, propagating "one-way" upward or downward along a preferred or principal direction, (ii) computing the interaction of the counterpropagating constituents, and (iii) recomposing the constituents into observables at the positions of interest. The Bremmer series [1] then synthesizes the constituents into a full-wave solution. Each term in the series represents a wave constituent that has traveled up and down along the principal direction a number of times equal to its order. Thus we are able to trace waves: evolution is no longer in time but now in the vertical coordinate, vertical being identified with the principal direction. The microlocal analysis of the one-way wave propagator can be found in Treves [2].

Applications of the generalized Bremmer series solution to the wave equation include (i) the identification and elimination of multiple scattered wave constituents and (ii) the formulation of various imaging and inverse scattering procedures in remote sensing. In general, the inverse scattering problem can be decomposed into a coupled inverse "contrast-source" or "reflectivity"–inverse "constituency" problem. With the aid of time-reversal mirrors, each pair of successive terms in the Bremmer series can be exploited to construct the reflectivity (see de Hoop [3]).

The generalized Bremmer series can be viewed as a full-wave extension of the (high-frequency) geometrical ray series representation of the wave field embedded

*Received by the editors February 7, 2002; accepted for publication (in revised form) July 30, 2002; published electronically January 17, 2003.
http://www.siam.org/journals/siap/63-3/40230.html
†Center for Wave Phenomena and Department of Mathematical and Computer Sciences, Colorado School of Mines, Golden, CO 80401-1887 (mdehoop@mines.edu). The research of this author was supported by the Consortium Project at the Center for Wave Phenomena.
‡Department of Mathematics and Ames Laboratory, 136 Wilhelm Hall, Iowa State University, Ames, IA 50011-3020 (gautesen@ameslab.gov).

in the Kirchhoff approximation (see Frazer [4]). The Maslov canonical operator is replaced by a Trotter product (see de Hoop, Le Rousseau, and Biondi [5]). The Bremmer series–Trotter product approach encompasses the microlocal, Kirchhoff–Maslov, representation of the wave field. Extensive lists of references to applications of the generalized Bremmer series in exploration and crustal seismology, ocean acoustics, and integrated optics can be found in Van Stralen, de Hoop, and Blok [6].

De Hoop [1] originally formulated the generalized Bremmer series modeling method in the time-*Laplace* domain. Owing to the fact that the medium can vary in the directions transverse to the preferred direction, pseudodifferential calculus became a natural tool to introduce the up and downgoing Green's functions: pseudodifferential operators appear in the directional (de)composition, in the downward and upward propagation or continuation, and in the interaction (reflection and transmission) between the counterpropagating constituents due to variations in medium properties in the preferred direction. The time-Laplace domain is not amenable to computations, however.

Various approaches have been developed over the years in the time-*Fourier* domain to approximate the operators appearing in the Bremmer series to make numerical computations feasible. An overview of the approaches based on rational (paraxial) expansions of the operator symbols can be found in Van Stralen, de Hoop, and Blok [6]. An overview of approaches based on phase-screen-like approximations of the operator symbols can be found in de Hoop, Le Rousseau, and Wu [7]. With these numerical approaches, however, critical "scattering-angle" phenomena such as the ones associated with rays the tangents to which become horizontal (for example, turning rays) cannot be modeled. With the approach proposed in this paper, this limitation is removed. In particular media, spectral analysis can be employed to find exact time-Fourier representations of mentioned operators (see Fishman, de Hoop, and van Stralen [8]).

In this paper, our goal is to gain analytic insight into the propagation and scattering of waves as described by the generalized Bremmer series—while developing a time-*Fourier* analysis of the constituent operators. We extend earlier results (Fishman, Gautesen, and Sun [9] and de Hoop and Gautesen [10]) in this direction that were derived in two dimensions to three (and higher) dimensions. Instead of using pseudodifferential operators in the time-Laplace domain, we will here employ microlocal and uniform asymptotics techniques combined in the time-Fourier domain. We focus our analysis on the development of a uniform asymptotic expansion of the transversal part of the one-way wave operator kernel (of the square-root Helmholtz type) and the associated one-way wave propagator. For the completion of the Bremmer coupling series we refer the reader to our earlier paper.

The uniform asymptotic expansions also provide the basis for a numerical scheme. Such a scheme would involve the computations of (i) a spatially varying effective index of refraction and (ii) a spatially varying effective "distance" in the transverse directions, and then applying the kernel. The effective index of refraction and the effective metric are computed along the bicharacteristics constrained to the plane spanned by transverse directions.

The outline of this paper is as follows. In the next section a summary of the method of directional decomposition, leading to a coupled system of one-way wave equations is given. In section 3, the medium is decomposed into thin slabs. In each thin slab we introduce a "characteristic" Green's function. In section 4 we introduce representations of the square-root operator and the one-way wave propagator in terms of the characteristic Green's function. The key effort is developing a uniform

asymptotic expansion of the characteristic Green's function. Such an expansion in the absence of transverse caustics is developed in section 5 and in the presence of transverse caustics in section 6. In both cases an "inner" (near-field) and "outer" (far-field) representation is derived upon which a matching procedure in a boundary layer is invoked. The latter synthesizes the uniformly valid expression. Section 7 summarizes the main result of the paper: the uniform asymptotic expansion for the square-root operator and the likewise expansion for the one-way wave propagator in higher dimensions. We conclude with a discussion (section 8).

2. Directional wave field decomposition. For the details on the derivation of the Bremmer coupling series solution of the acoustic wave equation, we refer the reader to de Hoop [1]. Here, we restrict ourselves to a summary of this wave field decomposition method.

Notation, transformations. We consider acoustic waves in a three-dimensional configuration. In this configuration, let p denote the pressure and $(v_{1,2}, v_3) = (v_1, v_2, v_3)$ the particle velocity. We introduce the Fourier transformation with respect to time t as

$$(\mathcal{F}\{p, v_{1,2}, v_3\})(x_{1,2}, x_3, \omega) = \int_{t\in\mathbb{R}_{\geq 0}} \{p, v_{1,2}, v_3\}(x_{1,2}, x_3, t)\ \exp(\mathrm{i}\omega t)\,\mathrm{d}t \tag{2.1}$$

for $\mathrm{Im}\{\omega\} > 0$. Under this transformation, assuming zero initial conditions, we have $\partial_t \to -\mathrm{i}\omega$.

In each subdomain of the configuration where the acoustic properties vary continuously with position, the acoustic wave field $\{p, v_{1,2}, v_3\}$ satisfies the system of partial differential equations

$$\partial_k p - \mathrm{i}\omega\rho\, v_k = f_k, \tag{2.2}$$

$$-\mathrm{i}\omega\kappa\, p + \partial_1 v_1 + \partial_2 v_2 + \partial_3 v_3 = q. \tag{2.3}$$

Here, ρ denotes the volume density of mass, κ the compressibility, q the volume source density of injection rate, and f_k the volume source density of force.

The spatial variation of the wave field along a direction of preference can now be expressed in terms of the variation of the wave field in the direction perpendicular to it. The direction of preference or principal direction is taken (globally) along the x_3-axis (or "vertical" axis) and the remaining ("transverse" or "horizontal") coordinates are denoted by (x_1, x_2) or $x_{1,2}$.

The reduced system of equations. Directional decomposition requires a separate handling of the horizontal or transverse component of the particle velocity. From (2.2) and (2.3) we obtain

$$v_{1,2} = -\mathrm{i}\rho^{-1}\omega^{-1}(\partial_{1,2} p - f_1), \tag{2.4}$$

leaving, upon substitution, the matrix differential equation $(I, J = 1, 2)$

$$(\partial_3\delta_{IJ} - \mathrm{i}\omega A_{IJ})\, F_J = N_I, \quad A_{IJ} = A_{IJ}(x_{1,2}, D_{1,2}; x_3), \quad D_{1,2} \equiv -\frac{\mathrm{i}}{\omega}\partial_{1,2}, \tag{2.5}$$

in which the elements of the acoustic field matrix[1] are given by

$$F_1 = p, \qquad F_2 = v_3, \tag{2.6}$$

[1] Present ocean-bottom seismic acquisition technology allows both p and v_3 to be measured.

the elements of the acoustic system's matrix operator by

$$A_{11} = A_{22} = 0, \tag{2.7}$$

$$A_{12} = \rho, \tag{2.8}$$

$$A_{21} = -D_1(\rho^{-1}D_1\,) - D_2(\rho^{-1}D_2\,) + \kappa, \tag{2.9}$$

and the elements of the notional source matrix by

$$N_1 = f_3, \qquad N_2 = D_1(\rho^{-1}f_1) + D_2(\rho^{-1}f_2) + q. \tag{2.10}$$

It is observed that the right-hand side of (2.4) and A_{IJ} contain the spatial derivatives $D_{1,2}$ with respect to the horizontal coordinates only. In the sequel of the paper it will become clear that $D_{1,2}$ has the interpretation of *horizontal slowness* operator. Further, it is noted that A_{12} is simply a multiplicative operator.

The coupled system of one-way wave equations. To distinguish up and downgoing constituents in the wave field, we shall construct an appropriate linear operator L_{IJ} with

$$F_I = L_{IJ}W_J, \tag{2.11}$$

which, with the aid of the commutation relation ([.,.] denotes the commutator)

$$(\partial_3 L_{IJ}) = [\partial_3, L_{IJ}], \tag{2.12}$$

transforms (2.5) into

$$L_{IJ}\,(\partial_3\delta_{JM} - \mathrm{i}\omega\Lambda_{JM})W_M = -(\partial_3 L_{IJ})\,W_J + N_I. \tag{2.13}$$

Transformation (2.11) should result in the diagonalization of the operator A_{IJ} in the sense that

$$A_{IJ}L_{JM} = L_{IJ}\Lambda_{JM}, \tag{2.14}$$

where Λ_{JM} is a diagonal matrix of operators. We denote L_{IJ} as the composition operator and W_M as the wave column matrix. The expression in parentheses on the left-hand side of (2.13) represents the two so-called *one-way* wave operators. The first term on the right-hand side of (2.13) is representative for the scattering due to variations of the medium properties in the vertical direction. The diffraction due to variations of the medium properties in the horizontal directions is contained in Λ_{JM} and, implicitly, in L_{IJ}. This diffraction comprises the multipathing of characteristics that commonly occurs in geophysical configurations.

To investigate whether solutions of (2.14) exist, we introduce the column matrix operators $L_I^{(\pm)}$ according to

$$L_I^{(+)} = L_{I1}, \qquad L_I^{(-)} = L_{I2}. \tag{2.15}$$

Upon writing the diagonal elements of Λ_{JM} as

$$\Lambda_{11} = \Gamma^{(+)}, \qquad \Lambda_{22} = \Gamma^{(-)}, \tag{2.16}$$

(2.14) decomposes into the two systems of equations

$$A_{IJ}L_J^{(\pm)} = L_I^{(\pm)}\Gamma^{(\pm)}. \tag{2.17}$$

By analogy with the case where the medium is translationally invariant in the horizontal directions, we shall denote $\Gamma^{(\pm)}$ as the *vertical slowness* operators. Notice that the operators $L_1^{(\pm)}$ synthesize the acoustic pressure and that the operators $L_2^{(\pm)}$ synthesize the vertical particle velocity. Through mutual elimination, the equations for $L_1^{(\pm)}$ and $L_2^{(\pm)}$ can be decoupled as follows:

$$A_{12}A_{21}L_1^{(\pm)} = L_1^{(\pm)}\Gamma^{(\pm)}\,\Gamma^{(\pm)}, \tag{2.18}$$

$$A_{21}A_{12}L_2^{(\pm)} = L_2^{(\pm)}\Gamma^{(\pm)}\,\Gamma^{(\pm)}. \tag{2.19}$$

The partial differential operators on the left-hand sides differ from one another in the case where the volume density of mass does vary in the horizontal directions.

To ensure that nontrivial solutions of (2.18)–(2.19) exist, one equation must imply the other. To construct a formal solution, an ansatz is introduced in the form of a commutation relation for one of the components $L_J^{(\pm)}$ that restricts the freedom in the choice for the other component. In the *acoustic-pressure normalization* analogue one assumes that $L_2^{(\pm)}$ can be chosen such that

$$[A_{12}L_2^{(\pm)}, A_{12}A_{21}] = 0. \tag{2.20}$$

In view of (2.19), $\Gamma^{(\pm)}$ must then satisfy

$$A_{12}A_{21} - \Gamma^{(\pm)}\,\Gamma^{(\pm)} = 0. \tag{2.21}$$

The commutation relation for $L_1^{(\pm)}$ follows as $[L_1^{(\pm)}, A_{12}A_{21}] = 0$ and a possible solution of (2.17) is

$$L_2^{(\pm)} = A_{12}^{-1}\Gamma^{(\pm)}, \qquad L_1^{(\pm)} = I. \tag{2.22}$$

Since $L_2^{(\pm)}$ as given by (2.22) satisfies (2.20), the ansatz is justified. The solutions of (2.21) are written as

$$\Gamma^{(+)} = -\Gamma^{(-)} = \Gamma = \mathrm{A}^{1/2} \ \text{ with } \ \mathrm{A} = A_{12}A_{21}. \tag{2.23}$$

Thus, the *composition* operator becomes

$$L = \begin{pmatrix} I & I \\ A_{12}^{-1}\Gamma & -A_{12}^{-1}\Gamma \end{pmatrix}. \tag{2.24}$$

Note that we have decomposed the pressure field according to

$$F_1 = F_1^{(+)} + F_1^{(-)} \ \text{ with } \ F_1^{(+)} = W_1, \quad F_1^{(-)} = W_2.$$

In terms of the inverse vertical slowness operator, $\Gamma^{-1} = \mathrm{A}^{-1/2}$, the *decomposition* operator then follows as

$$L^{-1} = \tfrac{1}{2}\begin{pmatrix} I & \Gamma^{-1}A_{12} \\ I & -\Gamma^{-1}A_{12} \end{pmatrix}. \tag{2.25}$$

Using the decomposition operator, (2.13) transforms into

$$(\partial_3\delta_{IM} - \mathrm{i}\omega\Lambda_{IM})\,W_M = -(L^{-1})_{IM}(\partial_3 L_{MJ})W_J + (L^{-1})_{IM}N_M, \tag{2.26}$$

which can be interpreted as a coupled system of one-way wave equations. The propagation is captured by the left-hand side. The coupling between the counter-propagating components, W_1 and W_2, is apparent in the first source-like term on the right-hand side. The waves are excited by the second term on the right-hand side. We have

$$-L^{-1}(\partial_3 L) = \begin{pmatrix} T & R \\ R & T \end{pmatrix}, \tag{2.27}$$

in which T and R represent the *transmission* and *reflection* operators, respectively: let $Y = A_{12}^{-1}\Gamma$ denote the *admittance* operator; then

$$R = -T = \tfrac{1}{2}Y^{-1}\,(\partial_3 Y). \tag{2.28}$$

The two-way Helmholtz equation. Suppose that the medium does *not* vary with x_3. Eliminating F_2 or v_3 from (2.5) then leads to the second-order equation for the pressure,

$$[\partial_3^2 + \omega^2 \mathrm{A}(x_{1,2}, D_{1,2})]\, F_1 = \mathrm{i}\omega\,\rho N_2 + \partial_3 N_1, \tag{2.29}$$

the *two-way* Helmholtz equation, where A is given by (2.23).

3. Decomposition of the configuration into thin slabs. We will now decompose the *medium* into (thin) slabs. Each slab in our three-dimensional configuration is assumed to be *invariant* in the direction of preference, x_3: the compressibility, κ, may vary in the transverse directions, whereas the density is assumed to be constant all together. However, the medium may vary from slab to slab, and hence the vertical coordinate x_3 becomes a parameter that identifies the slab in our further analysis.

The characteristic operator. As mentioned, in our thin-slab analysis, we will consider the following medium profile:

$$\rho = \text{const.}, \tag{3.1}$$

$$\kappa(x_{1,2}) = \kappa_0\, n^2(x_{1,2}); \tag{3.2}$$

thus, setting $\kappa_0 = \rho^{-1}c_0^{-2}$, the wave speed follows from

$$c^{-2}(x_{1,2}) = c_0^{-2} n^2(x_{1,2}),$$

where n denotes the index of refraction. The operator in (2.23) is then given by

$$\mathrm{A}(x_{1,2}, D_{1,2}) = -D_1^2 - D_2^2 + c_0^{-2} n^2(x_{1,2}). \tag{3.3}$$

We will denote A as the transverse Helmholtz or *characteristic* operator.

Factorization, Green's functions. We introduce the well-known Helmholtz equation and "characteristic" Green's function as (cf. (2.29))

$$\begin{aligned}[\partial_3^2 + \omega^2 \mathrm{A}(x_{1,2}, D_{1,2})]\, G(x_{1,2}, x_3 - x_3'; x_{1,2}') \\ = -\delta(x_1 - x_1')\delta(x_2 - x_2')\delta(x_3 - x_3').\end{aligned} \tag{3.4}$$

The vertical slowness operators $\Gamma^{(\pm)}$ factorize the Helmholtz operator (cf. (2.23)):

$$\partial_3^2 + \omega^2 \mathrm{A}(x_{1,2}, D_{1,2}) = [\partial_3 - \mathrm{i}\omega\,\Gamma^{(+)}(x_{1,2}, D_{1,2})]\,[\partial_3 - \mathrm{i}\omega\,\Gamma^{(-)}(x_{1,2}, D_{1,2})]. \tag{3.5}$$

The one-way Green's functions $\mathcal{G}^{(\pm)}$ associated with the two factors satisfy

$$\begin{aligned}[\partial_3 - \mathrm{i}\omega\,\Gamma^{(\pm)}(x_{1,2}, D_{1,2})]\, \mathcal{G}^{(\pm)}(x_{1,2}, x_3 - x_3'; x_{1,2}') \\ = \delta(x_1 - x_1')\delta(x_2 - x_2')\delta(x_3 - x_3').\end{aligned} \tag{3.6}$$

Vertical slowness as phase variable. Note that the Fourier representation of the causal Green's function G yields

$$G(x_{1,2}, x_3 - x_3'; x_{1,2}') = \frac{\omega}{2\pi c_0} \int_{\zeta\in\mathcal{Z}} \tilde{G}(x_{1,2}, x_{1,2}'; \zeta)\, \exp[\mathrm{i}\, \underbrace{(\omega/c_0)}_{k_0}\, |x_3 - x_3'|\zeta]\, \mathrm{d}\zeta. \tag{3.7}$$

Here, $\mathcal{Z}$ follows the real axis in the complex ζ-plane, below it for negative real parts and above it for positive real parts. Since

$$\omega^2 \mathrm{A}(x_{1,2}, D_{1,2}) = \partial_1^2 + \partial_2^2 + (\omega/c_0)^2\, n^2(x_{1,2}), \tag{3.8}$$

$\tilde{G}$ satisfies (cf. (3.4))

$$[\partial_1^2 + \partial_2^2 + (\omega/c_0)^2\, (n^2(x_{1,2}) - \zeta^2)]\, \tilde{G}(x_{1,2}, x_{1,2}'; \zeta) = -\delta(x_1 - x_1')\delta(x_2 - x_2'), \tag{3.9}$$

or, more formally,

$$-\omega^2[\mathrm{A}(x_{1,2}, D_{1,2}) - c_0^{-2}\zeta^2]\, \tilde{G}(x_{1,2}, x_{1,2}'; \zeta) = \delta(x_1 - x_1')\delta(x_2 - x_2'). \tag{3.10}$$

We can deform contour $\mathcal{Z}$ to a contour $\mathcal{Z}'$, say, such that the distance from a zero crossing of $n^2(x_{1,2}) - \zeta^2$ remains finite.

Observe the symmetry $\tilde{G}(x_{1,2}, x_{1,2}'; -\zeta) = \tilde{G}(x_{1,2}, x_{1,2}'; \zeta)$. Hidden inside the integral is a cut-off function in accordance with the microlocal representation of G.

4. Kernel representations in terms of the characteristic Green's function.

The one-way propagator. Using the image principle, we can express the one-way Green's functions in terms of the Green's function of the second-order Helmholtz equation,

$$\mathcal{G}^{(+)}(x_{1,2}, x_3 - x_3'; x_{1,2}') + \mathcal{G}^{(-)}(x_{1,2}, x_3 - x_3'; x_{1,2}') = -2\, \partial_3 G(x_{1,2}, x_3 - x_3'; x_{1,2}'). \tag{4.1}$$

Hence, for $x_3 > x_3'$,

$$\mathcal{G}^{(+)}(x_{1,2}, x_3 - x_3'; x_{1,2}') = -2\, \partial_3 G(x_{1,2}, x_3 - x_3'; x_{1,2}'). \tag{4.2}$$

In fact, $\mathcal{G} \equiv \mathcal{G}^{(+)}$ is the kernel of the (upward) one-way wave propagator. In view of (4.2) this kernel satisfies the property

$$\partial_3^{2j}\mathcal{G} = [-\omega^2 \mathrm{A}(x_{1,2}, D_{1,2})]^j \mathcal{G}, \qquad j = 1, 2, \ldots, \tag{4.3}$$

for $x_3 > x_3'$. We will pay special attention to the so-called thin-slab expansion of $\mathcal{G}$.

The vertical slowness or square-root operator. The vertical slowness or square-root operator Γ (see (2.23)) acts on the wave field as

$$(\Gamma\{W_1, W_2\})(x_{1,2}) = \int_{x_{1,2}'\in\mathbb{R}^2} \mathcal{C}(x_{1,2}, x_{1,2}')\, \{W_1, W_2\}(x_{1,2}')\, \mathrm{d}x_1'\mathrm{d}x_2', \tag{4.4}$$

where $\mathcal{C}$ denotes a Schwartz kernel. From this operator representation, we extract the left vertical slowness symbol through the Fourier transformation

$$\gamma(x_{1,2}, p_{1,2}) = \int_{x_{1,2}'\in\mathbb{R}^2} \mathcal{C}(x_{1,2}, x_{1,2}')\, \exp[-\mathrm{i}\omega(x_\sigma - x_\sigma')p_\sigma]\, \mathrm{d}x_1'\mathrm{d}x_2', \tag{4.5}$$

where the summation convention has been invoked for $\sigma \in \{1,2\}$. The left symbols of the horizontal slowness operators $D_{1,2}$ appear to be simply $p_{1,2}$. The relation between the left vertical slowness symbol and the horizontal slowness symbol constitutes the generalized slowness surface.

We will now focus on finding integral representations for the Schwartz kernel. First, note that the Schwartz kernel can be expressed in terms of the one-way Green's function,

$$\mathcal{C}^{(+)}(x_{1,2}, x'_{1,2}; x'_3) = -\lim_{x_3 \downarrow x'_3} \frac{\mathrm{i}}{\omega} \partial_3 \mathcal{G}^{(+)}(x_{1,2}, x_3 - x'_3; x'_{1,2}), \tag{4.6}$$

$$\mathcal{C}^{(-)}(x_{1,2}, x'_{1,2}; x'_3) = -\lim_{x_3 \uparrow x'_3} \frac{\mathrm{i}}{\omega} \partial_3 \mathcal{G}^{(-)}(x_{1,2}, x_3 - x'_3; x'_{1,2}). \tag{4.7}$$

With (4.2) we find that

$$\mathcal{C}(x_{1,2}, x'_{1,2}; x'_3) = -\lim_{x_3 \downarrow x'_3} \frac{2}{\mathrm{i}\omega}\, \partial_3^2 G(x_{1,2}, x_3 - x'_3; x'_{1,2}). \tag{4.8}$$

Note that $\mathcal{C}$ is dependent on x'_3 through the index of refraction. We will suppress this dependence in our notation.

The inverse vertical slowness operator. The inverse or reciprocal vertical slowness operator admits the kernel identification

$$\mathcal{A}_{-1/2}(x_{1,2}, x'_{1,2}) = -2\mathrm{i}\omega\, G(x_{1,2}, 0; x'_{1,2}). \tag{4.9}$$

From the inverse vertical slowness operator, the higher fractional powers of the characteristic operator can be obtained, viz., through the composition

$$\mathcal{A}^{j-1/2} = \mathrm{A}^j \mathcal{A}_{-1/2}. \tag{4.10}$$

5. Uniform asymptotic expansion of the characteristic Green's function: The absence of caustics.

The inner solution. The inner region is determined by the condition

$$\|(x_1 - x'_1, x_2 - x'_2)\| = \mathcal{O}(k_0^{-1})$$

and corresponds to the behavior of the kernels near their diagonals. The inner region is so close to the "source" at $x'_{1,2}$ that caustics have not (yet) formed.

We reconsider (3.4),

$$[\partial_k \partial_k + k_0^2 n^2(x_{1,2})]\, G(x_{1,2}, x_3 - x'_3; x'_{1,2}) = -\delta(x_1 - x'_1)\delta(x_2 - x'_2)\delta(x_3 - x'_3)$$

and introduce the relative coordinate

$$y_j = x_j - x'_j, \qquad j \in \{1,2,3\}.$$

We expand the index of refraction about (x'_1, x'_2) according to

$$\begin{aligned} n^2(y_{1,2} + x'_{1,2}) &= n^2(x'_{1,2}) + 2n(x'_{1,2})\,([y_1\partial_1 + y_2\partial_2]n)(x'_{1,2}) \\ &\quad + n(x'_{1,2})\,[y_1\partial_1 + y_2\partial_2]^2 n(x'_{1,2}) + (([y_1\partial_1 + y_2\partial_2]n)(x'_{1,2}))^2 + \ldots, \end{aligned} \tag{5.1}$$

where we differentiate $(\partial_{1,2})$ with respect to $x'_{1,2}$ while the argument of n and its derivatives is $x'_{1,2}$. We invoke the expansion of the Green's function in terms of $y_{1,2}$,

$$G = G_0 + G_1 + G_2 + \dots, \tag{5.2}$$

where the subscript indicates the order in $y = (y_1^2 + y_2^2 + y_3^2)^{1/2}$. Then

$$\begin{aligned}
&[\partial_k\partial_k + k_0^2 n^2(x'_{1,2})]\, G_0(x_{1,2}, x_3 - x'_3; x'_{1,2}) \\
&\qquad = -\delta(x_1 - x'_1)\delta(x_2 - x'_2)\delta(x_3 - x'_3), \qquad (5.3)\\
&[\partial_k\partial_k + k_0^2 n^2(x'_{1,2})]\, G_1(x_{1,2}, x_3 - x'_3; x'_{1,2}) \\
&\qquad = -2k_0^2 n(x'_{1,2})\,([y_1\partial_1 + y_2\partial_2]n)(x'_{1,2})\; G_0(x_{1,2}, x_3 - x'_3; x'_{1,2}), \qquad (5.4)\\
&[\partial_k\partial_k + k_0^2 n^2(x'_{1,2})]\, G_2(x_{1,2}, x_3 - x'_3; x'_{1,2}) \\
&\qquad = -2k_0^2 n(x'_{1,2})\,([y_1\partial_1 + y_2\partial_2]n)(x'_{1,2})\; G_1(x_{1,2}, x_3 - x'_3; x'_{1,2}) \qquad (5.5)\\
&\qquad\quad -k_0^2\left[n(x'_{1,2})\,[y_1\partial_1 + y_2\partial_2]^2 n(x'_{1,2})\right. \\
&\qquad\quad \left. +(([y_1\partial_1 + y_2\partial_2]n)(x'_{1,2}))^2\right]\; G_0(x_{1,2}, x_3 - x'_3; x'_{1,2}),
\end{aligned}$$

etc., with solutions obtained recursively as

$$\begin{aligned}
(5.6)\quad & G_0(x_{1,2}, x_3 - x'_3; x'_{1,2}) = \frac{\exp[\mathrm{i}k_0 n\, y]}{4\pi y}, \\
(5.7)\quad & G_1(x_{1,2}, x_3 - x'_3; x'_{1,2}) = \tfrac{1}{2}\mathrm{i}k_0 y\,([y_1\partial_1 + y_2\partial_2]n)\; G_0(x_{1,2}, x_3 - x'_3; x'_{1,2}), \\
& G_2(x_{1,2}, x_3 - x'_3; x'_{1,2}) = \frac{\mathrm{i}k_0 y}{24n}\left\{4n[y_1\partial_1 + y_2\partial_2]^2 n\right. \\
&\qquad +(3\mathrm{i}k_0 n y + 1)([y_1\partial_1 + y_2\partial_2]n)^2 - y^2[(\partial_1 n)^2 + (\partial_2 n)^2] \\
&\qquad \left. +\frac{(\partial_1 n)^2 + (\partial_2 n)^2 - 2n[(\partial_1)^2 + (\partial_2)^2]n}{(\mathrm{i}k_0 n)^2}(\mathrm{i}k_0 n y - 1)\right\} G_0. \qquad (5.8)
\end{aligned}$$

Inner expansion in midpoint coordinates. In the spirit of the Weyl calculus of kernel symbols (see [11, 21.6.5]), we can improve the above result by introducing the midpoint coordinates

$$\bar{x}_j = \tfrac{1}{2}(x_j + x'_j), \qquad j \in \{1, 2, 3\},$$

and re-expand the exponential according to

$$\begin{aligned}
&\exp[\mathrm{i}k_0 n(x'_{1,2})\, y] = \exp[\mathrm{i}k_0 n(\bar{x}_{1,2})\, y]\;\left\{1 - \tfrac{1}{2}\mathrm{i}k_0 y[y_1\partial_1 + y_2\partial_2]n\right. \\
&\qquad \left. + \tfrac{1}{8}\mathrm{i}k_0 y[y_1\partial_1 + y_2\partial_2]^2 n - \tfrac{1}{8}k_0^2 y^2([y_1\partial_1 + y_2\partial_2]n)^2 + \dots\right\}, \qquad (5.9)
\end{aligned}$$

where the argument of n is now $\bar{x}_{1,2}$. The expansion for G (cf. (5.2) and (5.6)–(5.8)) can then be rewritten as

$$\begin{aligned}
G = \frac{\exp[\mathrm{i}k_0 n\, y]}{4\pi y}\Bigg\{1 + \frac{\mathrm{i}k_0 y}{24n}\Bigg[\, & n[y_1\partial_1 + y_2\partial_2]^2 n + ([y_1\partial_1 + y_2\partial_2]n)^2 \\
& - y^2[(\partial_1 n)^2 + (\partial_2 n)^2] \\
& + \frac{(\partial_1 n)^2 + (\partial_2 n)^2 - 2n[(\partial_1)^2 + (\partial_2)^2]n}{(\mathrm{i}k_0 n)^2}(\mathrm{i}k_0 n y - 1)\Bigg] + \dots\Bigg\}. \qquad (5.10)
\end{aligned}$$

Note that the odd order terms (up to this order, G_1) have disappeared. The expansion above has an improved error estimate, here $\mathcal{O}(y^4)$: We now have (cf. (5.2)) $G \simeq G_0 + G_2$ and upon substitution it follows that

$$[\partial_k \partial_k + k_0^2 n^2(x_{1,2})]\, G = G_0(x_{1,2}, x_3 - x_3'; x_{1,2}')\ k_0^2 \mathcal{O}(y^4). \tag{5.11}$$

The outer solution. The outer region is determined by the condition

$$\|(x_1 - x_1', x_2 - x_2')\| = \mathcal{O}(1)$$

and corresponds to the behavior of the kernels away from their diagonals.

We reconsider (3.9),

$$[\partial_1^2 + \partial_2^2 + k_0^2(n^2(x_{1,2}) - \zeta^2)]\, \tilde{G}(x_{1,2}, x_{1,2}'; \zeta) = -\delta(x_1 - x_1')\delta(x_2 - x_2')$$

and we introduce the representation

$$\tilde{G}(x_{1,2}, x_{1,2}'; \zeta) = \mathsf{C}\, \exp(\mathrm{i}k_0 \psi), \tag{5.12}$$

where C is a yet-to-be-determined constant. We expand ψ into phase and amplitude contributions,

$$\psi = \phi + \underbrace{\frac{1}{\mathrm{i}k_0}\phi_1 + \frac{1}{(\mathrm{i}k_0)^2}\phi_2 + \dots}_{\text{amplitude}}. \tag{5.13}$$

Substituting this expansion into the partial differential equation, and collecting equal powers of $(\mathrm{i}k_0)$, results in the eikonal equation

$$p^2 + q^2 + \zeta^2 - n^2(x_{1,2}) = 0, \tag{5.14}$$

for the leading order; here $p \equiv \partial_{x_1}\phi$ and $q \equiv \partial_{x_2}\phi$. The next order terms yield the equation

$$2p\,(\partial_{x_1}\phi_1) + 2q\,(\partial_{x_2}\phi_1) + \partial_{x_1} p + \partial_{x_2} q = 0, \tag{5.15}$$

whereas the final order that we will account for implies the equation

$$2p\,(\partial_{x_1}\phi_2) + 2q\,(\partial_{x_2}\phi_2) + \partial_{x_1}^2\phi_1 + \partial_{x_2}^2\phi_1 + (\partial_{x_1}\phi_1)^2 + (\partial_{x_2}\phi_1)^2 = 0. \tag{5.16}$$

Amplitude expansion. It is convenient to remove the singularities from ϕ_1 and ϕ_2. This is accomplished by the change of functions,

$$\phi_1 = -\tfrac{1}{2}\log\phi + \psi_1, \tag{5.17}$$

$$\phi_2 = \tfrac{1}{8}\phi^{-1} + \psi_2. \tag{5.18}$$

With this change, (5.15)–(5.16) take the form

$$2p\,(\partial_{x_1}\psi_1) + 2q\,(\partial_{x_2}\psi_1) + \phi\,[\partial_{x_1}^2 + \partial_{x_2}^2]\log\phi = 0, \tag{5.19}$$

$$2p\,(\partial_{x_1}\psi_2) + 2q\,(\partial_{x_2}\psi_2) + \partial_{x_1}^2\psi_1 + \partial_{x_2}^2\psi_1 + (\partial_{x_1}\psi_1)^2 + (\partial_{x_2}\psi_1)^2 = 0, \tag{5.20}$$

supplemented with the initial conditions $\psi_1 = \psi_2 = 0$ at $x_{1,2} = x_{1,2}'$.

Expansion in $z = \mathcal{O}(k_0^{-1})$. We now make the assumption that the propagation distance satisfies

$$k_0 \underbrace{|x_3 - x_3'|}_{y_3 = z} = \mathcal{O}(1). \tag{5.21}$$

Thus we guarantee that the stationary point (where $\partial_\zeta \phi = 0$) of the integral representation (3.7) remains at $\zeta = 0$, and that

$$|\exp[\mathrm{i}k_0 |x_3 - x_3'|\zeta]| = \mathcal{O}(1). \tag{5.22}$$

We then expand the relevant functions about $\zeta = 0$, i.e.,

$$\phi = I_0 - \tfrac{1}{2}\zeta^2 I_1 - \tfrac{1}{8}\zeta^4 I_2 + \dots, \tag{5.23}$$

$$\psi_1 = \psi_{10} + \zeta^2 \psi_{11} + \dots, \tag{5.24}$$

$$\psi_2 = \psi_{20} + \dots, \tag{5.25}$$

where I_0, I_1, I_2 and $\psi_{10}, \psi_{11}, \psi_{20}$ are independent of ζ.

The phase function. Invoking expansion (5.23) into (5.14), the equations determining the phase function become

$$P^2 + Q^2 - n^2(x_{1,2}) = 0, \tag{5.26}$$

$$P\,(\partial_{x_1} I_1) + Q\,(\partial_{x_2} I_1) - 1 = 0, \tag{5.27}$$

$$P\,(\partial_{x_1} I_2) + Q\,(\partial_{x_2} I_2) - (\partial_{x_1} I_1)^2 - (\partial_{x_2} I_1)^2 = 0, \tag{5.28}$$

where $P = \partial_{x_1} I_0$ and $Q = \partial_{x_2} I_0$. With eikonal equation (5.26) is associated the Hamilton system

$$\begin{aligned} \frac{\mathrm{d}x_1}{\mathrm{d}\mu} &= P, \qquad \frac{\mathrm{d}P}{\mathrm{d}\mu} = (\partial_1 M)(x_{1,2}), \\ \frac{\mathrm{d}x_2}{\mathrm{d}\mu} &= Q, \qquad \frac{\mathrm{d}Q}{\mathrm{d}\mu} = (\partial_2 M)(x_{1,2}), \end{aligned} \tag{5.29}$$

where $M = \frac{1}{2}n^2$, supplemented by the initial conditions

$$(x_1, x_2)|_0 = (x_1', x_2'), \quad (P, Q)|_0 = (\alpha_1, \alpha_2), \quad \alpha_1^2 + \alpha_2^2 = n^2(x_{1,2}'). \tag{5.30}$$

The additional equations (5.27)–(5.28) comply with the initial conditions at $\mu = 0$: $I_j = 0$, $j = 0, 1, 2, \dots$.

In the Hamilton system (5.29), we expand the right-hand sides into a Taylor series about the "source" coordinates, $x_{1,2}'$:

$$\begin{aligned} \frac{\mathrm{d}P}{\mathrm{d}\mu} &= (\partial_1 M)(x_{1,2}') + y_1 \partial_1 (\partial_1 M)(x_{1,2}') + y_2 \partial_2 (\partial_1 M)(x_{1,2}'), \\ \frac{\mathrm{d}Q}{\mathrm{d}\mu} &= (\partial_2 M)(x_{1,2}') + y_1 \partial_1 (\partial_2 M)(x_{1,2}') + y_2 \partial_2 (\partial_2 M)(x_{1,2}'). \end{aligned} \tag{5.31}$$

We then evaluate the solutions to the Hamilton (see (5.29)–(5.31)) and eikonal (see (5.26)) equations for small values of μ. The parametric representation of the Hamil-

tonian flow follows as

$$
\begin{aligned}
y_1 &= P|_0\mu + \tfrac{1}{2}P_1\mu^2 + \tfrac{1}{3}P_2\mu^3 + \dots,\\
P &= \underbrace{\alpha_1}_{P|_0} + \underbrace{(\partial_1 M)}_{P_1}\,\mu + \tfrac{1}{2}\,\underbrace{[\alpha_1\partial_1 + \alpha_2\partial_2](\partial_1 M)}_{P_2}\,\mu^2 + \dots,\\
y_2 &= Q|_0\mu + \tfrac{1}{2}Q_1\mu^2 + \tfrac{1}{3}Q_2\mu^3 + \dots,\\
Q &= \underbrace{\alpha_2}_{Q|_0} + \underbrace{(\partial_2 M)}_{Q_1}\,\mu + \tfrac{1}{2}\,\underbrace{[\alpha_1\partial_1 + \alpha_2\partial_2](\partial_2 M)}_{Q_2}\,\mu^2 + \dots;
\end{aligned}
\tag{5.32}
$$

in these equations we differentiate $(\partial_{1,2})$ with respect to $x'_{1,2}$ while the argument of $(\partial_{1,2}M)$ is $x'_{1,2}$. For the purpose of the uniform matching, we will re-expand the solution about the transverse midpoint coordinates $\bar{x}_{1,2}$ and give results as needed later.

Solving system (5.32) for μ, α_1, α_2 in terms of y_1, y_2, yields

$$
\begin{aligned}
\mu = \frac{r_2}{n}\Bigg(1 - \frac{1}{2}\frac{[y_1\partial_1 + y_2\partial_2]n}{n} + \frac{1}{3}\left(\frac{[y_1\partial_1 + y_2\partial_2]n}{n}\right)^2\\
+ \frac{1}{8}\left(\frac{[y_1^\perp\partial_1 + y_2^\perp\partial_2]n}{n}\right)^2 - \frac{1}{6}\frac{[y_1\partial_1 + y_2\partial_2]^2 n}{n}\Bigg),
\end{aligned}
\tag{5.33}
$$

$$
\alpha_{1,2} = \frac{n}{r_2}\left(y_{1,2}(1 - \tfrac{1}{2}a_1^2) + y_{1,2}^\perp(a_1 + a_2)\right),
\tag{5.34}
$$

where the argument of n is $x'_{1,2}$,

$$
r_2 = (y_1^2 + y_2^2)^{1/2}
\tag{5.35}
$$

and

$$
y_1^\perp = -y_2, \quad y_2^\perp = y_1,
\tag{5.36}
$$

while

$$
\begin{aligned}
a_1 &= -\frac{[y_1^\perp\partial_1 + y_2^\perp\partial_2]n}{2n},\\
a_2 &= \frac{1}{12}\Bigg(\frac{([y_1^\perp\partial_1 + y_2^\perp\partial_2]n)([y_1\partial_1 + y_2\partial_2]n)}{n^2}\\
&\qquad -2\frac{[y_1^\perp\partial_1 + y_2^\perp\partial_2][y_1\partial_1 + y_2\partial_2]n}{n}\Bigg);
\end{aligned}
\tag{5.37}
$$

note that a_1 and a_2 are of first and second order in y, respectively.

With the aid of relation

$$
P\,\partial_{x_1} I_j + Q\,\partial_{x_2} I_j = \frac{\mathrm{d}I_j}{\mathrm{d}\mu}
$$

valid along a characteristic or ray, (5.26)–(5.28) take the form

$$\frac{\mathrm{d}I_0}{\mathrm{d}\mu} = n^2(x_{1,2}), \tag{5.38}$$

$$\frac{\mathrm{d}I_1}{\mathrm{d}\mu} = 1, \tag{5.39}$$

$$\frac{\mathrm{d}I_2}{\mathrm{d}\mu} = (\partial_{x_1} I_1)^2 + (\partial_{x_2} I_1)^2. \tag{5.40}$$

Explicit expansions of I_j near the "source" (μ small or, equivalently, r_2 small) are readily obtained from (5.38)–(5.40) using system (5.32) and solutions (5.33)–(5.34). (Basically, such a procedure encompasses an expansion of (the argument of) n^2 about the fixed initial point $(x'_{1,2})$ in terms of $y_{1,2}$; we then substitute the small μ expansion (5.32) for $y_{1,2}$ and re-expand the relevant coefficients about $\bar{x}_{1,2}$.) These expansions are only needed for matching the inner and outer solutions. They are given as needed later (see (5.54)–(5.58)).

The amplitude expansion. Invoking expansion (5.23)–(5.25) into (5.19)–(5.20), the equations determining the amplitude become

$$2P\,\partial_{x_1}\psi_{10} + 2Q\,\partial_{x_2}\psi_{10} + I_0[\partial^2_{x_1} + \partial^2_{x_2}]\log I_0 = 0, \tag{5.41}$$

$$2P\,\partial_{x_1}\psi_{20} + 2Q\,\partial_{x_2}\psi_{20} + (\partial_{x_1}\psi_{10})^2 + (\partial_{x_2}\psi_{10})^2 + [\partial^2_{x_1} + \partial^2_{x_2}]\psi_{10} = 0, \tag{5.42}$$

supplemented with the initial conditions

$$\psi_{10} = \psi_{20} = 0 \quad \text{at } x_{1,2} = x'_{1,2}. \tag{5.43}$$

The next order equation, for ψ_{11}, becomes

$$\begin{aligned} &2P\,\partial_{x_1}\psi_{11} + 2Q\,\partial_{x_2}\psi_{11} - (\partial_{x_1} I_1)(\partial_{x_1}\psi_{10}) - (\partial_{x_2} I_1)(\partial_{x_2}\psi_{10}) \\ &\quad - \tfrac{1}{2} I_1[\partial^2_{x_1} + \partial^2_{x_2}]\log I_0 - \tfrac{1}{2} I_0[\partial^2_{x_1} + \partial^2_{x_2}](I_1/I_0) = 0, \end{aligned} \tag{5.44}$$

supplemented with the initial conditions

$$\psi_{11} = 0 \quad \text{at } x_{1,2} = x'_{1,2}. \tag{5.45}$$

In (5.41)–(5.42) and (5.44),

$$P\,\partial_{x_1}\psi_{ij} + Q\,\partial_{x_2}\psi_{ij} = \frac{\mathrm{d}\psi_{ij}}{\mathrm{d}\mu}$$

along a characteristic or ray. Upon solving these equations, about the stationary point at $\zeta = 0$, we obtain the transform-domain expansion for the characteristic Green's function,

$$\begin{aligned} &\tilde{G}(x_{1,2}, x'_{1,2}; \zeta)\, \exp[-\mathrm{i}k_0(x_3 - x'_3)\zeta] = \mathsf{C}\,\frac{1}{\sqrt{I_0}} \exp[\mathrm{i}k_0(I_0 - \tfrac{1}{2}\zeta^2 I_1) + \psi_{10}] \\ &\left\{ 1 - \mathrm{i}k_0\zeta(x_3 - x'_3) + \frac{1}{\mathrm{i}k_0}\left(\frac{1}{8I_0} + \psi_{20} + \mathrm{i}k_0\zeta^2\left(\psi_{11} + \frac{I_1}{4I_0} + \tfrac{1}{2}[\mathrm{i}k_0(x_3 - x'_3)]^2\right)\right.\right. \\ &\qquad \left.\left. - \tfrac{1}{8}(\mathrm{i}k_0\zeta^2)^2 I_2 \right) + \dots \right\}. \end{aligned} \tag{5.46}$$

TABLE 5.1
Relevant equations.

I_0	I_1	I_2	ψ_{10}	ψ_{20}	ψ_{11}
(5.38)	(5.39)	(5.40)	(5.41)	(5.42)	(5.44)

Carrying out the inverse Fourier transform with the method of stationary phase then results in

$$G(x_{1,2}, x_3 - x_3'; x_{1,2}') = \mathsf{C}\left(\frac{-\mathrm{i}k_0}{2\pi}\right)^{1/2} \exp(\psi_{10}) \frac{\exp(\mathrm{i}k_0 I_0)}{(I_0 I_1)^{1/2}}$$
$$\left\{1 + \frac{1}{\mathrm{i}k_0 I_0}\left(\frac{3}{8}\left(1 - \frac{I_2 I_0}{I_1^2}\right) + \left(\frac{\psi_{11}}{I_1} + \psi_{20}\right) I_0 + \tfrac{1}{2}[\mathrm{i}k_0(x_3 - x_3')]^2 \frac{I_0}{I_1}\right) + \dots\right\}. \tag{5.47}$$

Effective index of refraction, effective metric and uniform asymptotic expansion. As in the two-dimensional case [10], for notational convenience, we introduce the *effective* index of refraction and *effective* horizontal distance as

$$\nu \equiv \left[\frac{I_0}{I_1}\right]^{1/2}, \tag{5.48}$$

$$\chi_1 \equiv [I_0 I_1]^{1/2}, \tag{5.49}$$

where the arguments are evaluated along the characteristics, whereas

$$r = [\chi_1^2 + z^2]^{1/2}. \tag{5.50}$$

Then a uniform asymptotic expansion is

$$G(x_{1,2}, x_3 - x_3'; x_{1,2}') = \frac{1}{4\pi r} \exp(\mathrm{i}k_0 \nu r) \exp(\psi_{10} r^2/\chi_1^2) \tag{5.51}$$
$$\left\{ 1 + \frac{r}{\mathrm{i}k_0 \nu \chi_1^2}\left(\frac{3}{8}\left(1 - \frac{\nu^3 I_2}{\chi_1}\right) + \nu(\nu\psi_{11} + \chi_1\psi_{20})\right) \right.$$
$$\left. + \frac{1}{8}\left(1 - \frac{\nu^3 I_2}{\chi_1}\right)\left(\frac{(x_3 - x_3')^2}{\chi_1^2 r^2}[\mathrm{i}k_0 \nu r (x_3 - x_3')^2 + r^2 + \chi_1^2]\right) + \dots \right\}.$$

The equations to be evaluated or solved are listed in Table 5.1.

In the *outer* region, $\chi_1 = \mathcal{O}(1)$, whence

$$\frac{r}{\chi_1} \sim 1 + \frac{(x_3 - x_3')^2}{2\chi_1^2} = 1 + \mathcal{O}(k_0^{-2}), \tag{5.52}$$

$$\nu r = \nu\chi_1 + \nu\frac{(x_3 - x_3')^2}{2\chi_1} + \mathcal{O}(k_0^{-4}), \tag{5.53}$$

and the uniform solution reduces to the outer solution (5.47) with

$$\mathsf{C} = \left(\frac{\mathrm{i}}{8\pi k_0}\right)^{1/2}.$$

On the *inner* region, $\chi_1 = \mathcal{O}(k_0^{-1})$, whence

$$\frac{1}{\chi_1^2}\left(1-\frac{\nu^3 I_2}{\chi_1}\right) \sim -\frac{1}{3}\left[\left(\frac{\partial_1 n}{n}\right)^2 + \left(\frac{\partial_2 n}{n}\right)^2\right], \tag{5.54}$$

$$\frac{\nu}{\chi_1^2}(\nu\psi_{11}+\chi_1\psi_{20}) \sim \frac{1}{12}\left[\frac{[\partial_1^2+\partial_2^2]n}{n} + \left(\frac{\partial_1 n}{n}\right)^2 + \left(\frac{\partial_2 n}{n}\right)^2\right], \tag{5.55}$$

$$\frac{\psi_{10}}{\chi_1^2} \sim \frac{1}{12}\left[-\frac{[\partial_1^2+\partial_2^2]n}{n} + \left(\frac{\partial_1 n}{n}\right)^2 + \left(\frac{\partial_2 n}{n}\right)^2\right], \tag{5.56}$$

$$\begin{aligned}\nu r \sim ny\Bigg(1+\frac{1}{24}\Bigg\{&\frac{[y_1\partial_1+y_2\partial_2]^2 n}{n} + \left(\frac{y_1\partial_1 n}{n}\right)^2 + \left(\frac{y_2\partial_2 n}{n}\right)^2 \\ &+\left[\left(\frac{\partial_1 n}{n}\right)^2 + \left(\frac{\partial_2 n}{n}\right)^2\right]\left(-y^2+\frac{(x_3-x_3')^4}{y^2}\right)\Bigg\}\Bigg),\end{aligned} \tag{5.57}$$

$$\frac{1}{r} \sim \frac{1}{y}\left(1-\frac{1}{24}\left[\left(\frac{\partial_1 n}{n}\right)^2 + \left(\frac{\partial_2 n}{n}\right)^2\right]\frac{r_2^4}{y^2}\right), \tag{5.58}$$

where r_2 was defined in (5.35), and the argument of n is $\bar{x}_{1,2}$. Substitution of these results into the uniform solution yields the inner solution (5.10).

The inner and outer solutions match when $\chi_1 = \mathcal{O}(k_0^{-2/3})$, which scaling defines the boundary layer.

6. Uniform asymptotic expansion of the characteristic Green's function: The presence of a caustic. In the generic case of a heterogeneous slab, caustics will form in the transverse directions. Following the Maslov approach, we note that there will always be two coordinates chosen from (y_1, y_2) and their Fourier duals (η_1, η_2) such that the solution in these coordinates remains asymptotically finite and meaningful. The transition from one doublet of coordinates to another is followed by the Keller–Maslov line bundle [11] that is accounted for in the solution's amplitude through a tensor product. We will discuss the mixed (y_1, η_2) case here; together with the previous section, all necessary combinations can be found by permutation of coordinates.

We reconsider (3.4) once again,

$$[\partial_k\partial_k + k_0^2 n^2(x_{1,2})]\, G(x_{1,2}, x_3 - x_3'; x_{1,2}') = -\delta(x_1-x_1')\delta(x_2-x_2')\delta(x_3-x_3'),$$

and introduce a slight change in notation,

$$y_{1,2} = x_{1,2} - x_{1,2}', \qquad z = x_3 - x_3'.$$

We write the Green's function in the form of an appropriate Fourier integral,

$$G(x_{1,2}, x_3 - x_3'; x_{1,2}') = \frac{k_0}{2\pi}\int_{\mathbb{R}} \hat{G}(y_1, \eta_2, z; x_{1,2}')\, \exp(\mathrm{i}k_0\eta_2 y_2)\, \mathrm{d}\eta_2. \tag{6.1}$$

We now distinguish amplitude and phase according to

$$\hat{G}(y_1, \eta_2, z; x_{1,2}') = A(y_1, \eta_2, z; x_{1,2}')\exp[\mathrm{i}k_0\phi(y_1, \eta_2, z; x_{1,2}')]. \tag{6.2}$$

The inner solution. We expand the index of refraction, $n^2(x'_{1,2}+y_{1,2})$, in the partial differential equation in y_2 about $-q$ where $q\equiv\partial_{\eta_2}\phi$. Upon substituting the Fourier representation into the Helmholtz equation (3.4), after several integrations by parts, we obtain the following equations.

Up to highest order (lowest order in $(\mathrm{i}k_0)^{-1}$), we recover the eikonal equation, viz.,

$$p^2+\eta_2^2+r^2-n^2(x'_1+y_1,x'_2-q)=0, \tag{6.3}$$

where $p\equiv\partial_{y_1}\phi$ and $r\equiv\partial_z\phi$. This is a pseudodifferential equation for ϕ. In the spirit of the solution method of characteristics, we deduce the Hamilton equations for the bicharacteristics,

$$\begin{aligned}
&\frac{\mathrm{d}y_1}{\mathrm{d}\lambda}=p, && \frac{\mathrm{d}p}{\mathrm{d}\lambda}=(\partial_1M)(x'_1+y_1,x'_2-q),\\
&\frac{\mathrm{d}\eta_2}{\mathrm{d}\lambda}=(\partial_2M)(x'_1+y_1,x'_2-q), && \frac{\mathrm{d}q}{\mathrm{d}\lambda}=-\eta_2,\\
&\frac{\mathrm{d}z}{\mathrm{d}\lambda}=r, && \frac{\mathrm{d}r}{\mathrm{d}\lambda}=0,
\end{aligned} \tag{6.4}$$

in which

$$M=\tfrac{1}{2}n^2.$$

The Hamilton system is supplemented with initial conditions at $\lambda=0$:

$$(y_1,\eta_2,z)|_0=(0,\beta_2,0),\quad (p,q,r)|_0=(\beta_1,0,\beta_3),\quad \beta_1^2+\beta_2^2+\beta_3^2=n^2(x'_{1,2}). \tag{6.5}$$

In the next order, we recover the transport-like equation for A, viz.,

$$(\partial_{y_1}^2+\partial_z^2)A+2\mathrm{i}k_0CA-DA=-\delta(y_1)\delta(z), \tag{6.6}$$

in which

$$\begin{aligned}
CA&=[p\,\partial_{y_1}+r\partial_z+(\partial_2M)(x'_1+y_1,x'_2-q)\,\partial_{\eta_2}]\,A\\
&\quad+\tfrac{1}{2}\left[\partial_{y_1}^2\phi+\partial_z^2\phi-(\partial_2^2M)(x'_1+y_1,x'_2-q)\,\partial_{\eta_2}^2\phi\right]A, \qquad (6.7)\\
DA&=\left[(\partial_2^2M)(x'_1+y_1,x'_2-q)\,\partial_{\eta_2}^2-(\partial_{\eta_2}q)(\partial_2^3M)(x'_1+y_1,x'_2-q)\,\partial_{\eta_2}\right]A\\
&\quad+\left[-\tfrac{1}{3}(\partial_{\eta_2}^2q)(\partial_2^3M)(x'_1+y_1,x'_2-q)+\tfrac{1}{4}(\partial_{\eta_2}q)^2(\partial_2^4M)(x'_1+y_1,x'_2-q)\right]A\\
&\quad+\mathcal{O}((\mathrm{i}k_0)^{-1}). \qquad (6.8)
\end{aligned}$$

Observe that on the inner region the y_1 and z derivatives are large, and, hence, the inner transport-like equation reduces to

$$(\partial_{y_1}^2+\partial_z^2)A+2\mathrm{i}k_0CA=-\delta(y_1)\delta(z)+\dots. \tag{6.9}$$

The phase function. First, we evaluate the solutions to the Hamilton (cf. (6.4)) and eikonal (cf. (6.3)) equations for small values of λ. The parametric representation of the Hamiltonian flow follows from (6.4) as

$$\begin{aligned}
&y_1=\beta_1\lambda+\tfrac{1}{2}(\partial_1M)\,\lambda^2+\dots, && p=\beta_1+(\partial_1M)\,\lambda+\dots,\\
&\eta_2=\beta_2+(\partial_2M)\,\lambda+\dots, && q=-\beta_2\lambda-\tfrac{1}{2}(\partial_2M)\,\lambda^2+\dots,\\
&z=\beta_3\lambda, && r=\beta_3
\end{aligned} \tag{6.10}$$

while, through integration of the canonical one-form along the bicharacteristic, the phase function is found to be

$$\phi = (\beta_1^2 + \beta_3^2)\,\lambda + [\beta_1(\partial_1 M) - \tfrac{1}{2}\beta_2(\partial_2 M)]\,\lambda^2 + \dots, \tag{6.11}$$

in which the substitution $(\partial_{1,2}M) = (\partial_{1,2}M)(x'_{1,2})$ is understood.

Solving system (6.10) subject to the constraint in (6.5) for $\lambda, \beta_1, \beta_2, \beta_3$ in terms of y_1, η_2, z yields

$$\lambda = (R/\gamma)[1 - \gamma^{-3}(\tfrac{1}{2}\gamma y_1(\partial_1 M) + \eta_2 R(\partial_2 M)) + \dots], \tag{6.12}$$

$$\beta_1 = (\gamma/R)[y_1 + \gamma^{-3}(-\tfrac{1}{2}\gamma z^2(\partial_1 M) + \eta_2 y_1 R(\partial_2 M)) + \dots], \tag{6.13}$$

$$\beta_2 = \eta_2 - (R/\gamma)(\partial_2 M) + \dots, \tag{6.14}$$

$$\beta_3 = (\gamma z/R)[1 + \gamma^{-3}(\tfrac{1}{2}\gamma y_1(\partial_1 M) + \eta_2 R(\partial_2 M)) + \dots], \tag{6.15}$$

in which $R \equiv (y_1^2 + z^2)^{1/2}$ and $\gamma \equiv [n^2(x'_{1,2}) - \eta_2^2]^{1/2}$. Substituting these solutions in the remaining equations of system (6.10) gives

$$\begin{aligned} p &= (\gamma/R)[y_1 + \gamma^{-3}(\tfrac{1}{2}\gamma(y_1^2 + R^2)(\partial_1 M) + \eta_2 y_1 R(\partial_2 M)) + \dots], \\ q &= (R/\gamma)[-\eta_2 + \gamma^{-4}\tfrac{1}{2}(\gamma^2 y_1\eta_2(\partial_1 M) + (\gamma^2 + 2\eta_2^2)R(\partial_2 M)) + \dots], \\ r &= (\gamma z/R)[1 + \gamma^{-3}(\tfrac{1}{2}\gamma y_1(\partial_1 M) + \eta_2 R(\partial_2 M)) + \dots], \end{aligned} \tag{6.16}$$

whereas the phase function (6.11) takes the form

$$\phi = \gamma R\,[1 + \gamma^{-3}(\tfrac{1}{2}\gamma y_1(\partial_1 M) + \tfrac{1}{2}\eta_2 R(\partial_2 M)) + \dots]. \tag{6.17}$$

The amplitude function. Having obtained the solution of the eikonal equation, we now proceed with solving the transport-like equation. First, observe the following property of functions F of $k_0\phi$:

$$\left[\partial_{y_1}^2 + \partial_z^2 + 2\mathrm{i}k_0 C\right] F = k_0^2(p^2 + r^2)$$

$$\left\{F'' + 2\mathrm{i}F' + \frac{1}{k_0\phi}\,(F' + \mathrm{i}F) + \frac{\eta_2(\partial_2 M)(x'_{1,2})}{2k_0\gamma^4}\,(F' + \mathrm{i}F - 4\mathrm{i}k_0\phi F') + \dots\right\}.$$

Using this property, the inner solution of (6.9) is constructed and found to be

$$\begin{aligned} A = \frac{\mathrm{i}}{4}\exp(-\mathrm{i}k_0\phi) \\ \left\{H_0^{(1)}(k_0\phi) + \frac{\eta_2(\partial_2 M)(x'_{1,2})}{2\gamma^4}k_0\phi^2[H_1^{(1)}(k_0\phi) - \mathrm{i}H_2^{(1)}(k_0\phi)] + \dots\right\}. \end{aligned} \tag{6.18}$$

The outer solution. We assume that our wave field is a transient phenomenon with dominant wave number k_0. The outer region is determined by the condition

$$\|(x_1 - x'_1, x_2 - x'_2)\| = \mathcal{O}(1)$$

and corresponds to the behavior of the kernels away from their diagonals.

Amplitude expansion. In the outer region the derivatives of the amplitude A in (6.2) are $\mathcal{O}(1)$. Thus we expand

$$A = \frac{1}{(k_0\phi)^{1/2}}\left\{A_0 + \frac{1}{\mathrm{i}k_0}\left(\frac{A_0}{8\phi} + A_1\right) + \dots\right\}. \tag{6.19}$$

With this definition, A_0 and A_1 are continuous near the "source" (at $x'_{1,2}$). Substitution of (6.19) into (6.6) and setting terms proportional to $k_0^{-n+1/2}$, $n = 0, 1, \ldots$, equal to zero yields the transport equations

$$LA_0 = 0, \tag{6.20}$$

$$\begin{aligned} LA_1 - \phi^{1/2} D(\phi^{-1/2} A_0) + \partial_{y_1}^2 A_0 + \partial_z^2 A_0 \\ + \frac{1}{\phi}(\partial_2 M)(x'_1 + y_1, x'_2 - q) \left(\partial_{\eta_2} A_0 + \frac{3}{4} \frac{q}{\phi} A_0 \right) \\ - \tfrac{1}{2}(\partial_{\eta_2} q)\, (\partial_2^2 M)(x'_1 + y_1, x'_2 - q) A_0 = 0, \end{aligned} \tag{6.21}$$

where

$$\begin{aligned} LA = 2\frac{\mathrm{d}A}{\mathrm{d}\lambda} + \Big\{ \phi\, [\partial_{y_1}^2 + \partial_z^2] \log\phi \\ - \frac{q}{\phi}(\partial_2 M)(x'_1 + y_1, x'_2 - q) - (\partial_{\eta_2} q)(\partial_2^2 M)(x'_1 + y_1, x'_2 - q) \Big\} A. \end{aligned} \tag{6.22}$$

The nonhomogeneous terms in (6.21) are continuous near the "source."

In preparation for matching the inner and outer solutions, we consider the small λ expansion of the solutions to (6.20)–(6.21):

$$\begin{aligned} A_0 &\to A_0|_0 \left(1 - \frac{3\beta_2 \lambda (\partial_2 M)(x'_{1,2})}{4(\beta_1^2 + \beta_3^2)} + \ldots \right) \\ &= A_0|_0 \left(1 - \frac{3\beta_2 (\partial_2 M)(x'_{1,2}) \phi}{4(\beta_1^2 + \beta_3^2)^2} + \ldots \right), \end{aligned} \tag{6.23}$$

$$A_1 \to A_1|_0 + \ldots. \tag{6.24}$$

Thus near the "source," (6.19) takes the form

$$A = \left(\frac{1}{k_0 \phi} \right)^{1/2} \left\{ A_0|_0 \left(1 + \frac{1}{8\mathrm{i} k_0 \phi} - \frac{3\beta_2 (\partial_2 M)(x'_{1,2})}{32(\beta_1^2 + \beta_3^2)^2} \left[8\phi + \frac{1}{\mathrm{i} k_0} \right] \right) + \frac{1}{\mathrm{i} k_0} A_1|_0 + \ldots \right\}. \tag{6.25}$$

Inner solution on the outer scale. In preparation for developing the inner solution on the outer region, we observe the asymptotic behavior of the amplitude given in (6.18): We have

$$A = \left(\frac{\mathrm{i}}{8\pi k_0 \phi} \right)^{1/2} \left\{ 1 + \frac{1}{8\mathrm{i} k_0 \phi} - \frac{3\eta_2 (\partial_2 M)(x'_{1,2})}{4\gamma^4} \left[\phi - \frac{5}{8\mathrm{i} k_0} \right] + \ldots \right\} \tag{6.26}$$

as $k_0 \phi$ becomes large. On the other hand, approaching the "source" as λ (i.e., ϕ) becomes small, in this expression, gives

$$\frac{\eta_2}{\gamma^4} \to \frac{\beta_2}{(\beta_1^2 + \beta_3^2)^2},$$

cf. (6.10).

Uniform asymptotic expansion. *Matching the inner solution on the overlapping region.* The overlapping region is governed by λ small and $k_0\lambda$ large, i.e., ϕ small and $k_0\phi$ large (or $R = \mathcal{O}(k_0^{-1/3})$, where R was defined just below (6.15)). Comparing (6.25) with (6.26) yields the initial conditions for A_0 and A_1,

$$A_0|_0 = \left(\frac{\mathrm{i}}{8\pi}\right)^{1/2}, \tag{6.27}$$

$$A_1|_0 = \left(\frac{\mathrm{i}}{8\pi}\right)^{1/2} \frac{9\beta_2(\partial_2 M)(x'_{1,2})}{16(\beta_1^2+\beta_3^2)^2}. \tag{6.28}$$

Thus the initial conditions for A_0 and A_1 are determined by matching the outer expansion to the inner solution on the overlapping region. It is only now that the outer solution is fully determined.

Uniform expansion. Finally, the uniform expansion is obtained by adding the outer solution ((6.1), (6.2), and (6.19)) to the inner solution ((6.1), (6.2), and (6.18)) and subtracting the matching terms on the overlapping region (equation (6.26)). However, in the inner region caustics will not have developed yet and, hence, there the noncaustic uniform asymptotic expansion of the previous section will apply.

Expansion in $z = \mathcal{O}(k_0^{-1})$. For use of the expansion of G in the kernels of the vertical slowness operator and the thin-slab propagator, we will have to make the assumption that the propagation distance satisfies

$$k_0 \underbrace{|x_3 - x'_3|}_{z} = \mathcal{O}(1). \tag{6.29}$$

Exploiting the small range of propagation to yield the thin-slab propagator, we expand

$$\phi = I_0 + \tfrac{1}{2}z^2 I_1 + \tfrac{1}{8}z^4 I_2 + \dots, \tag{6.30}$$

$$A_0 = A_{00} + z^2 A_{01} + \dots, \tag{6.31}$$

$$A_1 = A_{10} + \dots, \tag{6.32}$$

where I_0, I_1, I_2 and A_{00}, A_{01}, A_{10} are independent of z.

The phase function. Substituting the expansion for ϕ in (6.3) yields up to leading order

$$P^2 + \eta_2^2 - n^2(x'_1 + y_1, x'_2 - Q) = 0, \tag{6.33}$$

where $P = \partial_{y_1} I_0$ and $Q = \partial_{\eta_2} I_0$. The associated Hamilton system, i.e., the counterpart of (6.4) with the preferred (principal) components removed, becomes

$$\begin{aligned} &\frac{\mathrm{d}y_1}{\mathrm{d}\mu} = P, && \frac{\mathrm{d}P}{\mathrm{d}\mu} = (\partial_1 M)(x'_1 + y_1, x'_2 - Q), \\ &\frac{\mathrm{d}\eta_2}{\mathrm{d}\mu} = (\partial_2 M)(x'_1 + y_1, x'_2 - Q), && \frac{\mathrm{d}Q}{\mathrm{d}\mu} = -\eta_2, \end{aligned} \tag{6.34}$$

supplemented by the initial conditions (cf. (6.5))

$$(y_1, \eta_2)|_0 = (0, \alpha_2), \quad (P, Q)|_0 = (\alpha_1, 0), \quad \alpha_1^2 + \alpha_2^2 = n^2(x'_{1,2}). \tag{6.35}$$

The equation for the next order term follows as

$$\frac{\mathrm{d}I_1}{\mathrm{d}\mu} + I_1^2 = 0 \tag{6.36}$$

with solution

$$I_1 = \frac{1}{\mu}. \tag{6.37}$$

(The initial condition has been matched with the inner solution.)

The equation for I_2 follows as

$$\frac{\mathrm{d}I_2}{\mathrm{d}\mu} + 4I_2I_1 + (\partial_{y_1}I_1)^2 - (\partial_2^2M)(x_1' + y_1, x_2' - Q)(\partial_{\eta_2}I_1)^2 = 0. \tag{6.38}$$

This equation simplifies from a computational point of view upon scaling $I_2 = \mu^{-4}\bar{I}_2$; then

$$\frac{\mathrm{d}\bar{I}_2}{\mathrm{d}\mu} + [(\partial_{y_1}I_1)^2 - (\partial_2^2M)(x_1' + y_1, x_2' - Q)(\partial_{\eta_2}I_1)^2]\mu^4 = 0, \tag{6.39}$$

supplemented by the initial condition

$$\bar{I}_2|_0 = 0. \tag{6.40}$$

We evaluate the solutions to the Hamilton (equation (6.34)) and eikonal (equation (6.33)) equations for small values of μ. The parametric representation of the Hamiltonian flow follows as (compare with (6.10))

$$\begin{aligned} y_1 &= \alpha_1\mu + \tfrac{1}{2}(\partial_1M)\,\mu^2 + \ldots, & P &= \alpha_1 + (\partial_1M)\,\mu + \ldots, \\ \eta_2 &= \alpha_2 + (\partial_2M)\,\mu + \ldots, & Q &= -\alpha_2\mu - \tfrac{1}{2}(\partial_2M)\,\mu^2 + \ldots, \end{aligned} \tag{6.41}$$

while the leading-order constituent phase function is found to be (compare with (6.11))

$$I_0 = \alpha_1^2\mu + [\alpha_1(\partial_1M) - \tfrac{1}{2}\alpha_2(\partial_2M)]\,\mu^2 + \ldots, \tag{6.42}$$

in which the substitution $(\partial_{1,2}M) = (\partial_{1,2}M)(x_{1,2}')$ is understood.

Solving system (6.41) subject to the constraint in (6.35) for μ, α_1, α_2 in terms of y_1, η_2 yields

$$\mu = (|y_1|/\gamma)[1 - \gamma^{-3}(\tfrac{1}{2}\gamma y_1(\partial_1M) + \eta_2|y_1|(\partial_2M)) + \ldots], \tag{6.43}$$

$$\alpha_1 = (\gamma/|y_1|)[y_1 + \gamma^{-3}\eta_2y_1^2(\partial_2M) + \ldots], \tag{6.44}$$

$$\alpha_2 = \eta_2 - (|y_1|/\gamma)(\partial_2M) + \ldots, \tag{6.45}$$

in which $\gamma = [n^2(x_{1,2}') - \eta_2^2]^{1/2}$ as before. Substituting these solutions into (6.42) then yields (compare with (6.17))

$$I_0 = \gamma|y_1|[1 + \gamma^{-3}(\tfrac{1}{2}\gamma y_1(\partial_1M) + \tfrac{1}{2}\eta_2|y_1|(\partial_2M)) + \ldots]. \tag{6.46}$$

Amplitude expansion. Upon substituting (6.30)–(6.32) into (6.20)–(6.21), and collecting leading-order terms, the equations for A_{00} and A_{10} follow as

$$\widetilde{L}A_{00} = 0, \tag{6.47}$$

$$\widetilde{L}A_{10} + 2A_{01} - \frac{3}{4}\frac{Q}{I_0^2}(\partial_2M)(x_1' + y_1, x_2' - Q)\,A_{00} - \widetilde{D}A_{00} = 0, \tag{6.48}$$

where

$$\widetilde{L}A = 2\frac{\mathrm{d}A}{\mathrm{d}\mu} + \Bigg[I_0\,\partial_{y_1}^2 \log I_0 + I_1 - \frac{Q}{I_0}(\partial_2 M)(x_1' + y_1, x_2' - Q) - (\partial_{\eta_2} Q)(\partial_2^2 M)(x_1' + y_1, x_2' - Q) \Bigg] A, \tag{6.49}$$

$$\begin{aligned}\widetilde{D}A = {} & -\partial_{y_1}^2 A - \frac{1}{I_0}(\partial_2 M)(x_1' + y_1, x_2' - Q)\,\partial_{\eta_2} A \\ & + I_0^{1/2} \Bigg[(\partial_2^2 M)(x_1' + y_1, x_2' - Q) \left(\partial_{\eta_2}^2 + \frac{\partial_{\eta_2} Q}{2I_0} \right) \\ & - (\partial_2^3 M)(x_1' + y_1, x_2' - Q)((\partial_{\eta_2} Q)\partial_{\eta_2} + \tfrac{1}{3}(\partial_{\eta_2}^3 Q)) \\ & + \tfrac{1}{4}(\partial_2^4 M)(x_1' + y_1, x_2' - Q)\tfrac{1}{4}(\partial_{\eta_2} Q)^2 \Bigg] \left(\frac{A}{I_0^{1/2}} \right),\end{aligned} \tag{6.50}$$

supplemented with the initial conditions (compatible with (6.27)–(6.28))

$$A_{00}|_0 = \left(\frac{\mathrm{i}}{8\pi} \right)^{1/2}, \tag{6.51}$$

$$A_{10}|_0 = \left(\frac{\mathrm{i}}{8\pi} \right)^{1/2} \frac{9\beta_2 (\partial_2 M)(x_{1,2}')}{16(\beta_1^2 + \beta_3^2)^2}. \tag{6.52}$$

The next order equation, for A_{01}, yields (cf. (6.20))

$$\widetilde{L}A_{01} + 4I_1 A_{01} + \widetilde{L}_1 A_{00} = 0, \tag{6.53}$$

where

$$\begin{aligned}\widetilde{L}_1 A = {} & (\partial_{y_1} I_1)(\partial_{y_1} A) - (\partial_{\eta_2} I_1)(\partial_2^2 M)(x_1' + y_1, x_2' - Q)(\partial_{\eta_2} A) \\ & + \tfrac{1}{2} \left((\partial_{y_1}^2 I_1) - \frac{2P(\partial_{y_1} I_1)}{I_0} + \frac{P^2 I_1}{I_0^2} + 3I_2 - \frac{2I_1^2}{I_0} \right) A \\ & + \Bigg[\left(\frac{QI_1}{I_0^2} - \frac{(\partial_{\eta_2} I_1)}{I_0} \right) (\partial_2 M)(x_1' + y_1, x_2' - Q) \\ & + \left(\frac{Q(\partial_{\eta_2} I_1)}{I_0} - (\partial_{\eta_2}^2 I_1) \right) (\partial_2^2 M)(x_1' + y_1, x_2' - Q) \\ & + (\partial_{\eta_2} I_1)(\partial_{\eta_2} Q)(\partial_2^3 M)(x_1' + y_1, x_2' - Q) \Bigg] A.\end{aligned} \tag{6.54}$$

This equation simplifies from a computational point of view upon scaling $A_{01} = \mu^{-2} \bar{A}_{01}$; then

$$\widetilde{L}\bar{A}_{01} + \mu^2 \widetilde{L}_1 A_{00} = 0 \tag{6.55}$$

with the initial condition $\bar{A}_{01}|_0 = 0$.

We remark that the inhomogeneous term in (6.55) is continuous at the "source" $(x_{1,2}')$ and

$$A_{01} \to -\frac{3\alpha_2 (\partial_2 M)(x_{1,2}')}{8\alpha_1^4 \mu} A_{00}|_0 \quad \text{as} \quad \mu \to 0,$$

TABLE 6.1
Relevant equations.

I_0	I_1	I_2	A_{00}	A_{10}	A_{01}
(6.33)	(6.37)	(6.39)	(6.47)	(6.48)	(6.55)

cf. (6.41) for Q and (6.42) for I_0. The inhomogeneous term in (6.48) is continuous at the "source" $(x'_{1,2})$ since

$$-\frac{3}{4}\frac{Q}{I_0^2}(\partial_2 M)(x'_1+y_1,x'_2-Q)\to\frac{3}{4}\frac{\alpha_2(\partial_2 M)(x'_{1,2})}{\alpha_1^4\mu}\qquad\text{as}\qquad\mu\to 0.$$

Effective index of refraction and effective metric. Again, we introduce an *effective* index of refraction and *effective* horizontal distance as

$$\nu\equiv[I_0I_1]^{1/2}, \tag{6.56}$$

$$\chi_1\equiv\left[\frac{I_0}{I_1}\right]^{1/2}, \tag{6.57}$$

where the arguments are evaluated along the characteristics, whereas

$$r=[\chi_1^2+z^2]^{1/2}. \tag{6.58}$$

Then

$$\begin{aligned}G(x_{1,2},x_3-x'_3;x'_{1,2})&\\ =\frac{k_0}{2\pi}\int_{\mathbb{R}}\frac{1}{(k_0\nu r)^{1/2}}\exp[\mathrm{i}k_0(\nu r+\eta_2y_2)]&\left\{A_{00}+\frac{1}{\mathrm{i}k_0}\left(\frac{A_{00}}{8\nu r}+A_{10}\right)+\dots\right\}\mathrm{d}\eta_2,\end{aligned} \tag{6.59}$$

cf. (6.1), (6.2), (6.19), which represents the outer solution. The equations to be evaluated or solved are listed in Table 6.1.

7. Uniform asymptotic expansions of the vertical slowness operator and the one-way wave propagator.

The square-root operator kernel. Using (4.8), upon carrying out repeated differentiation, we arrive at the uniform asymptotic expansion of the square-root operator kernel,

$$\begin{aligned}\mathcal{C}(x_{1,2},x'_{1,2};x'_3)=\frac{\mathrm{i}}{2\pi\omega\chi_1^3}\exp(\mathrm{i}k_0\nu\chi_1)\exp(\psi_{10})&\\ \left\{(\mathrm{i}k_0\nu\chi_1+2\psi_{10})\left(1+\frac{1}{\mathrm{i}k_0\nu\chi_1}\left(\frac{3}{8}\left(1-\frac{\nu^3I_2}{\chi_1}\right)+\nu(\nu\psi_{11}+\chi_1\psi_{20})\right)\right)&\right.\\ \left.-\frac{1}{2}\left(1+\frac{\nu^3I_2}{\chi_1}\right)+\dots\right\},&\end{aligned} \tag{7.1}$$

in the absence of caustics, and the outer expansion,

$$\begin{aligned}\mathcal{C}(x_{1,2},x'_{1,2};x'_3)=\frac{\mathrm{i}k_0}{\omega\pi}\int_{\mathbb{R}}\exp[\mathrm{i}k_0(\nu\chi_1+\eta_2y_2)]&\\ \left\{\frac{1}{(\nu\chi_1)^{1/2}\chi_1^2}\left(\left(\mathrm{i}k_0\nu\chi_1-\frac{3}{8}\right)A_{00}+2\chi_1^2A_{01}+\nu\chi_1A_{10}+\dots\right)\right\}\mathrm{d}\eta_2,&\end{aligned} \tag{7.2}$$

in the presence of a caustic, to the order considered.

The propagator kernel. Using (4.2), we arrive at the uniform asymptotic expansion of the one-way propagator kernel,

$$\begin{aligned}(7.3)\quad &\mathcal{G}(x_{1,2}, x_3 - x_3'; x_{1,2}') \\ &= -2(x_3 - x_3')\left\{ \left(\frac{\mathrm{i}k_0\nu}{r} + \frac{2\psi_{10}}{\chi_1^2}\right) G(x_{1,2}, x_3 - x_3'; x_{1,2}') \right. \\ &\quad -\frac{1}{4\pi r^3}\exp(\mathrm{i}k_0\nu r)\exp\left(\frac{\psi_{10}r^2}{\chi_1^2}\right) \\ &\qquad \left.\left[-1 + \frac{1}{8\chi_1^2 r^2}\left(1 - \frac{\nu^3 I_2}{\chi_1}\right)(4\chi_1^4 + (x_3 - x_3')^2(r^2 + \chi_1^2))\right] + \dots \right\}\end{aligned}$$

in the absence of caustics, and the outer expansion,

$$\begin{aligned}(7.4)\quad &\mathcal{G}(x_{1,2}, x_3 - x_3'; x_{1,2}') = -\frac{k_0}{\pi}\int_{\mathbb{R}} \frac{(x_3 - x_3')}{r^2(k_0\nu r)^{1/2}} \\ &\qquad \exp[\mathrm{i}k_0(\nu r + \eta_2 y_2)]\left\{\left(\mathrm{i}k_0\nu r - \frac{3}{8}\right)A_{00} + \nu r A_{10} + 2r^2 A_{01} + \dots\right\}\mathrm{d}\eta_2,\end{aligned}$$

in the presence of a caustic, to the order considered. In both cases, we observe that

$$\lim_{x_3 \downarrow x_3'} \mathcal{G}(x_{1,2}, x_3 - x_3'; x_{1,2}') = \delta(x_{1,2} - x_{1,2}'),$$

as it should.

8. Discussion. One of the main objectives of directional wave field decomposition is the introduction of the concept of "tracing waves." A general theory for this, employing the complete generalized Bremmer coupling series, has been developed before. The application of the series, however, depends on solving an operator composition equation, the characteristic equation, and an associated one-way wave equation. In this paper, in smoothly varying media, we have obtained uniform asymptotic expansions for both solutions valid in the "high- and mid-frequency" wave regime.

The method of uniform asymptotics consists of three components: (i) the construction of a "far-field" or "outer" solution, representing an operator kernel away from its diagonal and obtained by microlocal techniques suppressing locally medium variations in the principal (here vertical) direction; (ii) the construction of a "near-field" or "inner" solution, representing the operator kernel near its diagonal and obtained mostly by Taylor-like expansions; (iii) matching the inner and outer solutions in a boundary layer to all orders considered.

The result is a one-way wave field representation that is truly more general than its microlocal counterpart. For example, the microlocal treatment of the one-way operator solutions to the characteristic equation would require cut-offs removing critical-angle scattering phenomena. Also, modal behavior is naturally included in our framework of uniform asymptotics. Conceptually, our theory is an intermediate between asymptotic-ray and full-wave theories in the sense that our theory is still asymptotic but valid in a much larger frequency band (see also Thomson [12]).

From a computational perspective, the uniform asymptotic one-way wave propagator falls into the category of propagators associated with the paraxial wave equation, the phase-screen or split-step Fourier approximation, the phase-shift-plus-interpolation

method, and so on. However, it does not suffer from any of the limitations of these approaches. A desirable feature of a closed-form solution as presented in this paper is its ease of use, in particular with a view to taking caustics into account. (In this context, for a comparison with a ray tracing approach, see Ziomek [13]).

Hidden in the uniform asymptotic expansions are certain aspects of homogenization: we have introduced an effective index of refraction and an effective metric, which follow from the actual medium variations and are evaluated by means of ray methods.

Throughout the paper, the configuration has been assumed to be three-dimensional. Previous two-dimensional results, obtained by more restrictive arguments, are recovered by assuming that $\partial_2 n \equiv 0$ and integrating the characteristic Green's function over y_2.

As a final remark, we indicate how variable density can be incorporated in the analysis. For details on how it affects the decomposition procedure, see de Hoop [1]. The key in the approach presented in this paper is the introduction of an effective wave speed, $c'^{-2}(x_{1,2},\omega) = c_0^{-2} n'^2(x_{1,2},\omega)$, with

$$n'^2(x_{1,2},\omega) = n^2(x_{1,2}) + c_0^2 \left[\frac{3[(D_1\rho)^2 + (D_2\rho)^2]}{4\rho^2} - \frac{(D_1^2 + D_2^2)\rho}{2\rho} \right], \quad \rho = \rho(x_1, x_2).$$

This change requires some straightforward adjustments of the asymptotic matching.

REFERENCES

[1] M.V. de Hoop, *Generalization of the Bremmer coupling series*, J. Math. Phys., 37 (1996), pp. 3246–3282.

[2] F. Treves, *Introduction to Pseudodifferential and Fourier Integral Operators*, 2, Plenum Press, New York, 1980.

[3] M.V. de Hoop, *Direct, leading-order asymptotic, inverse scattering based on the generalized Bremmer series*, in Mathematical and Numerical Aspects of Wave Propagation, SIAM, Philadelphia, 1998, pp. 249–253.

[4] L.N. Frazer, *Synthetic seismograms using multifold path integrals* - I. *Theory*, Geophys. J. R. Astr. Soc., 88 (1987), pp. 621–646.

[5] M.V. de Hoop, J.H. Le Rousseau, and B. Biondi, *Symplectic structure of wave-equation imaging: A path-integral approach based on the double-square-root equation*, Geophys. J. Int., to appear.

[6] M.J.N. van Stralen, M.V. de Hoop, and H. Blok, *Generalized Bremmer series with rational approximation for the scattering of waves in inhomogeneous media*, J. Acoust. Soc. Am., 104 (1998), pp. 1943–1963.

[7] M.V. de Hoop, J.H. Le Rousseau, and R.-S. Wu, *Generalization of the phase-screen approximation for the scattering of acoustic waves*, Wave Motion, 31 (2000), pp. 43–70.

[8] L. Fishman, M.V. de Hoop, and M.J.N. van Stralen, *Exact constructions of square-root Helmholtz operator symbols: The focusing quadratic profile*, J. Math. Phys., 41 (2000), pp. 4881–4938.

[9] L. Fishman, A.K. Gautesen, and Z. Sun, *Uniform high-frequency approximations of the square root Helmholtz operator symbol*, Wave Motion, 26 (1997), pp. 127–161.

[10] M.V. de Hoop and A.K. Gautesen, *Uniform asymptotic expansion of the generalized Bremmer series*, SIAM J. Appl. Math., 60 (2000), pp. 1302–1329.

[11] L. Hörmander, *The Analysis of Linear Partial Differential Operators* III, Springer–Verlag, Berlin, 1985.

[12] C.J. Thomson, *The 'gap' between seismic ray theory and 'full' wavefield extrapolation*, Geophys. J. Int., 137 (2001), pp. 364–380.

[13] L.J. Ziomek, *Sound-pressure level calculations using the RRA algorithm for depth-dependent speeds of sound valid at turning points and focal points*, IEEE J. Oceanic Eng., 19 (1994), pp. 242–248.

SIAM J. APPL. MATH.
Vol. 63, No. 3, pp. 801–817

ANALYTICAL AND NUMERICAL INVESTIGATION OF THE RESOLVENT FOR PLANE COUETTE FLOW*

MATTIAS LIEFVENDAHL† AND GUNILLA KREISS†

Abstract. We present new bounds for the solution of the resolvent equation for plane Couette flow. Both analytic methods and computation, using the Chebyshev tau spectral method, are used. The emphasis is on determining the Reynolds number-dependence of the estimates. The main result is the introduction of a weighted norm, which leads to optimal asymptotic behavior of the resolvent for large Reynolds numbers.

Key words. hydrodynamical stability, resolvent estimate, threshold amplitude

AMS subject classifications. 76E05, 35Q35, 47N20, 35P05

PII. S0036139901396759

1. Introduction.

Hydrodynamic stability of shear flows is a classic topic in applied mathematics with roots in the 19th century. We refer to the books [14], [1], [3], and [20] for an outline of the field and general background of this work.

In this paper we study plane Couette flow, i.e., flow of a viscous incompressible fluid between two moving planes. It is assumed that there is no pressure gradient, and so the stationary velocity profile is linear. The spectrum for Couette flow was thoroughly investigated in the period 1950–1970. We mention the papers [26], [5], and [7], which are concerned with bounds for the most unstable eigenvalues. These efforts culminated in a paper of Romanov [18], where it is shown that all eigenvalues λ satisfy $\operatorname{Re}\lambda \leq -\delta/R$ for some $\delta > 0$. Here R denotes the Reynolds number. In [18] the fact that the eigenvalues lie in the stable complex half-plane is combined with semigroup theory to prove the nonlinear stability of plane Couette flow for all Reynolds numbers. This is a rather special situation for shear flows; e.g., Poiseuille flow becomes unstable for $R > 5772$. This was shown in [23] and [19]. The accurate value of the critical Reynolds number was obtained in [15] with the same numerical method we use in this paper to study the norm of the resolvent.

Although plane Couette flow is stable to infinitesimal perturbations, it can be excited to turbulence by finite perturbations. There is a threshold value for the size of the perturbations. Below this value the perturbations decay to zero, and above the threshold the perturbation may lead to turbulence. The question of what the threshold value is and its dependence on the Reynolds number is the motivation for the present work. It has been known for a very long time that the threshold decreases with increasing Reynolds number. In the beginning of the 1990s the hypothesis that the threshold is proportional to $R^{-\beta}$ (for some $\beta > 0$) was starting to be investigated [25], [24]. Computations in [16] suggest a β-value of approximately 5/4. The asymptotic analysis of [2] suggests an approximate β-value of 1.

Another approach to this problem is found in [9], where a bound on the resolvent operator in the entire unstable half-plane is used to show nonlinear stability. The resolvent estimate depends on the Reynolds number. In [9] it is proven that Couette

*Received by the editors October 22, 2001; accepted for publication (in revised form) July 5, 2002; published electronically January 17, 2003.
http://www.siam.org/journals/siap/63-3/39675.html
†Department of Numerical Analysis and Computing Science, Royal Intitute of Technology, Stockholm S-100 44, Sweden (mli@nada.kth.se, gunillak@nada.kth.se).

flow is stable to perturbations of amplitude $\leq CR^{-21/4}$. Thus $21/4$ is an upper bound on the exponent β.

The proof in [9] is based on a resolvent estimate which is obtained by computation. In this paper, we analytically derive a resolvent estimate in a sector of the complex (s-)plane, which covers all but a bounded part of the unstable half-plane. In this estimate we include both the dependence on the Reynolds number and the complex number s. For the Fourier transformed resolvent equation we give an estimate in the entire unstable half-plane which holds for certain wave numbers. These results are not sufficient for stability investigations but should be seen as the start of more detailed investigations of the resolvent for Couette flow.

To obtain estimates directly applicable to nonlinear stability, we must resort to numerical methods. An important result of this paper is obtained in this way. The resolvent estimate of [9] is

$$\|\mathbf{u}\| \leq CR^2 \|\mathbf{f}\| . \tag{1.1}$$

Here $\mathbf{u}$ is the velocity field, $\mathbf{f}$ is the forcing, and we use the L^2-norm. It is well known that, with respect to this norm, the resolvent is increasingly nonnormal as R grows. This means that the eigenfunctions of the linear problem are increasingly nonorthogonal and that there exist initial perturbations such that the norm of the solution of the linear problem grows initially. The largest possible growth increases with R. By using a different norm, the normality properties of the resolvent and the orthogonality properties of the eigenfunctions change.

If the resolvent $\mathcal{R}$ is normal with respect to the inner product defining the norm, then with this norm

$$\|\mathcal{R}\|_N = \frac{1}{\operatorname{dist}(\Omega_U, \Sigma)}; \tag{1.2}$$

see [8, p. 277]. Here $\Omega_U = \{s \in \mathbb{C} : \operatorname{Re} s \geq 0\}$ is the unstable half-plane, and Σ is the spectrum. By considering the Fourier transformed eigenvalue problem with streamwise wavenumber equal to zero, it is easy to see that the eigenvalue bound of Romanov is sharp. Thus the right-hand side in (1.2) is proportional to R. If $\mathcal{R}$ is nonnormal, then its norm is larger.

In this paper we introduce new norms by weighting the different velocity components with R-dependent coefficients. We investigate several possibilities, and the best result is

$$\|\mathbf{u}\|_3 \leq CR \|\mathbf{f}\|_3 . \tag{1.3}$$

C denotes a generic constant. The norm with subscript 3 is introduced in section 2.2. We see that we gain a factor R in the right-hand side. By the above argument, this result is optimal in the sense that we can improve only the constant C; the exponent of R cannot be lower than one.

We obtain (1.3) by numerical computations in which we maximize the Fourier transformed resolvent over different wave numbers. The maximum occurs when the streamwise wave number equals zero, corresponding to disturbances without streamwise variations.

In [13], the inequality (1.3) is used to considerably improve the exponent $21/4$ for the threshold mentioned above. In the present paper we also investigate the weighting proposed in [9] and show that it does not lead to an improved exponent.

1.1. Outline of the paper. A definition of the continuous problem and definitions and notation for the various norms we will use are given in section 2.

Section 3 is devoted to the analytical resolvent estimate, which is given in Theorem 3.1. The well-known reformulation of the problem in terms of u_2 and η_2 (the normal component of the vorticity), instead of $\mathbf{u}$ and p, is presented in section 4. The equations are also Fourier transformed. In the same section we state an analytical estimate for the Fourier transformed problem, in Lemma 4.1. After this we turn to numerical methods, which are applied to the reformulated problem.

In section 5 we present results of computations of the different norms of the resolvent. The most important result is the optimal R-exponent in the resolvent estimate for the third modified norm (defined in section 2.2). For the discretization we use the Chebyshev spectral tau method developed in [15] for eigenvalue calculations. Our method is described in detail in [12]. In [17] computations of the norm of the resolvent (or equivalently the pseudospectra; see [17]) were performed for the Orr–Sommerfeld equation. There a Chebyshev spectral method developed in [6] was used. In [25] computation of the pseudospectra for the full Couette problem seems to have been made using the same method as in [17].

Some remarks on estimates of higher order space derivatives are stated in section 6.

2. Statement of the problem. Here we introduce the resolvent PDE. We also define the resolvent operator. (It is not obvious how to do this because of the special role of the pressure.) In section 2.2 we introduce notation for the norms we will use. These norms are constructed as a weighted combination of the norms of the components of the velocity field. The weights depend on the Reynolds number R.

2.1. The resolvent equation. We choose the coordinate system so that (nondimensionalized) Couette flow is given by

$$\mathbf{U} = \begin{pmatrix} x_2 \\ 0 \\ 0 \end{pmatrix} \tag{2.1}$$

in the domain

$$\Omega = \{\mathbf{x} \in \mathbb{R}^3 : -1 < x_2 < 1\}. \tag{2.2}$$

We use bold letters to denote vectors, and subscripts to identify the components, so x_2 is the second component of the vector $\mathbf{x}$.

The resolvent equation is derived by linearizing the Navier–Stokes equations at the flow (2.1) and then applying the Laplace transform. The result is

$$s\mathbf{u} + x_2 \frac{\partial \mathbf{u}}{\partial x_1} + \begin{pmatrix} u_2 \\ 0 \\ 0 \end{pmatrix} + \mathbf{grad}\, p = \frac{1}{R}\Delta \mathbf{u} + \mathbf{f}, \tag{2.3}$$

$$\operatorname{div} \mathbf{u} = 0,$$

with boundary conditions

$$\mathbf{u} = 0, \qquad \mathbf{x} \in \partial\Omega. \tag{2.4}$$

For this problem we are interested in the mapping of $\mathbf{f}$ to $\mathbf{u}$. This function is defined when the problem is uniquely solvable for $\mathbf{u}$. We introduce the notation

$$\mathcal{R}(s) : L^2(\Omega)^3 \to L^2(\Omega)^3, \qquad \mathcal{R}(s) : \mathbf{f} \to \mathbf{u}, \tag{2.5}$$

and call $\mathcal{R}$ the resolvent operator. According to the result in [18], $\mathcal{R}$ is well defined when $\operatorname{Re} s \geq -\delta/R$ for some $\delta > 0$. This means that all eigenvalues for Couette flow correspond to stable modes and have real part less than $-\delta/R$. The focus in this paper is on deriving bounds on the norm of $\mathcal{R}$ in the right half-plane ($\operatorname{Re} s \geq 0$) and tracking the R-dependence of these bounds.

The pressure has a special role in (2.3). It is not uniquely determined, but this is no problem from our point of view. See [11] and [27] for more information on the properties of the pressure. In [27, p. 68] a resolvent estimate, including the pressure, is given in Theorem 6.1. (For that result to hold, an additional condition on the pressure must of course be added to determine it uniquely.)

It suffices to study (2.3) with $\mathbf{f} \in C_0^\infty(\Omega)^3$ such that $\operatorname{div} \mathbf{f} = 0$. Less regular forcing can be treated by the standard closure argument for densely defined continuous operators, once the appropriate estimate has been derived in the smooth case. A forcing which is not solenoidal will be corrected by the pressure so that only the solenoidal part affects the velocity field [27].

2.2. Weighted norms. All norms used in this article are defined by inner products. The L^2-inner product for scalar functions has the following definition and notation:

$$(u, v) = \int_\Omega \overline{u}(x) v(x) dx, \qquad u, v \in L^2(\Omega).$$

We denote the complex conjugate of a function u by $\overline{u}$. For the L^2-inner product of vector functions we have the same notation:

$$(\mathbf{u}, \mathbf{v}) = \sum_{k=1}^{3} (u_k, v_k), \qquad \mathbf{u}, \mathbf{v} \in L^2(\Omega)^3.$$

Which of the above two inner products we have in mind will be clear from the arguments.

In this paper we will also consider norms with R-dependent weights of the components of $\mathbf{u}$. We define the following three modified norms:

$$\|\mathbf{u}\|_1^2 = \|u_1\|^2 + R^2 \|u_2\|^2 + R^2 \|u_3\|^2, \tag{2.6}$$

$$\|\mathbf{u}\|_2^2 = \|u_1\|^2 + R \|u_2\|^2 + \|u_3\|^2, \tag{2.7}$$

$$\|\mathbf{u}\|_3^2 = \|u_1\|^2 + R^2 \|u_2\|^2 + \|u_3\|^2. \tag{2.8}$$

We emphasize that subscripts on norms are always used to identify the different modified norms. The subscripts do NOT indicate p in the L_p-norms.

The modified norm (2.6) was suggested in [9], where it was conjectured that it leads to a more favorable dependence of the Reynolds number in the resolvent estimate. This turns out to be false, as we show in section 5. The two norms with subscripts 2 and 3, respectively, which only weight the second component of the velocity, turn out to be more suited for this purpose; see section 5.

3. The analytical estimate. Here we derive a resolvent estimate in a sector of the complex plane. The result is similar to that of [27, Chapter 1, section 5], the differences being that we track the Reynolds number-dependence of the estimate, we treat a specific flow and domain so that we can give the numerical value of all constants, and finally the domain in our case is unbounded; the result in [27] holds

for bounded domains. To simplify the presentation, we treat only the case $R \geq 1$. In the application to hydrodynamic stability we are interested in the behavior for large Reynolds numbers. When $R < 1$ we have a very viscous fluid. That simple situation is covered by a general and classical result of Serrin [21]; see also [20, section 5.6].

To derive our a priori estimate, we assume that $\mathbf{u}$, p is a smooth solution of (2.3). The estimate we derive will immediately imply uniqueness for the velocity field; the existence of a solution is established by standard methods [11], [22]. Now we scalar multiply the resolvent equation (2.3) with $\overline{\mathbf{u}}$ and integrate over Ω; this leads to

$$s\|\mathbf{u}\|^2 + \left(\mathbf{u}, x_2 \frac{\partial \mathbf{u}}{\partial x_1}\right) + (u_1, u_2) + (\mathbf{u}, \mathbf{grad}\, p) = \frac{1}{R}(\mathbf{u}, \Delta \mathbf{u}) + (\mathbf{u}, \mathbf{f}). \tag{3.1}$$

To simplify this equation, we use the following identities (easily derived by partial integration using the fact that $\mathbf{u} = 0$ on the boundary):

$$(\mathbf{u}, \mathbf{grad}\, p) = -(\operatorname{div} \mathbf{u}, p), \tag{3.2}$$

$$(\mathbf{u}, \Delta \mathbf{u}) = -\sum_{k=1}^{3} \left\| \frac{\partial \mathbf{u}}{\partial x_k} \right\|^2, \tag{3.3}$$

$$\operatorname{Re}\left(\mathbf{u}, x_2 \frac{\partial \mathbf{u}}{\partial x_1}\right) = 0. \tag{3.4}$$

In (3.2) we can use the fact that the velocity field is solenoidal, and thus the pressure disappears from (3.1). This simplification is unique for the L^2-norm.

We use (3.2)–(3.4) in (3.1) and take the real and imaginary parts of the resulting expression. After some rearrangement, we obtain

$$\operatorname{Re} s\|\mathbf{u}\|^2 + \frac{1}{R}\sum_{k=1}^{3} \left\| \frac{\partial \mathbf{u}}{\partial x_k} \right\|^2 = \operatorname{Re}\left[(\mathbf{u}, \mathbf{f}) - (u_1, u_2)\right], \tag{3.5}$$

$$\operatorname{Im} s\|\mathbf{u}\|^2 + \operatorname{Im}\left(\mathbf{u}, x_2 \frac{\partial \mathbf{u}}{\partial x_1}\right) = \operatorname{Im}\left[(\mathbf{u}, \mathbf{f}) - (u_1, u_2)\right]. \tag{3.6}$$

Using these two equations, we will derive the resolvent estimate of Theorem 3.1. Before we state the theorem, we define the following two sectors in the complex plane:

$$\Sigma_R = \left\{ s \in \mathbb{C} : \operatorname{Re} s - 3 + \frac{1}{2R}|\operatorname{Im} s| \geq 0 \right\},$$
$$\Sigma_0 = \{ s \in \mathbb{C} : \operatorname{Re} s - 3 - |\operatorname{Im} s| \geq 0 \}.$$

The sectors are plotted in Figure 3.1.

THEOREM 3.1. *If $s \in \Sigma_R$, then the solution of the resolvent equation satisfies*

$$\|\mathbf{u}\| \leq \frac{4\sqrt{2}R}{|s-3|}\|\mathbf{f}\|. \tag{3.7}$$

If, furthermore, $s \in \Sigma_0$, then we have the R-independent estimate

$$\|\mathbf{u}\| \leq \frac{\sqrt{2}}{|s-3|}\|\mathbf{f}\|. \tag{3.8}$$

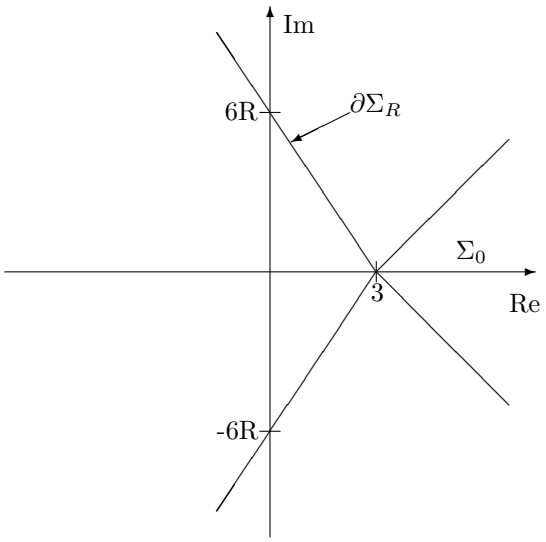


FIG. 3.1. *The sectors of the complex plane in which the estimates of Theorem* 3.1 *holds.*

Proof. Application of the Cauchy–Schwarz inequality and simple manipulations of (3.5) and (3.6) give the following inequalities:

$$(\operatorname{Re} s - 1)\,\|\mathbf{u}\|^2 + \frac{1}{R}\sum_{k=1}^{3}\left\|\frac{\partial \mathbf{u}}{\partial x_k}\right\|^2 \leq \|\mathbf{u}\|\,\|\mathbf{f}\|\,, \tag{3.9}$$

$$(|\operatorname{Im} s| - 1)\,\|\mathbf{u}\|^2 \leq \|\mathbf{u}\|\left\|\frac{\partial \mathbf{u}}{\partial x_1}\right\| + \|\mathbf{u}\|\,\|\mathbf{f}\|\,. \tag{3.10}$$

We first prove (3.8) and thus assume $s \in \Sigma_0$. We drop the derivative terms in (3.9) and cancel $\|\mathbf{u}\|$; this leads to

$$\|\mathbf{u}\| \leq \frac{1}{\operatorname{Re} s - 1}\,\|\mathbf{f}\|\,.$$

The following inequalities hold for $s \in \Sigma_0$:

$$\frac{1}{\operatorname{Re} s - 1} \leq \frac{1}{\operatorname{Re} s - 3} \leq \frac{\sqrt{2}}{|s-3|}.$$

We have now proven (3.8).

We now use the inequality $ab \leq a^2/4 + b^2$ on the first term in the right-hand side of (3.10). After some rearrangement, we obtain

$$\left(|\operatorname{Im} s| - \frac{5}{4}\right)\|\mathbf{u}\|^2 - \left\|\frac{\partial \mathbf{u}}{\partial x_1}\right\|^2 \leq \|\mathbf{u}\|\,\|\mathbf{f}\|\,.$$

We divide this inequality by R and add it to (3.9). This gives

$$\left(\operatorname{Re} s - 1 + \frac{1}{R}|\operatorname{Im} s| - \frac{5}{4R}\right)\|\mathbf{u}\|^2 \leq \left(1 + \frac{1}{R}\right)\|\mathbf{u}\|\,\|\mathbf{f}\|\,.$$

We use the fact that $R \geq 1$ and cancel $\|\mathbf{u}\|$. This leads to

$$\left(\operatorname{Re} s + \frac{1}{R}|\operatorname{Im} s| - \frac{9}{4}\right) \|\mathbf{u}\| \leq 2\,\|\mathbf{f}\|\,.$$

Using the defining inequality for Σ_R, we get

$$\|\mathbf{u}\| \leq \frac{4R}{|\operatorname{Im} s|}\,\|\mathbf{f}\|\,.$$

For $s \in \Sigma_R \setminus \Sigma_0$ we have

$$\frac{1}{|\operatorname{Im} s|} \leq \frac{\sqrt{2}}{|s-3|}.$$

This concludes the proof of (3.7) in $\Sigma_R \setminus \Sigma_0$. Since (3.8) implies (3.7) in Σ_0, the theorem is proven.

Remark. For the modified norms one cannot, in analogy with the above, take the modified inner product of $\mathbf{u}$ with the resolvent equation. The reason for this is that the pressure then is not eliminated. If one instead uses the above manipulations and the equivalence of norms,

$$c(R)\,\|\mathbf{u}\| \leq \|\mathbf{u}\|_X \leq C(R)\,\|\mathbf{u}\| \qquad (X = 1, 2 \text{ or } 3),$$

then a resolvent estimate is obtained in the same regions as in the theorem above. Now, however, the constant in the inequality will, asymptotically for large R, grow faster than in the case of the L^2-norm.

4. Transformation of the problem. The original problem (2.3) is an elliptic system coupled with the divergence "constraint." This is a PDE system with a very special structure. We shall instead consider the useful and well-known reformulation of the problem to one fourth order equation for the normal velocity u_2 and a second order equation for the normal vorticity. See [3] and [20]. Since the coefficients in the PDE do not depend on the x_1- and x_3-coordinates, the problem can be Fourier transformed in these variables. The Fourier transformed equation for the normal velocity is called the Orr–Sommerfeld equation. Following [20], we refer to the Fourier transformed equation for the normal vorticity as the Squire equation. We will also give an analytical estimate for the solution of the transformed problem in Lemma 4.1.

Let $\boldsymbol{\eta} = \mathbf{curl}\,\mathbf{u}$ and $\mathbf{g} = \mathbf{curl}\,\mathbf{f}$. The reduced and transformed system is

$$\begin{pmatrix} L_{OS} & 0 \\ -i\xi_3 & L_{SQ} \end{pmatrix} \begin{pmatrix} \hat{u}_2 \\ \hat{\eta}_2 \end{pmatrix} - s \begin{pmatrix} \hat{\Delta} & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \hat{u}_2 \\ \hat{\eta}_2 \end{pmatrix} = \begin{pmatrix} -\hat{\Delta}\hat{f}_2 \\ -\hat{g}_2 \end{pmatrix}. \tag{4.1}$$

Here we have introduced the differential operators

$$\begin{aligned} \hat{\Delta} &= \frac{\partial^2}{\partial x_2^2} - k^2, \\ L_{OS} &= \frac{1}{R}\hat{\Delta}^2 - i\xi_1 x_2 \hat{\Delta}, \\ L_{SQ} &= \frac{1}{R}\hat{\Delta} - i\xi_1 x_2 \end{aligned}$$

and the notation $k^2 = \xi_1^2 + \xi_3^2$, where ξ_1 and ξ_3 are the wave numbers in the x_1 and x_3 directions, respectively. The system (4.1) is supplemented with the boundary conditions

$$\hat{u}_2 = \hat{u}_2' = \hat{\eta}_2 = 0, \qquad x_2 = \pm 1. \tag{4.2}$$

The Fourier transforms of the remaining velocity components are given by

$$\hat{u}_1 = \frac{i\xi_1 \hat{u}_2' - i\xi_3 \hat{\eta}_2}{k^2}, \tag{4.3}$$

$$\hat{u}_3 = \frac{i\xi_3 \hat{u}_2' + i\xi_1 \hat{\eta}_2}{k^2}. \tag{4.4}$$

Considering ξ_1 and ξ_3 as parameters, the domain of the problem is $(-1, 1)$ instead of Ω. We will use the same notation for the L^2-norm in this case as in section 2.2. The domain we have in mind will be clear from the argument. For example, we have

$$\|\hat{\mathbf{u}}\|_1^2 = \int_{-1}^{1} \left(|\hat{u}_1|^2 + R^2 |\hat{u}_2|^2 + R^2 |\hat{u}_3|^2 \right) dx_2.$$

Using the expressions (4.3) and (4.4) for $\hat{u}_1$ and $\hat{u}_3$, we obtain for the Fourier transform of the velocity field

$$\|\hat{\mathbf{u}}\|^2 = \int_{-1}^{1} \left(|\hat{u}_2(x)|^2 + \frac{1}{k^2} |\hat{u}_2'(x)|^2 + \frac{1}{k^2} |\hat{\eta}_2(x)|^2 \right) dx. \tag{4.5}$$

This expression allows us to compute $\|\hat{\mathbf{u}}\|$ without explicitly determining $\hat{u}_1$ and $\hat{u}_3$.

From (4.1) we can derive estimates directly, for some parameter values, using only integration by parts, $\operatorname{div} \mathbf{f} = 0$, and the following Poincaré inequality:

$$\|\hat{\mathbf{u}}\| \leq 4 \|\hat{\mathbf{u}}'\|.$$

These estimates are collected in the following lemma.

LEMMA 4.1. *There is a constant C, independent of ξ_1, ξ_3, R, and $\operatorname{Re} s \geq 0$, such that*

$$\|\hat{\eta}_2\|^2 \leq \frac{CR^2}{1 + k^2} \left(\|\hat{u}_2\|^2 + \|\hat{g}_2\|^2 \right).$$

If at least one of the inequalities $|\xi_1| R \leq |\xi_3|^3$, $|\xi_1| R \leq 1/128$, or $\xi_1^2 + \xi_3^2 \geq R$ hold, then we also have the estimate

$$(1 + k^2) \|\hat{u}_2'\|^2 + \|\hat{u}_2\|^2 \leq \frac{CR^2}{1 + k^4} \|\hat{f}_2\|^2,$$

and it follows that there are constants C_0, C_2, and C_3 such that

$$\begin{aligned} \|\hat{\mathbf{u}}\| &\leq C_0 R^2 |\mathbf{f}\|, \\ \|\hat{\mathbf{u}}\|_2 &\leq C_2 R^{\frac{3}{2}} |\mathbf{f}\|_2, \\ \|\hat{\mathbf{u}}\|_3 &\leq C_3 R \|\mathbf{f}\|_3. \end{aligned}$$

We do not give the straightforward proof of the inequalities in Lemma 4.1.

In analogy with (2.5), we introduce notation for the mapping

$$\hat{\mathcal{R}}(s,\xi_1,\xi_3) : L^2(-1,1)^3 \to L^2(-1,1)^3, \qquad \hat{\mathcal{R}}(s,\xi_1,\xi_3) : \hat{\mathbf{f}} \to \hat{\mathbf{u}}.$$

For the L^2-norm we now show that $\|\mathcal{R}(s)\|$ can be determined by maximizing

$$\|\hat{\mathcal{R}}(s,\xi_1,\xi_3)\|$$

with respect to ξ_1 and ξ_3. For the modified norms this is also the case, and it is shown in a similar manner.

We need Plancherel's formula connecting the L^2-norm of the transformed function with that of the original unknown:

$$\|\mathbf{u}\|^2 = \frac{1}{4\pi^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \|\hat{\mathbf{u}}\|^2 \, d\xi_1 d\xi_3.$$

By definition, we have

$$\|\mathcal{R}(s)\|^2 = \sup_{\mathbf{f}\neq 0} \frac{\|\mathbf{u}\|^2}{\|\mathbf{f}\|^2}. \tag{4.6}$$

We also have

$$\|\hat{\mathbf{u}}\| = \|\hat{\mathcal{R}}(s,\xi_1,\xi_3)\hat{\mathbf{f}}\| \le \left(\sup_{\xi_1,\xi_3} \|\hat{\mathcal{R}}(s,\xi_1,\xi_3)\|\right) \|\hat{\mathbf{f}}\|. \tag{4.7}$$

Using (4.7) in Plancherel's formula and then (4.6), we obtain the inequality

$$\|\mathcal{R}(s)\| \le \sup_{\xi_1,\xi_3} \|\hat{\mathcal{R}}(s,\xi_1,\xi_3)\|. \tag{4.8}$$

Now we will show that there is actually equality in (4.8). This is done by constructing a family of functions $\mathbf{f}_\epsilon$, with corresponding solutions $\mathbf{u}_\epsilon$, such that

$$\lim_{\epsilon\to 0} \frac{\|\mathbf{u}_\epsilon\|}{\|\mathbf{f}_\epsilon\|} = \sup_{\xi_1,\xi_3} \|\hat{\mathcal{R}}(s,\xi_1,\xi_3)\|.$$

We denote by $\xi_1^{(0)}$ and $\xi_3^{(0)}$ the values for which the maximum of the norm occurs,[1] and by $\hat{\mathbf{f}}_0$ the corresponding forcing, and thus we have

$$\sup_{\xi_1,\xi_3} \|\hat{\mathcal{R}}(s,\xi_1,\xi_3)\| = \|\hat{\mathcal{R}}(s,\xi_1^{(0)},\xi_3^{(0)})\hat{\mathbf{f}}_0\|.$$

We will use a family of cut-off functions $\varphi_\epsilon \in C_0^\infty(\mathbb{R}^2)$ which satisfy

$$\begin{aligned} \varphi_\epsilon &\ge 0, \\ \int_{\mathbb{R}^2} \varphi_\epsilon^2 dx &= 1, \\ \varphi_\epsilon(\xi_1,\xi_3) &= 0 \qquad \text{if } (\xi_1-\xi_1^{(0)})^2 + (\xi_3-\xi_3^{(0)})^2 \ge \epsilon. \end{aligned}$$

[1] That the supremum is attained follows from the continuity and the decay for large $\xi_1^2+\xi_3^2$, which follow from the estimates of Lemma 4.1.

Now we take as our function $\mathbf{f}_\epsilon$ the inverse Fourier transform of

$$\hat{\mathbf{f}}_\epsilon(\xi_1, x_2, \xi_3) = \hat{\mathbf{f}}_0(x_2)\varphi_\epsilon(\xi_1, \xi_3).$$

We have the following:

$$\begin{aligned}&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \|\hat{\mathbf{u}}_\epsilon(\cdot, \xi_1, \xi_3)\|^2 \, d\xi_1 d\xi_3 \\ &\quad = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \varphi_\epsilon^2 \|\hat{\mathcal{R}}(s, \xi_1, \xi_3)\hat{\mathbf{f}}_0\|^2 d\xi_1 d\xi_3 \to \|\mathcal{R}(s, \xi_1^{(0)}, \xi_3^{(0)})\hat{\mathbf{f}}_0\|^2\end{aligned}$$

when $\epsilon \to 0$. Using the definition of $\|\mathcal{R}\|$ and Plancherel's formula, we see that we can obtain $\|\mathcal{R}\|$ by maximizing $\|\hat{\mathcal{R}}\|$ over ξ_1 and ξ_3:

$$\|\mathcal{R}(s)\| = \max_{\xi_1, \xi_3} \|\hat{\mathcal{R}}(s, \xi_1, \xi_3)\|. \tag{4.9}$$

5. Numerical results. In this section we numerically determine C and γ in the expression

$$\sup_{\operatorname{Re} s \geq 0} \|\mathcal{R}(s)\| \approx CR^\gamma. \tag{5.1}$$

This is done for the four norms introduced in section 2.2. For this optimization we use (4.9) and maximize

$$\|\hat{\mathcal{R}}(s, \xi_1, \xi_3)\| \tag{5.2}$$

over $\operatorname{Re} s \geq 0$, $\xi_1 \geq 0$, $\xi_3 \geq 0$.

The numerical calculations of the resolvent norm are done by discretizing (4.1) and (4.2) with a spectral Chebyshev tau method. The method is based on the classical paper [15]. In [12] our method is described in detail.

Because of symmetry, it is sufficient to search the first quadrant of the ξ_1-ξ_3 plane. The resolvent cannot have a local maximum (in the resolvent set); this maximum principle is given in [4, p. 230]. The decay estimate of Theorem 3.1 in the sector Σ_R, combined with the maximum principle, implies that the maximization of $\|\hat{\mathcal{R}}(s, \xi_1, \xi_3)\|$ can be restricted to $\operatorname{Re} s = 0$. We thus have to maximize the function over three real variables. However, as was noted in [25] and [9] in the case of the L^2-norm, the maximum seems always to occur for $s = 0$. We believe that this is the case for all norms considered in this paper, and in section 5.1 we present results supporting this fact.

In sections 5.2 and 5.3 we determine the values of C and γ in (5.1) for all four different norms. For all the norms except $\|\cdot\|_1$ the maximum of (5.2) occurs where the second part of Lemma 4.1 is applicable. The exponents of R in the computed resolvents agree with the resolvent bounds in the lemma.

5.1. The maximum occurs for $s = 0$. In this section we present computational results supporting the assumption

$$\sup_{\operatorname{Re} s = 0, \xi_1, \xi_3} \|\hat{\mathcal{R}}(s, \xi_1, \xi_3)\| = \sup_{\xi_1, \xi_3} \|\hat{\mathcal{R}}(0, \xi_1, \xi_3)\|.$$

We have performed extensive computations. In Figure 5.1 we see typical results. For a fixed Reynolds number we have plotted $\|\hat{\mathcal{R}}\|_3$ as a function of σ when $s = i\sigma$ for

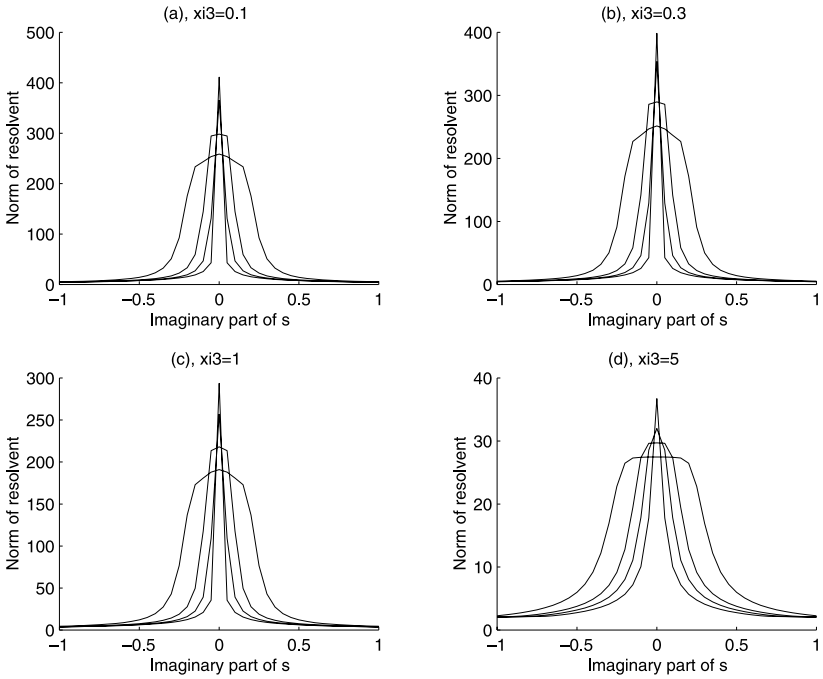


FIG. 5.1. *The value of* $\|\mathcal{R}(i\sigma, \hat{\xi}_1, \xi_3)\|_3$ *for* $R = 1000$. *For each plot,* ξ_3 *is fixed to the value shown above the plot. The four different curves in each plot correspond to* $\xi_1 = 0, 0.1, 0.2,$ *and* 0.4, *respectively. On the horizontal axis we have the imaginary part of* s *for* $\operatorname{Re} s = 0$. *All curves exhibit the typical behavior, with maximum at* $s = 0$. *The peak in plot* (a) *is near the global maximum; compare Table* 5.3.

some different values of ξ_1 and ξ_3. The third modified norm turns out to be the most interesting for our purposes; see section 5.3.

The corresponding computations for other Reynolds numbers and the other norms of this paper have been made with the same result: The maximum occurs for $s = 0$. The assumption is also supported by the fact that the peak at $s = 0$ is so sharp. Furthermore, only a small range of ξ_1 and ξ_3 give values of $\|\hat{\mathcal{R}}\|$ near the peak at $s = 0$; the norm decays quickly for other choices of ξ_1 and ξ_3.

5.2. The L^2- and first modified norms. For the standard norm, the following R-dependence was established in [9]:

$$\max_{\operatorname{Re} s \geq 0} \mathcal{R}(s) = 0.0152R^2. \tag{5.3}$$

From Table 5.1 we conclude that our computations yield the same result. In Figure 5.2 we show $\|\mathcal{R}(0, \hat{\xi}_1, \xi_3)\|$ as a function of ξ_1 and ξ_3. As was noted in [9], the maximum occurs for $\xi_1 = 0$, where the second part of Lemma 4.1 is applicable. Figure 5.2 can be compared with Figure 1 in [9]. There the resolvent, scaled by $k/(\xi_3 R^2)$, is plotted as a function of $\xi_1 R$ and k for $R = 500, 1000$. For a given R a simple change of variables connects $\xi_1 R$ and k to ξ_1 and ξ_3. Note that, along the vertical axis, where $\xi_1 = \xi_1 R = 0$, $k = \xi_3$. As expected, the contour lines of the two figures are very similar.

The object when constructing the modified norms is to find a norm where the normality of the problem does not degenerate as R increases. Such a norm would

TABLE 5.1

The norm of the resolvent for different Reynolds numbers in the L^2-norm and the first modified norm. The value of ξ_3 for which the maximum occurs is also given. The ξ_1-coordinate for the maximum is zero. The values are determined with three significant digits.

L^2-norm

R	ξ_3	$\|\mathcal{R}\|$
350	1.18	1860
1000	1.18	$1.52 \cdot 10^4$
3500	1.18	$1.86 \cdot 10^5$
10^4	1.18	$1.52 \cdot 10^6$

1st modified norm

R	ξ_3	$\|\mathcal{R}\|_1$
350	0	$2.45 \cdot 10^4$
1000	0	$2.00 \cdot 10^5$
3500	0	$2.45 \cdot 10^6$
10^4	0	$2.01 \cdot 10^7$

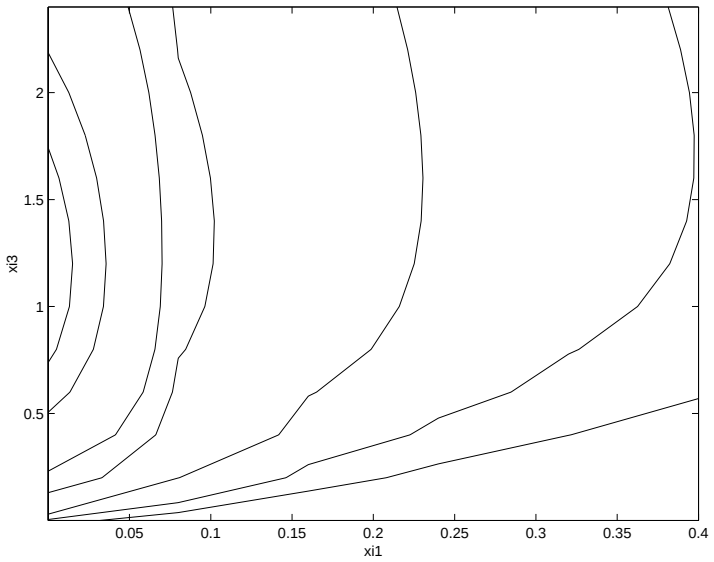


FIG. 5.2. *Contour plot of $\|\hat{\mathcal{R}}(0, \xi_1, \xi_3)\|$ (L^2-norm) for $R = 1000$. The contour lines represent values in the range 300 to 13000. The maximum occurs along the ξ_3-axis.*

yield a smaller exponent γ than in (5.3), where it is two. The first modified norm was suggested in [9] for this purpose. As we show below, however, the exponent is still two in this case.

In Table 5.1 we present $\|\hat{\mathcal{R}}\|_1$ for some values of the Reynolds number. From this table the values of C and γ in (5.3) can be reproduced by a least squares fit, and for the first modified norm we obtain

$$\|\mathcal{R}\|_1 = 0.198 R^{2.00}.$$

We see that the first modified norm gives the same exponent and a larger (worse) constant than the L^2-norm.

There are some difficulties in the investigation for the first modified norm. The maximum occurs for $\xi_1, \xi_3 \to 0$, as can be seen in Figure 5.3. Moreover, we get different limits depending on how we let ξ_1 and ξ_3 tend to zero. To find the maximum, we approach the origin along the ray $\xi_1 = \xi_3$. In Figure 5.4 we show the norm of $\hat{\mathcal{R}}$ on this ray. The peak is sharp, but it is clear that the limiting value is $2.00 \cdot 10^5$ (for $R = 1000$) as is shown in Table 5.1.

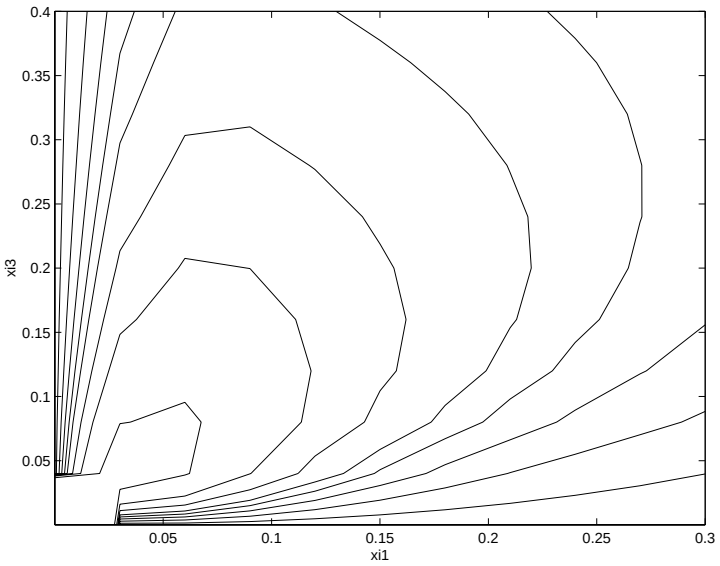


FIG. 5.3. *Contour plot of* $\|\hat{\mathcal{R}}(0,\xi_1,\xi_3)\|_1$ *for* $R = 1000$. *The contour lines represent values in the interval* $(1000, 17000)$. *The region around the origin (*$\xi_1 \leq 0.05$ *and* $\xi_3 \leq 0.05$*) is not resolved. A high peak is located around the origin; as shown in Table* 5.1, *the maximum is* $2 \cdot 10^5$. *See Figure* 5.4 *for the limiting behavior at the origin.*

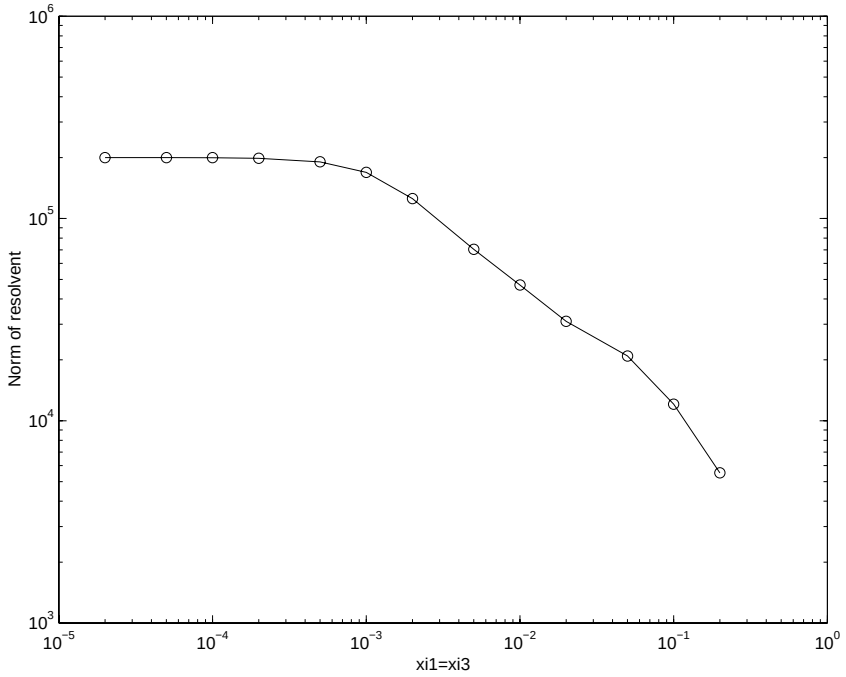


FIG. 5.4. *The values of* $\|\hat{\mathcal{R}}(0,\xi_1,\xi_3)\|_1$ *for* $\xi_1 = \xi_3$ *and* $R = 1000$. *In this logarithmic plot it is evident that the limiting value at* $\xi_1 = \xi_3 = 0$ *is* $2 \cdot 10^5$.

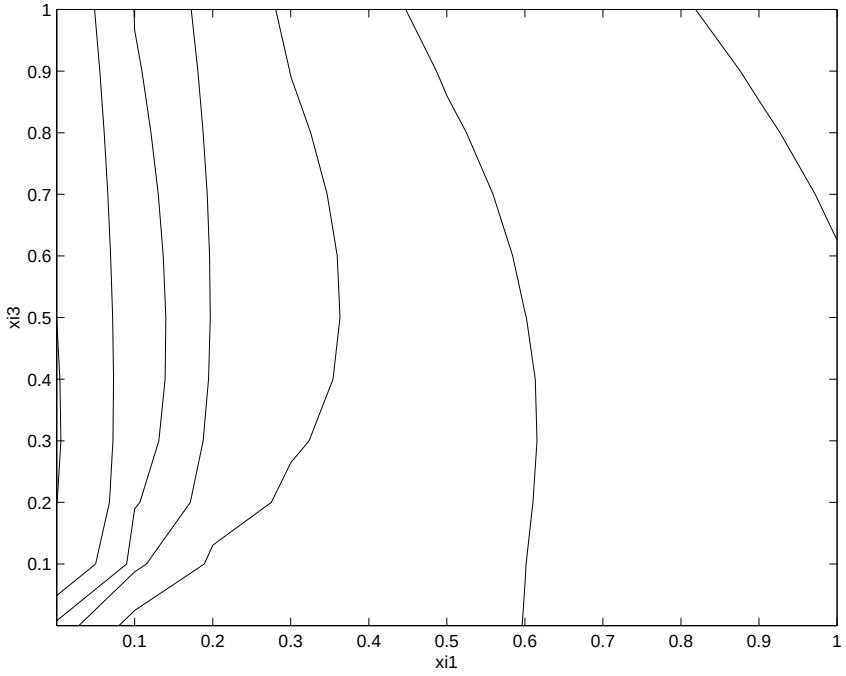


FIG. 5.5. *The values of* $\|\hat{\mathcal{R}}(0,\xi_1,\xi_3)\|_2$ *for* $R = 1000$ *in the* ξ_1*-*ξ_3 *plane. The contour lines represent values in the range* 150 *to* 2000*.*

5.3. The second and third modified norms. In this section we present our investigations of the second and third modified norms which achieve the goal and yield lower (better) exponents γ. In Figures 5.5 and 5.6 we see the plots in the ξ_1-ξ_3 plane for the second and third modified norms, respectively. In both cases the maximum occurs along the ξ_3 axis, where the second part of Lemma 4.1 is applicable. In Table 5.2 we list the values of $\|\mathcal{R}\|_2$ for different values of the Reynolds number. From this table we obtain

$$\|\mathcal{R}\|_2 = 0.056R^{1.52}.$$

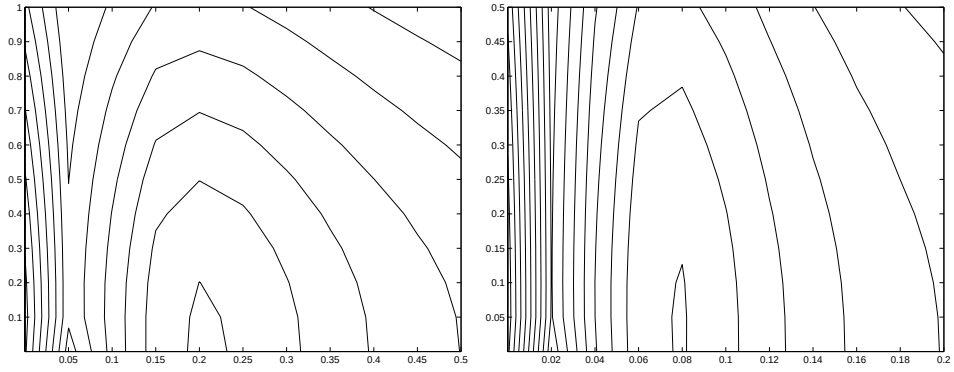


FIG. 5.6. *The values of* $\|\hat{\mathcal{R}}(0,\xi_1,\xi_3)\|_3$*. In the left panel we have* $R = 350$*, and in the right,* $R = 1000$*. The global maximum is along the* ξ_3 *(vertical) axis. Observe the local maximum on the* ξ_1*-axis. See Table* 5.3 *for location and values of the global and local maxima.*

TABLE 5.2

The norm of the resolvent for different Reynolds numbers for the second norm. The value of ξ_3 for which the maximum occurs is also given. The ξ_1-coordinate for the maximum is zero.

2nd modified norm

R	ξ_3	$\|\mathcal{R}\|_2$
350	0.40	420
1000	0.31	2080
3500	0.24	$1.41 \cdot 10^4$
10^4	0.18	$6.92 \cdot 10^4$

TABLE 5.3

The third modified norm. In the first table we have the norm of the resolvent for different Reynolds numbers. The maximum occurs for $\xi_1 = 0$. In the second table we have the local maxima which occur for $\xi_3 = 0$ and the ξ_1-values listed.

Global maxima

R	ξ_3	$\|\mathcal{R}\|_3$
350	0.046	145
1000	0.028	413
3500	0.016	1450
10000	0.009	4130

Local maxima

R	ξ_1	$\|\hat{\mathcal{R}}\|_3$
350	0.20	132
1000	0.07	383
3500	0.02	1340
10^4	0.01	3840

In Figure 5.6 we see that, in addition to the global maximum on the ξ_3 axis, there is a local maximum on the ξ_1 axis. In Table 5.3 we give the value and locations of the global and local maxima. For increasing Reynolds number both maxima move along the respective axes towards the origin. From the values in Table 5.3 we obtain

$$\|\mathcal{R}\|_3 = 0.413 R^{1.00}, \tag{5.4}$$

in agreement with Lemma 4.1. Note that here we have the optimal exponent $\gamma = 1.00$. The constant and the exponent in (5.4) are determined by a least squares fit of the model function CR^γ to the four values in Table 5.3. The agreement of the curve (5.4) with the computed values is very good.

6. Remarks on regularity. Using a resolvent estimate of the type (1.1) or (1.3), it is possible to derive an elliptic regularity estimate

$$\|\mathbf{u}\|_{H^2} \leq C(R) \|\mathbf{f}\| ,$$

which holds uniformly for $\operatorname{Re} s \geq 0$. Here we have the usual Sobolev H^2-norm in the left-hand side,

$$\|\mathbf{u}\|_{H^2}^2 \stackrel{def}{=} \sum_{|\alpha| \leq 2} \left\| \frac{\partial^\alpha \mathbf{u}}{\partial x^\alpha} \right\|^2 ,$$

where α denotes a multi-index. This will lead to a constant $C(R)$ which asymptotically grows faster than R^γ. Here $\gamma = 2$ for the L^2-norm, and $\gamma = 1$ for the third modified norm. More interesting from our point of view of nonlinear stability is to have an estimate of some second derivatives satisfying the following two criteria. First, the constant in the estimate should be proportional to R^γ. Second, via Sobolev inequalities, it should be possible to bound the supremum norm with the combination

of second and lower order derivatives. One possibility for doing this, used in [9], is

$$\|\mathbf{u}\|_{\tilde{H}}^2 \stackrel{def}{=} \|\mathbf{u}\|^2 + \frac{1}{R}\sum_{k=1}^{3}\left\|\frac{\partial \mathbf{u}}{\partial x_k}\right\|^2 + \frac{1}{R^2}\left(\left\|\frac{\partial^2 \mathbf{u}}{\partial x_1^2}\right\|^2 + \left\|\frac{\partial^2 \mathbf{u}}{\partial x_1 \partial x_2}\right\|^2 + \left\|\frac{\partial^2 \mathbf{u}}{\partial x_2 \partial x_3}\right\|^2 + \left\|\frac{\partial^2 \mathbf{u}}{\partial x_3^2}\right\|^2\right),$$

which bounds the supremum norm according to

$$\sup_{\mathbf{x}\in\Omega} |\mathbf{u}(\mathbf{x})| \leq C_s R^{3/4} \|\mathbf{u}\|_{\tilde{H}};$$

see [10, Appendix 3]. As is shown in [9], this norm also satisfies the first criterion above,

$$\|\mathbf{u}\|_{\tilde{H}} \leq C R^2 \|\mathbf{f}\|.$$

In [13] a variation of this approach is used for the third modified norm.

7. Conclusions. The analytical resolvent estimate of Theorem 3.1 exhibits the expected property, deteriorating as the Reynolds number increases. The sector in which it holds shrinks, and the constant grows. However, in the sector Σ_0, which does not depend on the Reynolds number, the estimate is independent of the Reynolds number.

The result of Theorem 3.1 does not cover the entire unstable half-plane; still, it is known that the resolvent is bounded there. To improve the estimate with analytical techniques, it may be fruitful to consider the reformulated problem of section 4. One such result is Lemma 4.1. It is the reformulated problem which is studied in the investigations of the spectrum referred to in the introduction. We believe, however, that deriving a completely analytical resolvent estimate in the entire unstable half-plane would be extremely complicated.

We therefore employ numerical methods, which yield the estimate implying stability, as was found in [9]. The main result of our computations is the optimal exponent 1.00 of the Reynolds number in (5.4). This is obtained by weighting the second component of the velocity field with a coefficient that depends on the Reynolds number. In [13] we use (5.4) to improve the theoretical bound on the perturbation threshold, as discussed in the introduction.

REFERENCES

[1] S. Chandrasekhar, *Hydrodynamic and Hydromagentic Stability*, International Series of Monographs on Physics, Clarendon, Oxford, 1961.
[2] S. J. Chapman, *Subcritical transition in channel flows*, J. Fluid Mech., 451 (2002), pp. 35–97.
[3] P. G. Drazin and W. H. Reid, *Hydrodynamic Stability*, Cambridge University Press, London, 1982.
[4] N. Dunford and J. Schwartz, *Linear Operators, Part* I, Pure Appl. Math. 7, Interscience, New York, 1958.
[5] A. P. Gallagher and A. M. Mercer, *On the behaviour of small disturbances in plane Couette flow*, J. Fluid Mech., 13 (1962), pp. 91–100.
[6] T. Herbert, *Die Neutrale Fläche der Ebenen Poiseuille-Strömung*, Habilitationsschrift, Universität Stuttgart, Stuttgart, Germany, 1977.
[7] D. D. Joseph, *Eigenvalue bounds for the Orr-Sommerfeld equation*, J. Fluid Mech., 33 (1968), pp. 617–621.

[8] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, 1966.
[9] G. Kreiss, A. Lundbladh, and D. S. Henningson, *Bounds for threshold amplitudes in subcritical shear flow*, J. Fluid Mech., 270 (1994), pp. 175–198.
[10] H.-O. Kreiss and J. Lorenz, *Initial-Boundary Value Problems and the Navier–Stokes Equations*, Pure Appl. Math. 136, Academic Press, Boston, 1989.
[11] O. A. Ladyzhenskaya, *The Mathematical Theory of Viscous Incompressible Flow*, 2nd ed., Mathematics and Its Applications, Vol. 2, Gordon and Breach, New York, 1969.
[12] M. Liefvendahl, *Stability Results for Viscous Shock Waves and Plane Couette Flow*, Thesis TRITA-NA-0132, Royal Institute of Technology, Stockholm, 2001.
[13] M. Liefvendahl and G. Kreiss, *Bounds of the threshold amplitude for plane Couette flow*, J. Nonlinear Math. Phys., 9 (2002), pp. 311–324.
[14] C. C. Lin, *The Theory of Hydrodynamical Stability*, Cambridge Monogr. Mech. Appl. Math., Cambridge University Press, London, 1955.
[15] S. A. Orszag, *Accurate solution of the Orr–Sommerfeld stability equation*, J. Fluid Mech., 50 (1971), pp. 689–703.
[16] S. C. Reddy, P. J. Schmid, J. S. Baggett, and D. S. Henningson, *On stability of streamwise streaks and transition thresholds in plane channel flows*, J. Fluid Mech., 365 (1998), pp. 269–303.
[17] S. C. Reddy, P. J. Schmid, and D. S. Henningson, *Pseudospectra of the Orr–Sommerfeld operator*, SIAM J. Appl. Math., 53 (1993), pp. 15–47.
[18] V. A. Romanov, *Stability of plane-parallel Couette flow*, Funct. Anal. Appl., 7 (1973), pp. 137–146.
[19] H. Schlichting, *Berechnung der anfachung kleiner störungen bei der plattenströmung*, ZAMM Z. Angew. Math. Mech., 13 (1933), pp. 171–174.
[20] P. J. Schmid and D. S. Henningson, *Stability and Transition in Shear Flows*, Appl. Math. Sci. 142, Springer, New York, 2001.
[21] J. Serrin, *Mathematical principles of classical fluid mechanics*, in Handbuch der Physik Vol. VIII/1, Springer-Verlag, Berlin, 1959, pp. 125–263.
[22] R. Temam, *Navier–Stokes Equations, Theory and Numerical Analysis*, North–Holland, Amsterdam, 1984.
[23] W. Tollmien, *General instability criterion of laminar velocity distributions*, Tech. Memor. Nat. Adv. Comm. Aero., 792 (1936).
[24] L. N. Trefethen, S. J. Chapman, D. S. Henningson, Á. Meseguer, T. Mullin, and F. T. M. Nieuwstadt, *Threshold Amplitudes for Transition to Turbulence in a Pipe*, Tech. report NA-00/17, Oxford University Computing Laboratory, Oxford, UK, 2000.
[25] L. N. Trefethen, A. E. Trefethen, S. C. Reddy, and T. A. Driscoll, *Hydrodynamic stability without eigenvalues*, Science, 261 (1993), pp. 578–584.
[26] W. Wasow, *On small disturbances in plane Couette flow*, J. Research Nat. Bur. Standards, 51 (1953), pp. 195–202.
[27] V. I. Yudovich, *The Linearization Method in Hydrodynamical Stability Theory*, Trans. Math. Monogr. 74, American Mathematical Society, Providence, RI, 1989.

SIAM J. APPL. MATH.
Vol. 63, No. 3, pp. 818–833

CONGESTION ON MULTILANE HIGHWAYS*

J. M. GREENBERG†, A. KLAR‡, AND M. RASCLE§

Abstract. We present a new model for traffic on a multilane freeway (with n lanes). Our basic descriptors are the car density ρ (in cars/mile), taken across all lanes in the freeway, and the average car velocity u (in miles/hour). The flux of cars across all lanes is given by $\rho u = \sum_{i=1}^{n} \rho_i u_i$, where ρ_i is the car density in the ith lane, and u_i the velocity of cars in the ith lane. We shall track only ρ and u and not what is going on in each individual lane.

On such multilane freeways, one often observes distinct stable equilibrium relationships between car velocity and density. Prototypical situations involve two equilibria,

$$v = v_1(\rho) > v = v_2(\rho), \quad 0 \le \rho < \rho_{\max},$$

where $v_1(\cdot)$ and $v_2(\cdot)$ are monotone decreasing and satisfy $v_1(\rho_{\max}) = v_2(\rho_{\max}) = 0$. The upper curve is typically stable for densities satisfying $0 \le \rho \le \rho_1$, whereas the lower curve is stable for densities satisfying $\rho_2 \le \rho \le \rho_{\max}$. Our interest is in the situation where $0 < \rho_2 \le \rho_1 < \rho_{\max}$ and $v_2(\rho_2) \le v_1(\rho_1)$.

In this paper we present a model that incorporates both equilibrium curves and a simple switching mechanism which allows cars to transit from one equilibrium curve to the other. This switching mechanism, when combined with the continuity equation, produces relaxation or self-excited oscillations in the system, and these oscillations are what interests us here.

Key words. microscopic and macroscopic traffic models, multiple equilibria, self-excited oscillations, travelling waves

AMS subject classifications. 35L45, 90B20

PII. S0036139901396309

1. Introduction. In this paper we present a new model for traffic on a multilane freeway with n lanes. Our basic descriptors are the car density ρ (in cars/mile), taken across all lanes in the freeway, and the average car velocity u (in miles/hour). The flux of cars across all lanes is given by $\rho u = \sum_{i=1}^{n} \rho_i u_i$, where ρ_i is the car density in the ith lane, and u_i the velocity of cars in the ith lane. We shall track only ρ and u and not what is going on in each individual lane. This model simplification will ultimately yield a one-dimensional model.

On such multilane freeways, one often observes distinct stable equilibrium relationships between auto velocity and density. Prototypical situations involve two equilibria,

$$v = v_1(\rho) > v = v_2(\rho), \quad 0 \le \rho < \rho_{\max}, \tag{1.1}$$

where $v_1(\cdot)$ and $v_2(\cdot)$ are monotone decreasing and satisfy $v_1(\rho_{\max}) = v_2(\rho_{\max}) = 0$. The upper curve is typically stable for densities satisfying $0 \le \rho \le \rho_1$, whereas the

*Received by the editors October 10, 2001; accepted for publication (in revised form) July 6, 2002; published electronically January 17, 2003.

http://www.siam.org/journals/siap/63-3/39630.html

†Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA 15213-3890 (greenber@andrew.cmu.edu). This research was partially supported by the Applied Mathematical Sciences Program, the U.S. Department of Energy, and the U.S. National Science Foundation.

‡Fachbereich Mathematik, Technische Universität Darmstadt, Schlossgarten Strasse 7, 64289 Darmstadt, Germany (klar@mathematik.tu-darmstadt.de.) This research was partially supported by the German research foundation (DFG).

§Laboratoire J.A. Dieudonné, UMR CNRS No. 6621, Université de Nice, Parc Valrose, F-06108, Nice Cedex 02, France (rascle@math.unice.fr). This research was partially supported by the CNRS-NSF.

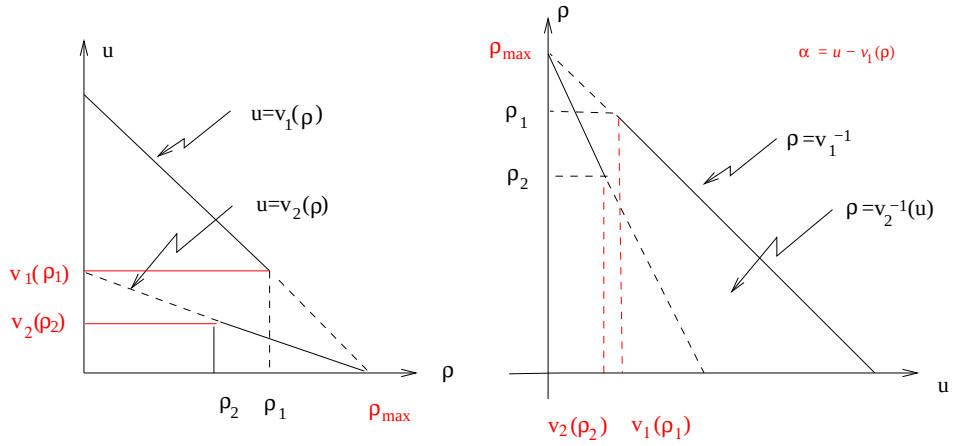


FIG. 1.1. *Two equilibrium curves $v_1(\rho)$ and $v_2(\rho)$.*

lower curve is stable for densities satisfying $\rho_2 \le \rho \le \rho_{\max}$. Our interest is in the situation where $0 < \rho_2 \le \rho_1 < \rho_{\max}$ and $v_2(\rho_2) \le v_1(\rho_1)$; see Figure 1.1.

The explanation for the two curves is quite simple. For high density congested traffic, lane changing and passing is difficult and dangerous, and this yields the slower equilibrium curve. On the other hand, when the traffic is less dense, lane changing and passing becomes easier, and this yields the faster equilibrium curve.

In this paper we present a model that incorporates both equilibrium curves and a simple switching mechanism which allows cars to transit from one equilibrium curve to the other.

Once again, our basic descriptors are the car density ρ and velocity u. We also track

$$\alpha = u - v_1(\rho),$$

which represents the discrepancy between the actual car speed and the uncongested equilibrium speed.

Our governing equations are

$$\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho u) = 0 \tag{1.2}$$

and

$$\frac{\partial \alpha}{\partial t} + u\frac{\partial \alpha}{\partial x} = \begin{cases} \dfrac{-\alpha}{\epsilon}, & \rho < R(u), \\ \dfrac{((v_2 - v_1)(\rho) - \alpha)}{\epsilon}, & \rho \ge R(u). \end{cases} \tag{1.3}$$

Here, $u \to R(u)$ is a monotone nondecreasing function defined on $0 \le u$ and satisfying

$$R(u) = \rho_2, \qquad 0 \le u \le v_2(\rho_2), \qquad \text{and} \quad R(u) = \rho_1,\ v_1(\rho_1) \le u; \tag{1.4}$$

see Figure 1.2.

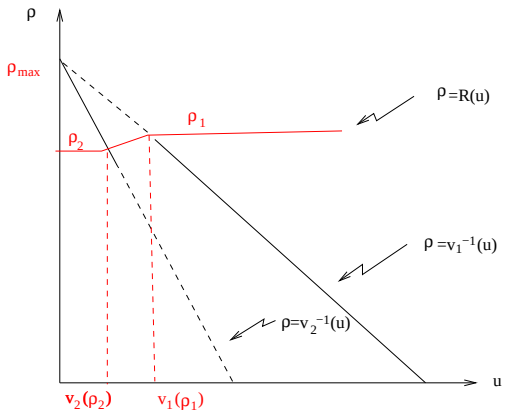


FIG. 1.2. *Switching curve* $R(u)$.

For experimental data and the choice of the switch curve, we refer to the work of Kerner [7, 8]. In his thesis, Sopasakis [9] gave an argument supporting the choice $\rho_2 = \rho_1$ and $R(u) \equiv \rho_2,\ 0 \le u$.

The motivation for system (1.2), (1.3) is as follows:

1. When there is no source term, i.e., the right-hand side of (1.3) is set to zero, our model is the one introduced in [3]. This model turns out to be the rigorous hydrodynamic limit of the microscopic follow-the-leader system (1.20)–(1.23) (see below) with no right-hand side; see [2] and also [10].
2. In the case with a source term of the form $-1/\epsilon(V(\rho) - u)$, the above result remains true; for details, see [6] and [2].
3. At least formally, the system we propose to study here is the limit of the microscopic system (1.23), when the size of cars goes to zero.

We note that (1.2) and (1.3) imply that u satisfies

$$\frac{\partial u}{\partial t} + (u + \rho v_1'(\rho))\frac{\partial u}{\partial x} = \begin{cases} \dfrac{v_1(\rho) - u}{\epsilon}, & \rho < R(u), \\ \dfrac{v_2(\rho) - u}{\epsilon}, & \rho \ge R(u). \end{cases} \tag{1.5}$$

One motivation for the switching mechanism hypothesized here is as follows. We assume that there are two natural modes in which drivers can operate. The first is the fast mode and is characterized by the equilibrium curve $\rho \to v_1(\rho)$, and the second is the slow mode characterized by the slow curve $\rho \to v_2(\rho)$. What we are hypothesizing in (1.5) is that if the current state of traffic, (u, ρ), lies below the switch curve $\rho = R(u)$, drivers' preferences will migrate towards the fast curve $u = v_1(\rho)$, whereas if the traffic state, (u, ρ), lies above $\rho = R(u)$, their preferences will migrate towards the slow curve $u = v_2(\rho)$.

An alternative approach would be to hypothesize that, for all densities $0 \le \rho \le \rho_{\max}$, the preferred state of an average driver is characterized by the homogenized equilibrium curve

$$v(\rho) = a(\rho)v_1(\rho) + (1 - a(\rho))v_2(\rho),$$

where $a(0) = 1$, $a'(\rho) \le 0$ for $0 \le \rho \le \rho_{\max}$, and $a(\rho_{\max}) = 0$. This latter approach has been used in multiclass models of traffic flow; see, for instance, [4, 5] and many other references.

For $0 < \rho \le \rho_{\max}$, the system (1.2), (1.4), (1.5) is strictly hyperbolic, with distinct wave speeds $c_1 = u + \rho v_1'(\rho) < c_2 = u$. Variants of this relaxation model with one equilibrium and no switch curve have been studied by Aw and Rascle [3], Argall et al. [1], Greenberg [6], and Aw et al. [2]. The principal results of those investigations relevant to us here are that for any initial data $\rho_0(\cdot)$ and $u_0(\cdot)$ satisfying

$$0 \le u_0(x) \le v_1(\rho_0(x)) \quad \text{and} \quad 0 \le \rho_0(x) \le \rho_{\max} \tag{1.6}$$

the system (1.2), (1.4), (1.5) has an appropriately defined weak solution satisfying (1.6) for all future times. Thus the model presented here has no signals propagating faster than the car velocities and yields none of the velocity reversals seen in the Payne–Whitham models. These two observations are the basic strength of this class of second-order model.

For simplicity, we restrict our attention to spatially periodic solutions—the ring road scenario. We shall also work with a Lagrangian reformulation of the system. When discretized, this Lagrangian system yields a follow-the-leader–type model.

We let l be the spatial period of our data $\rho_0(\cdot) > 0$ and assume that

$$\int_0^l \rho_0(\xi)d\xi = M \tag{1.7}$$

is an integer. For any real number $m \in [0, M]$ we let $x^0(m)$ be the unique solution of

$$m = \int_0^{x^0(m)} \rho_0(\xi)d\xi \tag{1.8}$$

and $x(m,t)$ be the solution of

$$\frac{\partial x}{\partial t}(m,t) = \bar{u}(m,t) \stackrel{def}{=} u(x(m,t),t) \quad \text{and} \quad x(m,0) = x^0(m). \tag{1.9}$$

Here, ρ and u are solutions of (1.2), (1.4), and (1.5). The continuity equation (1.2), when combined with (1.8) and (1.9), yields

$$m = \int_{x(0,t)}^{x(m,t)} \rho(\xi,t)d\xi, \tag{1.10}$$

and (1.10) in turn implies that

$$\bar{\rho}(m,t) \stackrel{def}{=} \rho(x(m,t),t) \quad \text{and} \quad \bar{\gamma}(m,t) \stackrel{def}{=} \frac{\partial x}{\partial m}(m,t) \tag{1.11}$$

satisfy

$$\bar{\rho}(m,t)\bar{\gamma}(m,t) \equiv 1. \tag{1.12}$$

Additionally, (1.9) implies that $\bar{\gamma}$ and $\bar{u}$ satisfy

$$\frac{\partial \bar{\gamma}}{\partial t}(m,t) = \frac{\partial \bar{u}}{\partial m}(m,t). \tag{1.13}$$

Finally, if we let

$$\bar{\alpha}(m,t) \stackrel{def}{=} \alpha(x(m,t),t) = \bar{u}(m,t) - V_1(\bar{\gamma}(m,t)), \tag{1.14}$$

then (1.3) implies

$$\frac{\partial \bar{\alpha}}{\partial t}(m,t) = \begin{cases} -\dfrac{\bar{\alpha}(m,t)}{\epsilon}, & \bar{\gamma}(m,t) > \dfrac{1}{R(\bar{u}(m,t))}, \\ \dfrac{((V_2 - V_1)(\bar{\gamma}(m,t)) - \bar{\alpha}(m,t))}{\epsilon}, & \bar{\gamma}(m,t) \le \dfrac{1}{R(\bar{u}(m,t))}, \end{cases} \tag{1.15}$$

where

$$V_1(\bar{\gamma}) \stackrel{def}{=} v_1\left(\frac{1}{\bar{\gamma}}\right) \quad \text{and} \quad V_2(\bar{\gamma}) \stackrel{def}{=} v_2\left(\frac{1}{\bar{\gamma}}\right). \tag{1.16}$$

In what follows, we assume that the functions $V_1(\cdot)$ and $V_2(\cdot)$ defined in (1.16) are increasing and concave on $[L \stackrel{def}{=} 1/\rho_{\max}, \infty)$ and satisfy

$$0 = V_2(L^+) = V_1(L^+),$$

$$0 < V_2^{(p)}(\bar{\gamma}) < V_1^{(p)}(\bar{\gamma}) \quad \text{for } L < \bar{\gamma} < \infty \text{ and } p = 0, 1, \tag{1.17}$$

and the limit relations

$$\lim_{\bar{\gamma}\to\infty} (V_i(\bar{\gamma}),\ V_i^{(p)}(\bar{\gamma})) = (v_i^\infty, 0), \qquad i \text{ and } p = 1, 2, \tag{1.18}$$

where $v_2^\infty < v_1^\infty$. The parameter L is interpreted as the length of a typical car on the roadway.

Equations (1.13)–(1.15) also combine to give

$$\frac{\partial \bar{u}}{\partial t}(m,t) - V_1'(\bar{\gamma}(m,t))\frac{\partial \bar{u}}{\partial m}(m,t) = \begin{cases} \dfrac{V_1(\bar{\gamma}(m,t)) - \bar{u}(m,t)}{\epsilon}, & \bar{\gamma}(m,t) > \dfrac{1}{R(\bar{u}(m,t))}, \\ \dfrac{V_2(\bar{\gamma}(m,t)) - \bar{u}(m,t)}{\epsilon}, & \bar{\gamma}(m,t) \le \dfrac{1}{R(\bar{u}(m,t))}. \end{cases} \tag{1.19}$$

1.1. The follow-the-leader model. In [6], Greenberg showed that for the Lagrangian system (1.9)–(1.19) the appropriate stable spatial differencing scheme was downwind; see also [2]. Moreover, such differencing, with $\Delta m = 1$ (recall that cars are discrete), yields

$$\frac{dx_m}{dt} = \bar{u}_m, \tag{1.20}$$

$$\bar{\gamma}_m = x_{m+1} - x_m, \tag{1.21}$$

$$\bar{\rho}_m = \frac{1}{\bar{\gamma}_m}, \tag{1.22}$$

and

$$\frac{d\bar{u}_m}{dt} - V_1'(x_{m+1}-x_m)(\bar{u}_{m+1}-\bar{u}_m)$$

$$= \begin{cases} \dfrac{V_1(x_{m+1}-x_m)-\bar{u}_m}{\epsilon}, & x_{m+1}-x_m > \dfrac{1}{R(\bar{u}_m)}, \\ \dfrac{V_2(x_{m+1}-x_m)-\bar{u}_m}{\epsilon}, & x_{m+1}-x_m \le \dfrac{1}{R(\bar{u}_m)}. \end{cases} \tag{1.23}$$

This latter system implies that

$$\bar{\alpha}_m \stackrel{def}{=} \bar{u}_m - V_1(x_{m+1}-x_m) \tag{1.24}$$

satisfies

$$\frac{d\bar{\alpha}_m}{dt} = \begin{cases} -\dfrac{\bar{\alpha}_m}{\epsilon}, & x_{m+1}-x_m > \dfrac{1}{R(\bar{u}_m)}, \\ \dfrac{((V_2-V_1)(x_{m+1}-x_m)-\bar{\alpha}_m)}{\epsilon}, & x_{m+1}-x_m \le \dfrac{1}{R(\bar{u}_m)}. \end{cases} \tag{1.25}$$

These equations hold for $1 \le m \le M$ and $x_{M+1}(t) = x_1(t) + l$, where again l is the spatial period of our original data $\rho_0(\cdot)$ and $u_0(\cdot)$. The initial positions of the cars are constrained to satisfy

$$x_{m+1}(0) - x_m(0) \ge L \stackrel{def}{=} \frac{1}{\rho_{\max}}, \tag{1.26}$$

and these numbers are related to $\rho_0(\cdot)$ by

$$\int_{x_m(0)}^{x_{m+1}(0)} \rho_0(\xi)\,d\xi \stackrel{def}{=} \bar{\rho}_m^0(x_{m+1}(0)-x_m(0)) = 1. \tag{1.27}$$

In section 2 we analyze a first-order integration scheme for the system (1.20)–(1.22), (1.24), and (1.25). We obtain estimates which guarantee that

$$L \le x_{m+1}(t) - x_m(t) \quad \text{and} \quad 0 \le u_m(t) \le V_1(x_{m+1}(t)-x_m(t)) \tag{1.28}$$

for all $t \ge 0$. These estimates guarantee the consistency of the model. In section 3 we present some simulations with the discrete model. Here we see the persistent periodic wave trains separating congested regions of slow-moving traffic from regions of less dense faster-moving traffic. The waves separating these regions are analyzed in section 4. In that section we revert to continuum model (1.9) and (1.11)–(1.19) because it is analytically easier to work with.

2. A priori estimates. In this section we establish a priori estimates for solutions of (1.20)–(1.22), (1.24), and (1.25). We integrate these equations with a first-order Euler scheme. Specifically, we let Δt be our time step, $t_n = n\Delta t$, and for any function $f_m(\cdot)$ we let f_m^n denote the approximate value of $f_m(\cdot)$ at t_n. Our integration scheme is

$$x_m^{n+1} = x_m^n + \Delta t u_m^n, \tag{2.1}$$

$$\bar{\gamma}_m^{n+1} = x_{m+1}^{n+1} - x_m^{n+1}, \tag{2.2}$$

$$\bar{\rho}_m^{n+1} = \frac{1}{\left(x_{m+1}^{n+1} - x_m^{n+1}\right)}, \tag{2.3}$$

$$\bar{\alpha}_m^{n+1} = \left(\bar{u}_m^{n+1} - V_1(x_{m+1}^{n+1} - x_m^{n+1})\right), \tag{2.4}$$

where

$$\bar{\alpha}_m^{n+1} = \left(1 - \frac{\Delta t}{\epsilon}\right)\bar{\alpha}_m^n + \Delta t(V_2 - V_1)(x_{m+1}^n - x_m^n)H\left(\bar{\rho}_m^n - R(\bar{u}_m^n)\right)/\epsilon \tag{2.5}$$

and

$$H(s) = \begin{cases} 0, & s < 0, \\ 1, & s \geq 0. \end{cases} \tag{2.6}$$

These equations hold for $1 \leq m \leq M$ and

$$x_{M+1}^{n+1} = x_1^{n+1} + l. \tag{2.7}$$

Throughout, we assume that

$$0 \leq \Delta t V_1'(L) \leq \frac{1}{2} \quad \text{and} \quad 0 \leq \frac{\Delta t}{\epsilon} \leq \frac{1}{2}. \tag{2.8}$$

Remark. Recall that in section 1 we assumed $\Delta m = 1$ in order to obtain the follow-the-leader model. If instead we had allowed any $0 < \Delta m$, our (2.2) and (2.3) would have been replaced by $\bar{\gamma}_m^{n+1} = (x_{m+1}^{n+1} - x_m^{n+1})/\Delta m$ and $\bar{\rho}_m^{n+1} = \Delta m/(x_{m+1}^{n+1} - x_m^{n+1})$. Our basic integration scheme (2.1), (2.5) would be the same, but (2.8) would be modified to $\frac{\Delta t}{\Delta m} V_1'(L) \leq \frac{1}{2}$.

THEOREM 2.1. *Suppose that* (2.8) *holds and that for* $1 \leq m \leq M$

$$L \leq x_{m+1}^n - x_m^n \quad \text{and} \quad 0 \leq u_m^n \leq V_1(x_{m+1}^n - x_m^n). \tag{2.9}$$

Then (2.9) *holds for* n *replaced by* $n+1$.

Proof. The identities (2.1)–(2.6) imply that

$$\bar{\gamma}_m^{n+1} = \bar{\gamma}_m^n + \Delta t\left(\bar{u}_{m+1}^n - \bar{u}_m^n\right) \tag{2.10}$$

and

$$\begin{aligned} \bar{u}_m^{n+1} = {} & V_1\left(\bar{\gamma}_m^{n+1}\right) + \left(\bar{u}_m^n - V_1\left(\bar{\gamma}_m^n\right)\right)\left(1 - \frac{\Delta t}{\epsilon}\right) \\ & + \left(V_2 - V_1\right)\left(\bar{\gamma}_m^n\right) H\left(\bar{\rho}_m^n - R\left(\bar{u}_m^n\right)\right)\frac{\Delta t}{\epsilon}, \end{aligned} \tag{2.11}$$

and the inequalities

$$\left.\begin{aligned} & L \leq \bar{\gamma}_m^n, \qquad 1 \leq m \leq M, \\ & 0 \leq \bar{u}_m^n = V_1\left(\bar{\gamma}_m^n\right) + \bar{\alpha}_m^n \ \text{ and } \ \bar{\alpha}_m^n \leq 0, \ \ 1 \leq m \leq M, \end{aligned}\right\} \tag{2.12}$$

imply that

$$\bar{\gamma}_m^{n+1} \geq F\left(\bar{\gamma}_m^n\right) \stackrel{def}{=} \bar{\gamma}_m^n - \Delta t V_1\left(\bar{\gamma}_m^n\right). \tag{2.13}$$

The fact that Δt satisfies (2.8) implies that $F(\cdot)$ is monotone increasing on $[L, \infty)$, and thus (2.9) and (2.13) imply

$$\bar{\gamma}_m^{n+1} \geq F(L) = L, \tag{2.14}$$

as desired. On the other hand, the inequalities

$$\bar{u}_m^n - V_1\left(\bar{\gamma}_m^n\right) \leq 0, \qquad \left(V_2 - V_1\right)\left(\bar{\gamma}_m^n\right) \leq 0, \tag{2.15}$$

and (2.11) imply that

$$\bar{\alpha}_m^{n+1} = \bar{u}_m^{n+1} - V_1\left(\bar{\gamma}_m^{n+1}\right) \leq 0. \tag{2.16}$$

The identity (2.11), when combined with (2.10), yields

$$\begin{aligned} \bar{u}_m^{n+1} = & \left(1 - \frac{\Delta t}{\epsilon}\right)\bar{u}_m^n + \left(V_1\left(\bar{\gamma}_m^n + \Delta t\left(\bar{u}_{m+1}^n - \bar{u}_m^n\right)\right) - V_1\left(\bar{\gamma}_m^n\right)\right) \\ & + \left(\frac{\Delta t}{\epsilon}\right)\left(1 - H\left(\bar{\rho}_m^n - R(\bar{u}_m^n)\right)\right) V_1(\bar{\gamma}_m^n) \\ & + \left(\frac{\Delta t}{\epsilon}\right) H\left(\bar{\rho}_m^n - R(\bar{u}_m^n)\right) V_2(\bar{\gamma}_m^n) \end{aligned} \tag{2.17}$$

or

$$\begin{aligned} \bar{u}_m^{n+1} = & \left(1 - \frac{\Delta t}{\epsilon} - \Delta t V_1'(\delta_m^n)\right)\bar{u}_m^n + \Delta t V_1'(\delta_m^n)\bar{u}_{m+1}^n \\ & + \left(\frac{\Delta t}{\epsilon}\right)\left(1 - H(\bar{\rho}_m^n - R(\bar{u}_m^n))\right)V_1(\bar{\gamma}_m^n) + \left(\frac{\Delta t}{\epsilon}\right) H(\bar{\rho}_m^n - R(\bar{u}_m^n))V_2(\bar{\gamma}_m^n) \end{aligned} \tag{2.18}$$

for some $\delta_m^n \geq \min\left(\gamma_m^{n+1}, \gamma_m^n\right) \geq L$, and (2.18) together with (2.6) and (2.8) and $u_m^n \geq 0$, $1 \leq m \leq M$, imply that $\bar{u}_m^{n+1} \geq 0$. This concludes the proof of Theorem 2.1. □

The estimates contained in Theorem 2.1 guarantee that the densities

$$\rho_m^n = \frac{1}{x_{m+1}^n - x_m^n}, \qquad 1 \leq m \leq M, \tag{2.19}$$

satisfy

$$0 \leq \rho_m^n \leq \rho_{\max}. \tag{2.20}$$

These estimates further imply that the approximate solutions defined in (2.1)–(2.7) converge to solutions of the follow-the-leader model (1.20)–(1.22), (1.24), (1.25) as $\Delta t \to 0^+$.

3. Simulations. All computations in this section were run with the following equilibrium relations:

$$v_1(\rho) = v_1^\infty \left(1 - \frac{\rho}{\rho_{\max}}\right) \quad \text{and} \quad v_2(\rho) = v_2^\infty \left(1 - \frac{\rho}{\rho_{\max}}\right). \tag{3.1}$$

These transform to

$$V_1(\gamma) = v_1^\infty \left(1 - \frac{L}{\gamma}\right) \quad \text{and} \quad V_2(\gamma) = v_2^\infty \left(1 - \frac{L}{\gamma}\right), \tag{3.2}$$

where $L = 1/\rho_{\max}$. The specific parameters used were

$$v_1^\infty = 100 \text{ feet/sec } = \frac{100 \times 3600}{5280} = 68.1818\ldots \text{ mph}, \tag{3.3}$$

$$v_2^\infty = 40 \text{ feet/sec} = \frac{40 \times 3600}{5280} = 27.2727\ldots \text{ mph}, \tag{3.4}$$

and

$$L = 15 \text{ feet}. \tag{3.5}$$

The latter number corresponds to a maximum car density of

$$\rho_{\max} = \frac{1}{15} \text{ cars/foot } = \frac{5280}{15} = 352 \text{ cars/mile}. \tag{3.6}$$

We used the constant switch curve introduced by Sopasakis [9]:

$$\gamma(u) = \gamma_*, \quad 0 \le u, \tag{3.7}$$

with $\gamma_* = 20$ feet. For initial data, we chose three sets of data:

$$x_m^{(k)}(0) = 20m + .1 \sin\left(\frac{km\pi}{200}\right) \tag{3.8}$$

for $-\infty \le m \le \infty$ and $k = 1, 2$, and 3. The observation that

$$x_{400}^{(k)}(0) = 8000 \text{ feet } = 1.5151\ldots \text{ miles} \tag{3.9}$$

and

$$x_{m+400}^{(k)}(0) = x_m^{(k)}(0) + 8000 \tag{3.10}$$

implies that we may interpret the data as initial data for a ring road with 400 cars which is of length 1.5151... miles. We chose constant initial velocities

$$u_m^{(k)}(0) = .5(V_1(\gamma_*) + V_2(\gamma_*)), \quad 1 \le m \le 400, \tag{3.11}$$

or

$$u_m^{(k)}(0) = 17.5 \text{ feet/sec } = 11.931818\ldots \text{ mph}, \quad 1 \le m \le 400. \tag{3.12}$$

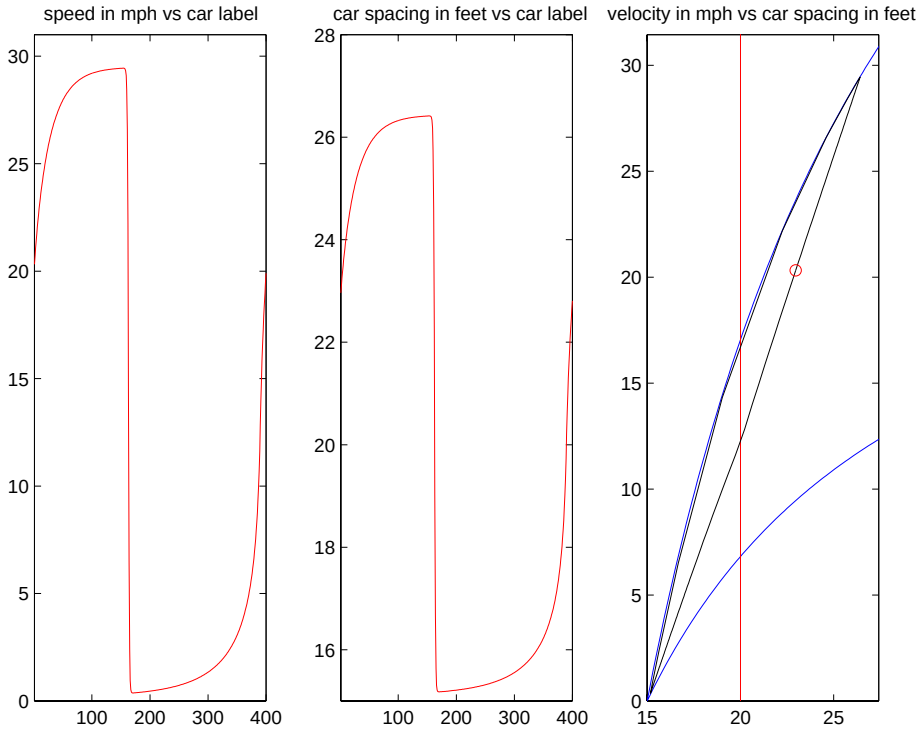


FIG. 3.1. *Periodic solutions at* $t = 2$ *for initial data* $k = 1$.

These data guarantee points on both sides of the switch curve. Simulations were run with relaxation times

$$\epsilon = 1, 2, 4, \text{ and } 8. \tag{3.13}$$

Below, we show the long-time spatially and temporally periodic solutions at time $t = 2$ hours when $\epsilon = 8$ seconds. Figures 3.1, 3.2, and 3.3 correspond to the initial data indexed by $k = 1, 2$, and 3, respectively. At earlier times the solution indexed by each particular k had k discontinuities per period. This phenomenon persisted to $t = 2$ hours for the solution indexed by $k = 2$, but the solution corresponding to the index $k = 3$ converged, by $t = 2$ hours, to a solution with one discontinuity per period.

The first two frames in each figure are self-explanatory. In the third frame of each figure we plot the curve $m \to (\gamma_m = x_{m+1} - x_m, u_m)$. This curve is shown in black. The blue curves are the equilibrium curves $\gamma \to (\gamma, V_1, (\gamma))$ and $\gamma \to (\gamma, V_2(\gamma))$, and the red curve is the image of $u \to (20, u)$. The red circle is the image of (γ_1, u_1). Complete animations of all of these simulations may be found at http://www.math.cmu.edu/∼plin/congestion/. The discontinuities in the profiles propagate at the speed

$$c \simeq 227.6 \pm .1 \text{ cars/minute}. \tag{3.14}$$

An analysis of these solutions may be found in section 4.

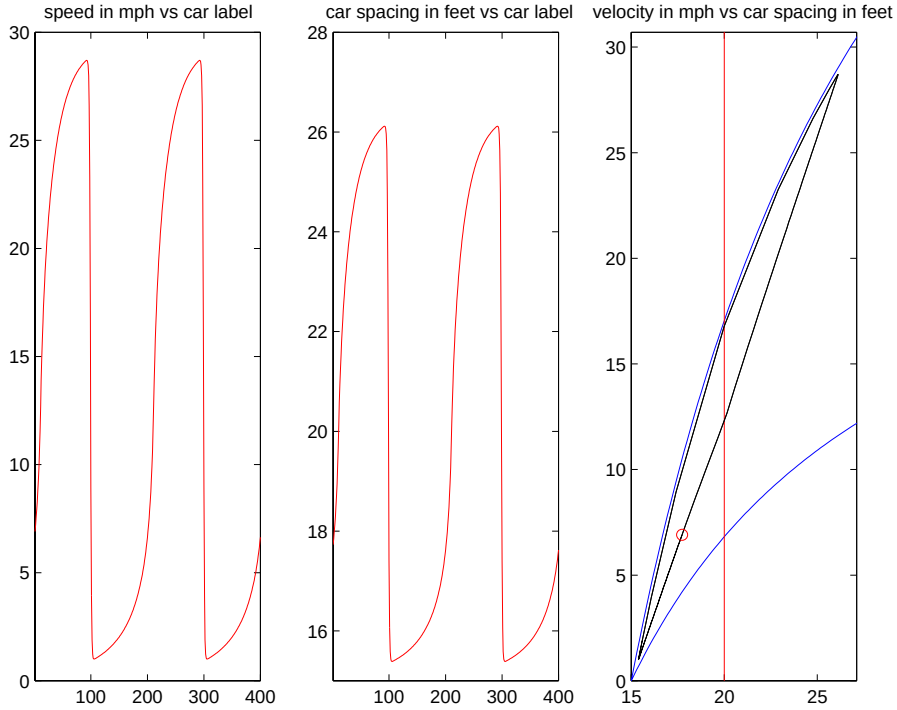


FIG. 3.2. *Periodic solutions at* $t = 2$ *for initial data* $k = 2$.

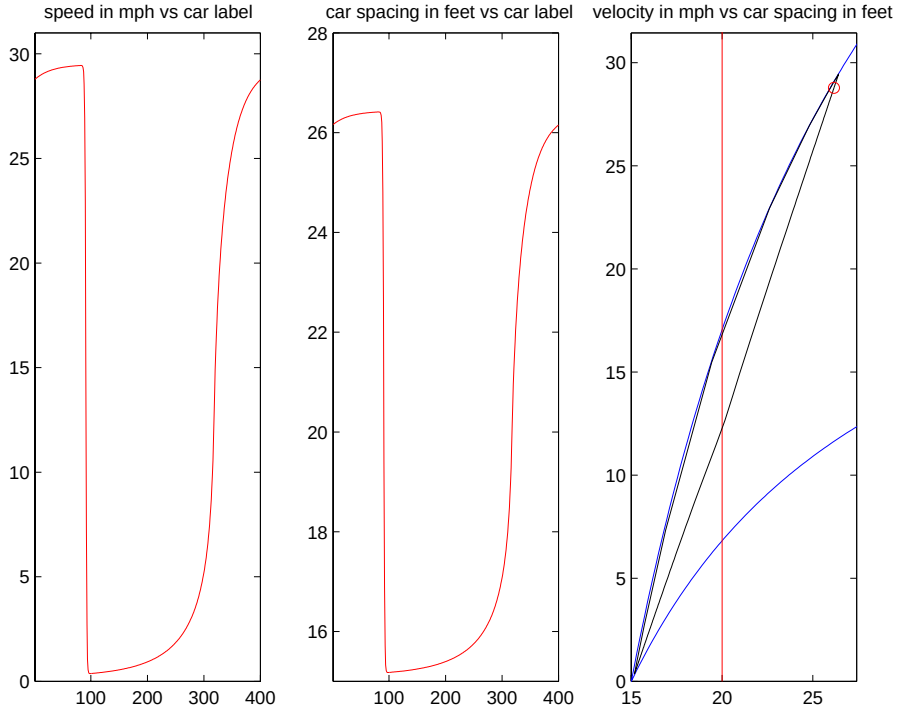


FIG. 3.3. *Periodic solutions at* $t = 2$ *for initial data* $k = 3$.

4. Travelling waves. The wave trains obtained in section 3 are basically discrete approximations to travelling wave solutions to the continuum equations (1.9)–(1.19). In this section our goal is to show that the continuum system (1.9)–(1.19) actually supports such travelling waves. For definiteness we shall assume that the switch curve introduced in (1.4) is the one derived by Sopasakis in [9], namely, the curve

$$R(u) = \rho_*, \quad 0 \le u. \tag{4.1}$$

With this choice of switch curve, the Lagrangian equations become

$$\frac{\partial \bar\gamma}{\partial t} - \frac{\partial \bar u}{\partial m} = 0 \quad \text{and} \quad \frac{\partial \bar u}{\partial t} - V_1'(\bar\gamma)\frac{\partial \bar u}{\partial m} = \begin{cases} \dfrac{V_1(\bar\gamma) - \bar u}{\epsilon}, & \bar\gamma > \gamma_* = \dfrac{1}{\rho_*}, \\[2ex] \dfrac{V_2(\bar\gamma) - \bar u}{\epsilon}, & \bar\gamma \le \gamma_* = \dfrac{1}{\rho_*}. \end{cases} \tag{4.2}$$

Once again,

$$V_1(\bar\gamma) = v_1\left(\frac{1}{\bar\gamma}\right) \quad \text{and} \quad V_2(\bar\gamma) = v_2\left(\frac{1}{\bar\gamma}\right), \tag{4.3}$$

and we assume that both V_1 and V_2 are increasing and concave on $[L, \infty)$ and satisfy

$$0 = V_2(L^+) = V_1(L^+) \quad \text{and} \quad 0 < V_2^{(p)}(\bar\gamma) < V_1^{(p)}(\bar\gamma) \quad \text{for } L < \bar\gamma < \infty \text{ and } p = 0, 1 \tag{4.4}$$

and the limit relations

$$\lim_{\bar\gamma \to \infty} (V_i(\bar\gamma),\ V_i^{(p)}(\bar\gamma)) = (v_i^\infty, 0), \quad i \text{ and } p = 1, 2, \tag{4.5}$$

where $0 < v_2^\infty < v_1^\infty$. L is related to $\rho_{\max}$ by $L = 1/\rho_{\max}$.

We start by describing the portion of the wave trains in which both $\bar\gamma$ and $\bar u$ are increasing in m. These solutions are functions of

$$\xi = m + ct \tag{4.6}$$

and are normalized so that

$$\bar\gamma(0) = \gamma_* \quad \text{and} \quad V_2(\gamma_*) < u_* < V_1(\gamma_*). \tag{4.7}$$

Once again, $\gamma_* = 1/\rho_*$ (see (4.1)). Equation $(4.2)_1$ implies that $\bar u = u_* + c(\bar\gamma - \gamma_*)$, while $(4.2)_2$ yields

$$c\,(c - V_1'(\bar\gamma))\,\frac{d\bar\gamma}{d\xi} = \begin{cases} \dfrac{V_1(\bar\gamma) - u_* - c(\bar\gamma - \gamma_*)}{\epsilon}, & \bar\gamma > \gamma_*, \\[2ex] \dfrac{V_2(\bar\gamma) - u_* - c(\bar\gamma - \gamma_*)}{\epsilon}, & \bar\gamma \le \gamma_*. \end{cases} \tag{4.8}$$

The requirement that $\bar\gamma$ be increasing in ξ implies that $\bar\gamma$ must satisfy $d\bar\gamma/d\xi(0^-) \ge 0$ and $d\bar\gamma/d\xi(0^+) \ge 0$. Equations (4.7) and (4.8) then imply that these latter inequalities may be met only if

$$c = V_1'(\gamma_*). \tag{4.9}$$

In what follows, we let $\Gamma_* > L$ be the unique solution of

$$V_1'(\Gamma_*) = V_2'(L^+). \tag{4.10}$$

If $L < \gamma_* < \Gamma_*$, we let

$$\bar{U} = V_1'(\gamma_*)(\gamma_* - L) \tag{4.11}$$

and note that for $V_2(\gamma_*) < u_* < \bar{U}$ the equation

$$u_* + V_1'(\gamma_*)(\bar{\gamma} - \gamma_*) = V_2(\bar{\gamma}) \tag{4.12}$$

has a unique solution $\gamma_- \in (L, \gamma_*)$ satisfying

$$V_1'(\gamma_*) > V_2'(\gamma_-). \tag{4.13}$$

On the other hand, if $\Gamma_* < \gamma_*$, we let $\gamma_l \in (L, \gamma_*)$ be the unique solution of

$$V_2'(\gamma_l) = V_1'(\gamma_*) \tag{4.14}$$

and

$$\bar{U} = V_1'(\gamma_*)(\gamma_* - \gamma_l) + V_2(\gamma_l) \tag{4.15}$$

and note that, for $V_2(\gamma_*) < u_* < \bar{U}$, (4.11) has a unique solution $\gamma_- \in (\gamma_l, \gamma_*)$ satisfying (4.12).

In what follows, we assume that the parameter u_* in (4.8) satisfies $V_2(\gamma_*) < u_* \leq \bar{U}$, where $\bar{U}$ is defined in (4.10) or (4.14) as appropriate.

We now note that $(4.2)_2$, when combined with (4.8), implies that the profile $\bar{\gamma}$ must satisfy

$$V_1'(\gamma_*)\left(V_1'(\gamma_*) - V_1'(\bar{\gamma})\right)\frac{d\bar{\gamma}}{d\xi} = \begin{cases} \dfrac{V_1(\bar{\gamma}) - u_* - V_1'(\gamma_*)(\bar{\gamma} - \gamma_*)}{\epsilon}, & \bar{\gamma} > \gamma_*, \\ \dfrac{V_2(\bar{\gamma}) - u_* - V_1'(\gamma_*)(\bar{\gamma} - \gamma_*)}{\epsilon}, & \bar{\gamma} \leq \gamma_*. \end{cases} \tag{4.16}$$

Once again, we normalize the profile by insisting that (4.7) hold. Noting that sign $(V_1'(\gamma_*) - V_1'(\bar{\gamma})) =$ sign $(\bar{\gamma} - \gamma_*)$, that

$$V_1(\bar{\gamma}) - u_* - V_1'(\gamma_*)(\bar{\gamma} - \gamma_*) > 0, \quad \gamma_* < \bar{\gamma} < \gamma_+, \tag{4.17}$$

where $\gamma_* < \gamma_+$ is the unique solution of

$$V_1(\gamma_+) - u_* - V_1'(\gamma_*)(\gamma_+ - \gamma_*) = 0, \tag{4.18}$$

and finally that

$$V_2(\bar{\gamma}) - u_* - V_1'(\gamma_*)(\bar{\gamma} - \gamma_*) < 0, \quad \gamma_- < \bar{\gamma} < \gamma_*, \tag{4.19}$$

where γ_- is defined in (4.11), we see that (4.15) and (4.16) have a unique increasing solution defined on $(-\infty, \infty)$. For $\xi < 0$ the solution is given by the quadrature formula

$$\epsilon V_1'(\gamma_*) \int_{\bar{\gamma}(\xi)}^{\gamma_*} \frac{(V_1'(\eta) - V_1'(\gamma_*))d\eta}{(u_* + V_1'(\gamma_*)(\eta - \gamma_*) - V_2(\eta))} = -\xi, \tag{4.20}$$

and for $\xi > 0$ the solution is given by

$$\epsilon V_1'(\gamma_*) \int_{\gamma_*}^{\bar{\gamma}(\xi)} \frac{(V_1'(\gamma_*) - V_1'(\eta))d\eta}{(V_1(\eta) - u_* - V_1'(\gamma_*)(\eta - \gamma_*))} = \xi. \tag{4.21}$$

4.1. Periodic profiles. For any $\bar{\gamma} \in (\gamma_-, \gamma_*)$, we let $\Gamma(\bar{\gamma}) > \gamma_*$ be the unique solution of

$$V_1(\Gamma(\bar{\gamma})) - V_1(\bar{\gamma}) = V_1'(\gamma_*)(\Gamma(\bar{\gamma}) - \bar{\gamma}) \tag{4.22}$$

and note that

$$\frac{d\Gamma(\bar{\gamma})}{d\bar{\gamma}} = \frac{(V_1'(\bar{\gamma}) - V_1'(\gamma_*))}{(V_1'(\Gamma(\bar{\gamma})) - V_1'(\gamma_*))} < 0. \tag{4.23}$$

We are now in a position to define the periodic wave trains. For $-|\xi_a| < \xi \le 0$, $\bar{\gamma}(\xi)$ is given by (4.20), and $|\xi_a|$ is given by

$$\epsilon V_1'(\gamma_*) \int_{\bar{\gamma}_a}^{\gamma_*} \frac{(V_1'(\eta) - V_1'(\gamma_*))d\eta}{(u_* + V_1'(\gamma_*)(\eta - \gamma_*) - V_2(\eta))} \stackrel{def}{=} |\xi_a|, \tag{4.24}$$

where $\gamma_- < \bar{\gamma}_a < \gamma_*$. For $0 \le \xi \le \xi_{\Gamma(\bar{\gamma}_a)}$, $\bar{\gamma}(\xi)$ is given by (4.21), and $\xi_{\Gamma(\bar{\gamma}_a)}$ is given by

$$\epsilon V_1'(\gamma_*) \int_{\gamma_*}^{\Gamma(\bar{\gamma}_a)} \frac{(V_1'(\gamma_*) - V_1'(\eta))d\eta}{(V_1(\eta) - u_* - V_1'(\gamma_*) * (\eta - \gamma_*))} \stackrel{def}{=} \xi_{\Gamma(\bar{\gamma}_a)}. \tag{4.25}$$

We extend these solutions to all ξ via

$$\bar{\gamma}(\xi) = \bar{\gamma}(\xi + \xi_{\Gamma(\bar{\gamma}_a)} + |\xi_a|). \tag{4.26}$$

The extended solution is a proper weak solution to (4.2). The relations (4.9) and (4.22) imply that the Rankine–Hugoniot relations for (4.2) hold across the discontinuities

$$\xi = m + V_1'(\gamma_*)t = \xi_{\Gamma(\bar{\gamma}_a)} \pm n(\xi_{\Gamma(\bar{\gamma}_a)} + |\xi_a|), \quad n = 0, 1, \ldots . \tag{4.27}$$

Equation (4.22) also implies that

$$V_1'(\bar{\gamma}_a) > V_1'(\gamma_*) = \frac{V_1(\Gamma(\bar{\gamma}_a)) - V_1(\bar{\gamma}_a)}{\Gamma(\bar{\gamma}_a) - \bar{\gamma}_a} > V_1'(\Gamma(\bar{\gamma}_a)), \tag{4.28}$$

and thus across these discontinuities the Lax entropy condition is satisfied. Recalling that the particular solutions of interest to us must be M periodic, we see that (4.24) and (4.25) imply that for some integer $k \ge 1, \bar{\gamma}_a$ and u_* must be such that

$$\begin{aligned} k\epsilon V_1'(\gamma_*) \Bigg[& \int_{\bar{\gamma}_a}^{\gamma_*} \frac{(V_1'(\eta) - V_1'(\gamma_*))d\eta}{(u_* + V_1'(\gamma_*)(\eta - \gamma_*) - V_2(\eta))} \\ & + \int_{\gamma_*}^{\Gamma(\bar{\gamma}_a)} \frac{(V_1'(\gamma_*) - V_1'(\eta))d\eta}{(V_1(\eta) - u_* - V_1'(\gamma_*)(\eta - \gamma_*))} \Bigg] = M. \end{aligned} \tag{4.29}$$

The condition that $x(M, t) = x(1, t) + l$ implies that $\bar{\gamma}_a$ and u_* must also satisfy

$$\begin{aligned} k\epsilon V_1'(\gamma_*) \Bigg[& \int_{\bar{\gamma}_a}^{\gamma_*} \frac{(V_1'(\eta) - V_1'(\gamma_*))\eta d\eta}{(u_* + V_1'(\gamma_*)(\eta - \gamma_*) - V_2(\eta))} \\ & + \int_{\gamma_*}^{\Gamma(\bar{\gamma}_a)} \frac{(V_1'(\gamma_*) - V_1'(\eta))\eta d\eta}{(V_1(\eta) - u_* - V_1'(\gamma_*)(\eta - \gamma_*))} \Bigg] = l. \end{aligned} \tag{4.30}$$

We conclude this section with an analysis of (4.29) and (4.30). We first note that the integer $k \geq 1$ in these equations is equal to the number of discontinuities of $\bar{\gamma}(\cdot)$ per period. We also note that instead of using u_* and $\bar{\gamma}_a$ as our basic parameters we may instead use γ_- and $\bar{\gamma}_a$. With this choice,

$$\begin{aligned} &u_* + V_1'(\gamma_*)(\eta - \gamma_*) - V_2(\eta) \\ &\qquad = V_2(\gamma_-) + V_1'(\gamma_*)(\eta - \gamma_-) - V_2(\eta) > 0, \quad \gamma_- < \eta < \gamma_*, \end{aligned} \tag{4.31}$$

and

$$\begin{aligned} &V_1(\eta) - u_* - V_1'(\gamma_*)(\eta - \gamma_*) \\ &\qquad = V_1(\eta) - V_2(\gamma_-) - V_1'(\gamma_*)(\eta - \gamma_-) > 0, \quad \gamma_* < \eta < \Gamma(\gamma_-), \end{aligned} \tag{4.32}$$

and solving (4.29) and (4.30) is equivalent to finding $\bar{\gamma}_a \in (\gamma_-, \gamma_*)$ and $\gamma_- < \gamma_*$ such that

$$\begin{aligned} k\epsilon V_1'(\gamma_*) \Bigg[&\int_{\bar{\gamma}_a}^{\gamma_*} \frac{(V_1'(\eta) - V_1'(\gamma_*))d\eta}{(V_2(\gamma_-) + V_1'(\gamma_*)(\eta - \gamma_-) - V_2(\eta))} \\ &+ \int_{\gamma_*}^{\Gamma(\bar{\gamma}_a)} \frac{(V_1'(\gamma_*) - V_1'(\eta))d\eta}{(V_1(\eta) - V_2(\gamma_-) - V_1'(\gamma_*)(\eta - \gamma_-))} \Bigg] = M \end{aligned} \tag{4.33}$$

and

$$\begin{aligned} k\epsilon V_1'(\gamma_*) \Bigg[&\int_{\bar{\gamma}_a}^{\gamma_*} \frac{(V_1'(\eta) - V_1'(\gamma_*))\eta d\eta}{(V_2(\gamma_-) + V_1'(\gamma_*)(\eta - \gamma_-) - V_2(\eta))} \\ &+ \int_{\gamma_*}^{\Gamma(\bar{\gamma}_a)} \frac{(V_1'(\gamma_*) - V_1'(\eta))\eta d\eta}{(V_1(\eta) - V_2(\gamma_-) - V_1'(\gamma_*)(\eta - \gamma_-))} \Bigg] = l. \end{aligned} \tag{4.34}$$

In what follows, we let $L_1(\gamma_-, \bar{\gamma}_a, \gamma_*)$ and $L_2(\gamma_-, \bar{\gamma}_a, \gamma_*)$ be the functions defined by the left-hand sides of (4.33) and (4.34), respectively. If $L < \gamma_* < \Gamma_*$ (see (4.9)), the functions L_1 and L_2 are well defined for $\gamma_- \in (L, \gamma_*)$ and $\bar{\gamma}_a \in (\gamma_-, \gamma_*)$, whereas if $\Gamma_* \leq \gamma_*$, these functions are well defined for $\gamma_- \in (\gamma_l, \gamma_*)$ (see (4.13)) and $\bar{\gamma}_a \in (\gamma_-, \gamma_*)$. In either case, the observation that $\lim_{\bar{\gamma}_a \to \gamma_*^-} \Gamma(\bar{\gamma}_a) = \gamma_*$ implies that $L_1(\gamma_-, \gamma_*^-, \gamma_*) = L_2(\gamma_-, \gamma_*^-, \gamma_*) = 0$. We further note that for $\gamma_- < \bar{\gamma}_a < \gamma_*$

$$\begin{aligned} \frac{\partial L_1}{\partial \bar{\gamma}_a}(\gamma_-, \bar{\gamma}_a, \gamma_*) = k\epsilon V_1'(\gamma_*) \Bigg[&\frac{(V_1'(\gamma_*) - V_1'(\bar{\gamma}_a))}{(V_2(\gamma_-) + V_1'(\gamma_*)(\bar{\gamma}_a - \gamma_*) - V_2(\bar{\gamma}_a))} \\ &+ \frac{d\Gamma(\bar{\gamma}_a)}{d\bar{\gamma}_a} \frac{(V_1'(\gamma_*) - V_1'(\Gamma(\bar{\gamma}_a)))}{(V_1(\Gamma(\bar{\gamma}_a)) - V_2(\gamma_-) - V_1'(\gamma_*)(\Gamma(\bar{\gamma}_a) - \gamma_-))} \Bigg]. \end{aligned} \tag{4.35}$$

The last identity, together with $d\Gamma/d\bar{\gamma}_a(\bar{\gamma}_a) < 0$, implies that $\partial L_1/\partial \bar{\gamma}_a(\gamma_-, \bar{\gamma}_a, \gamma_*) < 0$. The fact that

$$\lim_{\bar{\gamma}_a \to \gamma_-^+} L_1(\gamma_-, \bar{\gamma}_a, \gamma_*) = +\infty \tag{4.36}$$

then guarantees that for each $L < \gamma_*$ and admissible γ_- there is a unique $\bar{\gamma}_a(\gamma_-, \gamma_*, M)$ such that (4.33) holds. Thus, solving (4.33) and (4.34) is equivalent to finding an admissible $\gamma_- < \gamma_*$ so that

$$L_2(\gamma_-, \bar{\gamma}_a(\gamma_-, \gamma_*, M), \gamma_*) = l. \tag{4.37}$$

The integral mean-value theorem, when combined with the definition of $\bar{\gamma}_a(\gamma_-, \gamma_*, M)$, guarantees that

$$M\bar{\gamma}_a(\gamma_-, \gamma_*, M) \le L_2(\gamma_-, \bar{\gamma}_a(\gamma_-, \gamma_*, M), \gamma_*) = Mg \tag{4.38}$$

for some $g \in (\bar{\gamma}_a(\gamma_-, \gamma_*, M),\ \Gamma(\bar{\gamma}_a(\gamma_-, \gamma_*, M)))$. These observations, together with $\gamma_- < \bar{\gamma}_a(\gamma_-, \gamma_*, M)$ and $\Gamma(\bar{\gamma}_a(\gamma_-, \gamma_*, M)) < \Gamma(\gamma_-)$, imply that (4.37) has no solutions for

$$l < \begin{cases} ML & \text{if } \gamma_* < \Gamma_* \text{ (see (4.9) and (4.22))}, \\ M\gamma_l & \text{if } \gamma_* \ge \Gamma_* \text{ (see (4.9), (4.13), and (4.22))} \end{cases} \tag{4.39}$$

and

$$l > \begin{cases} M\Gamma(L) & \text{if } \gamma_* < \Gamma_* \text{ (see (4.9) and 4.22))}, \\ M\Gamma(\gamma_l) & \text{if } \gamma_* \ge \Gamma_* \text{(see (4.9), (4.13), and (4.22))}. \end{cases} \tag{4.40}$$

These estimates on the range of $\gamma_- \to L_2(\gamma_-, \bar{\gamma}_a(\gamma_-, \gamma_*, M), \gamma_*)$, though not particularly sharp, are all we could manage with this degree of generality on the functions $V_1(\cdot)$ and $V_2(\cdot)$.

REFERENCES

[1] B. Argall, E. Cheleshkin, J. M. Greenberg, C. Hinde, and P.-J. Lin, *A rigorous treatment of a follow-the-leader traffic model with traffic lights present,* SIAM J. Appl. Math, 63 (2002), pp. 149–168.

[2] A. Aw, A. Klar, T. Materne, and M. Rascle, *Derivation of continuum traffic flow models from microscopic follow-the-leader models,* SIAM J. Appl. Math., 63 (2002), pp. 259–278.

[3] A. Aw and M. Rascle, *Resurrection of "second order" models of traffic flow,* SIAM J. Appl. Math., 60 (2000), pp. 916–938.

[4] P. Bagnerini and M. Rascle, *Un modèle hyperbolique homogénéisé pour le trafic routier*, C. R. Acad. Sci. Paris, to appear.

[5] C. F. Daganzo, *A behavioral theory of multi-lane traffic flow. Part* I*: Long homogeneous freeway sections,* Transp. Res. B, 36 (2002), pp. 131–158.

[6] J. M. Greenberg, *Extensions and amplifications of a traffic model of Aw and Rascle,* SIAM J. Appl. Math., 62 (2001), pp. 729–745.

[7] B. S. Kerner, *Experimental features of self-organization in traffic flow,* Phys. Rev. Lett., 81 (1998), pp. 3797–3800.

[8] B. S. Kerner, *Congested traffic flow,* Transp. Res. Rec., 1678 (1998), pp. 160–167.

[9] A. Sopasakis, *Unstable flow and modeling,* Math. Comput. Modelling, to appear.

[10] M. Zhang, *A nonequilibrium traffic model devoid of gas like behaviour,* Transp. Res. B, to appear.

SIAM J. APPL. MATH.
Vol. 63, No. 3, pp. 834–849

ACOUSTIC PROPAGATION IN DISPERSIONS AND THE GEOMETRIC THEORY OF DIFFRACTION*

O. G. HARLEN†, M. J. HOLMES‡, M. J. W. POVEY‡, AND B. D. SLEEMAN†

Abstract. The ultrasonic characterization of emulsions relies principally upon the theory of thermoacoustic scattering. For a single spherical particle of radius a suspended in a homogeneous medium, the theory provides an exactly soluble solution to the scattering problem. Unfortunately, direct computations with this solution are often ill-conditioned for certain ranges of the acoustic wave number K and thermal wave number L. Recently the authors have developed a low frequency (i.e., $|Ka|, |La| \ll 1$) approach to the theory based on a potential theory technique. This low frequency theory is rapidly convergent and overcomes computation ill-conditioning. In this paper, we pursue the single particle theory and consider the region in which $|Ka| \ll 1$ and $|La| \gg 1$, i.e., the high frequency range of the thermal wavelength. To achieve this, we employ the geometric theory of diffraction. We show that the high frequency solution agrees well with both the experimental measurements and the exact solution in the region where $|Ka| \ll 1$ and $|La| \gg 1$. As in the low frequency case, the high frequency solution may be applied to arbitrary scattering domains.

Key words. Helmholtz equation, Poisson equation, ultrasound spectroscopy, geometrical theory of diffraction, low frequency limit, high frequency limit

AMS subject classifications. 35C10, 35J05, 35P25, 76Q05

PII. S0036139902404670

1. Introduction.

Ultrasound techniques for the characterization of colloidal systems are gaining wide acceptance, and many new ultrasound instruments have recently appeared on the market. The bulk of these instruments claim to accurately measure colloidal particle size distribution, in the case of oil-in-water emulsions at concentrations of up to 30 percent. All these instruments determine particle size from a measurement of ultrasonic attenuation as a function of frequency, together with thermophysical data for the continuous and dispersed phases such as thermal diffusivity. The claims depend on an accurate model of acoustic scattering in such systems. A recent review [1] suggests that the reason for the accuracy of these instruments is that scattering in water based systems is dominated by thermal scattering and that multiple scattering of the thermal field is small. The review authors then suggest that the ECAH model (named for Epstein–Carhart [2] and Allegra–Hawley [3]) upon which most instruments are based is too mathematically complex to be of use in cases where particle-particle interactions are important. In a previous paper [4] we showed that the real deficiency of ECAH is not its complexity but the fact that it is ill-conditioned, thereby leading to unreliability in its numerical predictions, despite the fact that it is an exact theory for a single spherical particle. We demonstrated that, in the case of silicone oil-in-water emulsions, ECAH was accurate to only a few volume percent of the dispersed phase and that the increasing overestimation of the attenuation by ECAH at higher concentrations was a result of unjustifiable approximations made in

*Received by the editors March 25, 2002; accepted for publication (in revised form) July 10, 2002; published electronically January 17, 2003. This research is supported by the Engineering and Physical Research Council, UK Soft Solids Programme grant GR/N17058.
http://www.siam.org/journals/siap/63-3/40467.html
†Department of Applied Mathematics, University of Leeds, Leeds, LS2 9JT, England (o.g.harlen@leeds.ac.uk, b.d.sleeman@leeds.ac.uk).
‡Procter Department of Food Science, University of Leeds, Leeds, LS2 9JT, England (amtmh@amsta.leeds.ac.uk, m.j.w.povey@leeds.ac.uk).

the multiple scattering model. In addition, high precision arithmetic was needed to overcome the numerical instability of solutions to ECAH.

To overcome the ill-conditioning in ECAH, we developed in [4] a low frequency solution to the problem of ultrasound propagation in dispersions. We adopted a conventional approach first formulated by Lord Rayleigh [5], refined by Epstein and Carhart [2] and Allegra and Hawley [3]. We call this the ECAH approach, and our low frequency approximation the low frequency potential scattering theory (LFPST). LFPST is based on an asymptotic approach first introduced by Kleinman [6], and we showed that it accurately reproduces experimental results and numerical results from ECAH computed using 24 digit precision arithmetic. The importance of this problem relates to the use of ultrasound for characterizing soft solid and particulate material in a wide range of industries including foods [7, 8] and pharmaceuticals [9]. Since very many such applications lie in the low frequency acoustic limit, an accurate approximation in this limit is of great potential utility. The starting point for LFPST is a set of linearized mass, momentum, and energy equations [4, 10] that reduce to a set of Helmholtz equations, coupled through transmission boundary conditions at the particle surface. It should be pointed out here that this solution at present includes only thermal scattering and that the visco-inertial scattering terms have been omitted for the sake of simplicity. The ECAH solution is not only ill-conditioned [4, 10, 11, 12], but also, because it is based on spherical harmonics, cannot easily be generalized to account for arbitrary particle shape. On the other hand, LFPST is well-conditioned, can provide solutions to arbitrary accuracy, and lends itself to boundary-integral and finite-element techniques that permit solution of the single scattering problem for arbitrary particle shape. In our previous papers [4, 11] we showed that, while it is true that many applications of ultrasound to food and similar emulsions lie in the acoustic low frequency limit, a great proportion of experimental data is high frequency regarding the thermal branch of the problem. In this paper we develop a novel asymptotic description appropriate for the low frequency acoustic limit ($Ka \ll 1$, where K is the acoustic wave vector and a is the particle radius) but where, nevertheless, the problem is high frequency in the thermal field ($La \gg 1$, where L is the thermal wave vector). Examples of such data are given later. We employ the LFPST method for the low frequency acoustic branch (expansion of the acoustic field in powers of iK [4, 11]) and the geometric theory of diffraction (expansion of the thermal field utilizing inverse powers of iL [4]) for the high frequency thermal branch. An iterative system of Poisson equations is formed, whose solution, dependent upon the boundary conditions, provides coefficient values for the field expansions. This approach is not intrinsically constrained by geometrical aspects of the domain. We show, by making an appropriate scaling, that the boundary conditions may be decoupled to second order, significantly simplifying the analysis. In section two, we introduce the problem of thermoacoustic scattering. In section three, we give details of the low frequency perturbation solution, and in section four, we present the geometric theory of diffraction which is applied to the high frequency perturbation solution. In section five, we combine the results of the previous two sections and introduce the form of our perturbation solution together with the associated analytical solution of the problem. Finally, in section six, we present the far field pattern for the solution and compare it with experimental results and the exact solution.

2. Thermoacoustic scattering. We now introduce the model equations and boundary conditions that quantify the problem, keeping our variables consistent with those of [4] as much as possible. As in [4], we consider the scattering of a plane

sound wave of radial frequency ω by a droplet with boundary B that has contrasting thermal and compressive properties. For simplicity we will neglect the effects of viscosity, as these are of secondary importance in oil-water emulsions, though these can be incorporated into the formulation [1]. We define the domains D_1 and D_2 to be regions outside and inside the drop, respectively, and $\mathbf{n}$ to be outward normal on B. The linearized equation for momentum gives

$$i\omega\rho\nabla\Phi + \nabla P = 0, \tag{2.1}$$

where Φ is the velocity potential, P is the pressure perturbation, and ρ is density. Hence the pressure perturbation and velocity potential are related by

$$P = -i\omega\rho\Phi. \tag{2.2}$$

Conservation of energy gives

$$i\omega T + \gamma\sigma\nabla^2 T + \frac{(\gamma-1)}{\beta}\nabla^2\Phi = 0, \tag{2.3}$$

where T is the temperature perturbation, γ the ratio of specific heats, β the coefficient of thermal expansion, and $\sigma = \tau/\rho C_P$ the thermal diffusivity. (τ is thermal conductivity and C_P the specific heat at constant pressure.) Finally, the combination of mass conservation and the first law of thermodynamics gives

$$i\omega P = -\frac{\rho v^2}{\gamma}\nabla^2\Phi + \frac{i\omega v^2\beta\rho}{\gamma}T, \tag{2.4}$$

where v is the adiabatic sound speed. Substituting (2.2) into (2.4) gives

$$\left(\nabla^2 + \frac{\gamma\omega^2}{v^2}\right)\Phi = i\omega\beta T,$$

and substituting for T in (2.3) gives a biharmonic equation for the velocity potential Φ,

$$\nabla^4\Phi + \left(\frac{\omega^2\gamma}{v^2} + i\frac{\omega}{\sigma}\right)\nabla^2\Phi + i\frac{\omega^3}{\sigma v^2}\Phi = 0. \tag{2.5}$$

Equations for the pressure, velocity, density, and temperature variations are obtained from the conservation of mass, momentum, and energy. Equation (2.5) may be factorized as

$$(\nabla^2 + K^2)(\nabla^2 + L^2)\Phi = 0, \tag{2.6}$$

where, in the limit of small $\sigma\omega/v^2$,

$$K \simeq \frac{\omega}{v}\left(1 + i\frac{(\gamma-1)\sigma\omega}{2v^2}\right), \qquad L \simeq \left(\frac{\omega}{2\sigma}\right)^{1/2}(1+i). \tag{2.7}$$

Here the wave number K corresponds to the acoustic wave of length $2\pi v/\omega$ with a small attenuation $\alpha = \omega^2\sigma(\gamma-1)/2v^3$. The wave number L corresponds to a much shorter wavelength disturbance due to heat conduction called the *thermal wave*.

We can therefore express the velocity potential as the sum of a compressional wave potential φ and a thermal wave potential ψ:

$$\Phi = e^{-i\omega t}\left(\varphi + \psi + \varphi_0\right),$$

where $\varphi_0 = e^{iKz}$ is the incoming wave and φ and ψ satisfy separate Helmholtz equations

$$(\nabla^2 + K^2)\varphi = 0 \quad \text{and} \quad (\nabla^2 + L^2)\psi = 0 \quad \text{in} \quad D_1. \tag{2.8}$$

The pressure and temperature perturbations are given by

$$P = -i\omega\rho e^{-i\omega t}(\varphi + \psi + \varphi_0), \qquad T = e^{-i\omega t}(\Gamma_c\varphi + \Gamma_t\psi + \varphi_0),$$

where

$$\Gamma_c = \frac{-iK^2\omega(\gamma-1)}{\beta(\omega + i\gamma\sigma K^2)} \quad \text{and} \quad \Gamma_t = \frac{-iL^2\omega(\gamma-1)}{\beta(\omega + i\gamma\sigma L^2)}.$$

Equivalent equations hold inside the droplet, so that

$$\Phi = e^{-i\omega t}\left(\varphi' + \psi'\right),$$

$$(\nabla^2 + K'^2)\varphi' = 0 \quad \text{and} \quad (\nabla^2 + L'^2)\psi' = 0 \quad \text{in} \quad D_2. \tag{2.9}$$

We use primes to denote quantities in the droplet phase. The wave numbers K' and L' are given by (2.7), but using the parameter values of the droplet. Finally we apply boundary conditions on the potential fields φ, ψ , φ', and ψ', where φ and ψ must satisfy the Sommerfeld radiation condition at infinity (see [15, 17]):

$$\lim_{r\to\infty} r\left(\frac{\partial\varphi}{\partial r} - iK\varphi\right) = 0, \qquad \lim_{r\to\infty} r\left(\frac{\partial\psi}{\partial r} - iL\psi\right) = 0. \tag{2.10}$$

In addition, we require that the normal velocity, pressure, temperature, and heat flux be continuous at the droplet boundary, giving the following four boundary conditions on the boundary B:

(a) normal velocity

$$\frac{\partial}{\partial n}\left(e^{iKz} + \varphi + \psi\right) = \frac{\partial}{\partial n}\left(\varphi' + \psi'\right), \tag{2.11}$$

(b) pressure

$$e^{iKz} + \varphi + \psi = \hat{\rho}\left(\varphi' + \psi'\right), \quad \hat{\rho} = \frac{\rho'}{\rho}, \tag{2.12}$$

(c) temperature

$$\Gamma_c\left(e^{iKz} + \varphi\right) + \Gamma_t\psi = \Gamma'_c\varphi' + \Gamma'_t\psi', \tag{2.13}$$

(d) heat flux

$$\Gamma_c\frac{\partial}{\partial n}\left(e^{iKz} + \varphi\right) + \Gamma_t\frac{\partial}{\partial n}\psi = \hat{\tau}\left(\Gamma'_c\frac{\partial}{\partial n}\varphi' + \Gamma'_t\frac{\partial}{\partial n}\psi'\right), \quad \hat{\tau} = \frac{\tau'}{\tau}. \tag{2.14}$$

For the case of a spherical droplet, a solution for the fields φ, ψ, φ', and ψ' may be found as a Rayleigh series expansion in spherical harmonics (see [1]),

$$\varphi = \sum_{n=0}^{\infty} i^n(2n+1)A_n h_n\,(Kr)\,P_n(\cos\theta), \quad \varphi' = \sum_{n=0}^{\infty} i^n(2n+1)A'_n j_n\,(K'\,r)\,P_n(\cos\theta),$$

$$\psi = \sum_{n=0}^{\infty} i^n(2n+1)B_n h_n\,(Lr)\,P_n(\cos\theta), \quad \psi' = \sum_{n=0}^{\infty} i^n(2n+1)B'_n j_n\,(L'\,r)\,P_n(\cos\theta),$$

$$\varphi_0 = \sum_{n=0}^{\infty} i^n(2n+1)j_n\,(Kr)\,P_n(\cos\theta) = e^{iKz}, \tag{2.15}$$

where $j_n(z)$ and $h_n(z)$ are, respectively, the nth order spherical Bessel and Hankel functions [18]. The coefficients $A_n, B_n, A_n{}', \text{and } B_n{}'$ are found by enforcing the boundary conditions (2.11)–(2.14).

3. Low frequency potential scattering. As mentioned in the introduction, while the ECAH solution given by (2.15) is exact [4], it suffers from a number of limitations, most significantly that its implementation leads to ill-conditioning problems. To eliminate this problem, the authors of [4] sought an asymptotic solution in the limit when the particle size a is such that $|Ka|$ and $|La| \ll 1$. This was achieved by appealing to a method first presented by Kleinman [6] for the solution of low frequency acoustic scattering problems. The method reduced the Helmholtz equation to a sequence of potential problems by taking an integral transformation. We now give a brief outline of the process. First, in solving the Helmholtz equations,

$$\left(\nabla^2 + K^2\right)\varphi = 0 \quad \text{in} \quad D_1, \tag{3.1}$$

$$\left(\nabla^2 + K'^2\right)\varphi' = 0 \quad \text{in} \quad D_2, \tag{3.2}$$

in which φ satisfies the radiation condition at infinity in D_1, we introduce the function

$$\tilde{\varphi} = e^{-iKr}\varphi, \tag{3.3}$$

which is regular at infinity, and we write (3.1) in the form

$$\nabla^2\tilde{\varphi} = -\frac{2iK}{r}\frac{\partial}{\partial r}\,(r\tilde{\varphi}). \tag{3.4}$$

A real valued function is said to be regular at infinity if

$$\lim r \to \infty, \qquad |\, rf(p)\,| < \infty, \qquad \text{and} \quad \lim r \to \infty \left| r^2 \frac{\partial(p)}{\partial r} \right| < \infty,$$

where $r = |x|$ is the magnitude of the position vector x of an arbitrary point. Boundary conditions on the surface connecting the two domains are transformed in terms of the variable $\tilde{\varphi}$. To complete the process, $\tilde{\varphi}$ and φ' are expressed as regular perturbation expansions:

$$\tilde{\varphi} = \sum_{n=0}^{\infty} (iKa)^n\tilde{\varphi}_n, \qquad \varphi' = \sum_{n=0}^{\infty} (iK'a)^n\varphi'_n. \tag{3.5}$$

For $|Ka| < \ln(2)$ this series converges (see [6]), and in practice it is rapidly convergent. This offers a significant improvement on ECAH formulation since the error is bounded by $O(|Ka|^{m+1})$ if an mth order solution is used [4]. In other words, as long as $|Ka|$ is small, we are guaranteed to obtain an accurate estimation of the potential with only a few terms. In contrast, due to the ill-conditioned nature of the ECAH system, it is difficult to know if we have obtained an accurate estimation of the scattered potential.

If we substitute the expansions (3.5) into (3.2) and (3.4), we obtain, for $m \geq 0$, the equivalent system:

$$\nabla^2 \tilde{\varphi}_m = -\frac{2}{ar}\frac{\partial}{\partial r}(r\tilde{\varphi}_{m-1}) \qquad \text{in } D_1, \tag{3.6}$$

together with

$$\nabla^2 \varphi'_m = \frac{1}{a^2}\varphi'_{m-2} \qquad \text{in } D_2, \tag{3.7}$$

with the appropriate modifications to the boundary conditions. In (3.6) and (3.7) we understand that

$$\tilde{\varphi}_{-1} = \tilde{\psi}_{-1} = 0 \quad \text{and} \quad \varphi'_{-1} = \psi'_{-1} = \varphi'_{-2} = \psi'_{-2} = 0.$$

If the fields are given this asymptotic description, then clearly the solution will only be valid in the long wavelength limit. In this paper we also seek to examine the short wavelength limit for which $|La| \gg 1$; this is addressed by the geometric theory of diffraction [13] which we now introduce.

4. The geometric theory of diffraction. Partial differential equations play a crucial part in many branches of mathematics, physics, and industry; however, since explicit exact solutions exist for only relatively few problems, numerical and asymptotic techniques were developed in a bid to elucidate the underlying physics or mechanisms. The notion of asymptotic solutions was first conceived to provide the functional dependence of partial differential equations upon their parameters and data. The main feature of the method relies on *rays*, which are curves along which the terms of the asymptotic expansion satisfy ordinary differential equations. This method was successfully applied by Lewis and Keller [13] to the Helmholtz equation, revealing the phenomenon of diffraction, from which the method gets its name. As a starting point consider the following equation:

$$(\nabla^2 + L^2 n^2(x))\psi = 0; \tag{4.1}$$

here the function $n(x)$ relates to the refractive index of the media. In the case of homogeneous media, $n(x) = constant$ and as such admits plane wave solutions of the form

$$\psi = \tilde{\psi} e^{in(x)\underline{L}\cdot\mathbf{r}}, \tag{4.2}$$

where $\underline{L} = L\hat{L}$ is the propagation vector, $\hat{L}$ the unit vector, and L the wave number. On the basis of (4.2), solutions are sought of the form

$$\psi = \tilde{\psi} e^{iLs(x)}; \tag{4.3}$$

substituting this expression into (4.1) and canceling the factor $\exp(iLs)$ yields

$$-L^2\{(\nabla s)^2 - n^2\}\tilde{\psi} + 2iL\nabla s \cdot \nabla\tilde{\psi} + iL\tilde{\psi}\triangle s + \triangle\tilde{\psi} = 0. \tag{4.4}$$

For large values of L we assume an expansion in inverse powers of iLa of the form

$$\tilde{\psi} = \sum_{m=0}^{\infty} (iLa)^{-m} \tilde{\psi}_m, \tag{4.5}$$

in contrast with the expansions in (3.5). Inserting this expression into (4.4) and equating in powers of m gives

$$-(La)^2\{(\nabla s)^2 - n^2\}\tilde{\psi}_{m+1} + 2iLa\nabla s \cdot \nabla\tilde{\psi}_m + iLa\tilde{\psi}_m \triangle s + a\triangle\tilde{\psi}_{m-1} = 0. \tag{4.6}$$

Again here we understand that $\tilde{\psi_m} = 0$ for $m = -1, -2, \ldots$. For $m = -1$ we have

$$\{(\nabla s)^2 - n^2\}\tilde{\psi}_0 = 0, \tag{4.7}$$

and if $\tilde{\psi}_0 \neq 0$ yields the Eiconal equation,

$$(\nabla s)^2 = n^2(x). \tag{4.8}$$

For $m = 0$, in order to obtain vanishing coefficients, we have

$$2\nabla s \cdot \nabla\tilde{\psi}_0 + \tilde{\psi}_0 \triangle s = 0, \tag{4.9}$$

and for $m = 1, 2, \ldots$,

$$2\nabla s \cdot \nabla\tilde{\psi}_m + \tilde{\psi}_m \triangle s = -\triangle\tilde{\psi}_{m-1}. \tag{4.10}$$

Equations (4.9)–(4.10) are called transport equations and are analogous to (3.6) in the low frequency case. For surfaces of constant phase defined by $s(x) = constant$, curves or rays orthogonal to them provide solutions to the Eiconal equation. One may then express the equation of a ray in the form

$$x = (x_1, x_2, x_3) = x(\zeta), \tag{4.11}$$

and the orthogonality condition is expressed by

$$\frac{dx_j}{d\zeta} = \lambda s_{x_j}, \quad j = 1, 2, 3. \tag{4.12}$$

Here $\lambda(x)$ is an arbitrary proportionality factor and for the choice $\lambda = n^{-1}$ ensures that the parameter ζ is simply the arc length along a ray. Utilizing (4.8) and (4.12), we may write

$$\frac{ds(x)}{d\zeta} = \nabla s \cdot \frac{dx}{d\zeta} = \lambda(\nabla s)^2 = \lambda n^2; \tag{4.13}$$

upon integrating this expression with respect to ζ, we have the solution as

$$s(x(\zeta)) = s(x(\zeta_0)) + \int_{\zeta}^{\zeta_0} \lambda(x(t))n^2(x(t))dt, \tag{4.14}$$

and for the choice $\lambda = n^{-1}$ this reduces to

$$s(\zeta) = s(\zeta_0) + \int_{\zeta}^{\zeta_0} n(t)dt. \tag{4.15}$$

Similarly, the transport equations (4.9)–(4.10) may be expressed in terms of the parameter ζ. The crucial aspect of this formulation is that the original partial differential equation has been confined to the solution of a system of ordinary differential equations, and that the geometry of the problem manifests itself only in terms of the parameter ζ. In this paper we assume our media are homogenous, in the sense that the rays are straight lines. As a simple example to illustrate this, we consider a spherical geometry; this has rays which are simply radial lines, and therefore

$$s(r) = \pm r + constant. \tag{4.16}$$

This furnishes us with the following forms for (4.3):

$$\psi = \tilde{\psi} e^{\pm iLr}. \tag{4.17}$$

A full treatment of the theory is given in [13]. We are now in a position to formulate (2.8)–(2.9) together with the associated boundary conditions (2.11)–(2.14).

5. Formulation. Based on the previous sections, we now apply the following transformations:

$$\tilde{\varphi} = e^{-iK(r-a)}\varphi,$$

$$\tilde{\psi} = e^{-iL(r-a)}\psi, \qquad \tilde{\psi}' = e^{iL'(r-a)}\psi'.$$

Thus (2.8)–(2.9) become

$$\nabla^2\tilde{\varphi} = -\frac{2iK}{r}\frac{\partial}{\partial r}\left(r\tilde{\varphi}\right), \tag{5.1}$$

$$\nabla^2\varphi' + {K'}^2\varphi' = 0, \tag{5.2}$$

$$2iL\nabla s_1 \cdot \nabla\tilde{\psi} + iL\tilde{\psi}\triangle s_1 + \triangle\tilde{\psi} = 0, \tag{5.3}$$

$$2iL'\nabla s_2 \cdot \nabla\tilde{\psi}' + iL'\tilde{\psi}'\triangle s_2 + \triangle\tilde{\psi}' = 0, \tag{5.4}$$

where $s_1(r) = r + constant$ and $s_2(r) = -r + constant$. The boundary conditions of the system now transform to the following:

$$iK\cos\theta e^{iKa\cos\theta} + \left\{iK\tilde{\varphi} + \frac{\partial\tilde{\varphi}}{\partial r}\right\} + \left\{iL\tilde{\psi} + \frac{\partial\tilde{\psi}}{\partial r}\right\} = \frac{\partial\varphi'}{\partial r} + \left\{\frac{\partial\tilde{\psi}'}{\partial r} - iL'\tilde{\psi}'\right\}, \tag{5.5}$$

$$e^{iKa\cos\theta} + \tilde{\varphi} + \tilde{\psi} = \hat{\rho}\{\varphi' + \tilde{\psi}'\}, \tag{5.6}$$

$$\Gamma_c\{e^{iKa\cos\theta} + \tilde{\varphi}\} + \Gamma_t\tilde{\psi} = \Gamma'_c\varphi' + \Gamma'_t\tilde{\psi}', \tag{5.7}$$

$$\Gamma_c\left\{iK\cos\theta e^{iKa\cos\theta}+iK\tilde{\varphi}+\frac{\partial\tilde{\varphi}}{\partial r}\right\}+\Gamma_t\left\{iL\tilde{\psi}+\frac{\partial\tilde{\psi}}{\partial r}\right\}$$

$$=\hat{\tau}\left\{\Gamma_c'\frac{\partial\varphi'}{\partial r}+\Gamma_t'\left\{\frac{\partial\tilde{\psi}'}{\partial r}-iL'\tilde{\psi}'\right\}\right\}. \tag{5.8}$$

We observe that for large La, by dividing the boundary conditions in (5.7)–(5.8) by Γ_t, we have the following ratios:

$$\frac{\Gamma_c}{\Gamma_t}\simeq\frac{(1-\gamma)K^2}{L^2-\gamma K^2}.$$

Thus we define

$$\frac{\Gamma_c}{\Gamma_t}\simeq\frac{G_cK^2}{L^2},$$

where $G_c=(1-\gamma)$, and similarly,

$$\frac{\Gamma_c'}{\Gamma_t}\simeq\frac{G_c'K'^2}{L^2},\qquad G_c'=\frac{(\gamma'-1)\beta}{\beta'},$$

$$\frac{\Gamma_t'}{\Gamma_t}\simeq G_t'=\frac{\sigma\beta}{\sigma'\beta'}.$$

The boundary conditions given in (5.7)–(5.8) now may be written as

$$G_c\frac{(Ka)^2}{(La)^2}\{e^{iKa\cos\theta}+\tilde{\varphi}\}+\tilde{\psi}=G_c'\frac{(K'a)^2}{(La)^2}\varphi'+G_t'\tilde{\psi}', \tag{5.9}$$

$$G_c\frac{(Ka)^2}{(La)^2}\left\{iK\cos\theta e^{iKa\cos\theta}+iKa\tilde{\varphi}+\frac{\partial\tilde{\varphi}}{\partial r}\right\}+\left\{iLa\tilde{\psi}+\frac{\partial\tilde{\psi}}{\partial r}\right\}$$

$$=\hat{\tau}\left\{G_c'\frac{(K'a)^2}{(La)^2}\frac{\partial\varphi'}{\partial r}+G_t'\left\{\frac{\partial\tilde{\psi}'}{\partial r}-iLa'\tilde{\psi}'\right\}\right\}. \tag{5.10}$$

Recognizing this scaling proved to be particularly important in that it simplifies the boundary conditions significantly. Essentially it decouples the thermal field from the acoustic field to low order. One can see that the acoustic potential contributes to the temperature perturbation only at second order in Ka. Consequently the thermal field will be of order $(Ka)^2$, and thus it will only affect the scattered potential at this order.

We now introduce expansions of the following form:

$$\tilde{\varphi}=\sum_{n=0}^{\infty}\sum_{m=0}^{\infty}\frac{(iKa)^n}{(iLa)^m}\tilde{\varphi}_{nm},\qquad\varphi'=\sum_{n=0}^{\infty}\sum_{m=0}^{\infty}\frac{(iK'a)^n}{(iL'a)^m}\varphi'_{nm}, \tag{5.11}$$

$$\tilde{\psi} = \sum_{n=0}^{\infty}\sum_{m=0}^{\infty}\frac{(iKa)^n}{(iLa)^m}\tilde{\psi}_{nm}, \qquad \tilde{\psi}' = \sum_{n=0}^{\infty}\sum_{m=0}^{\infty}\frac{(iK'a)^n}{(iL'a)^m}\tilde{\psi}'_{nm}. \tag{5.12}$$

Following the subsequent substitution into (5.1)–(5.4) and the associated boundary conditions (5.5)–(5.6) and (5.9)–(5.10), we obtain the following equivalent transport equations:

$$\nabla^2\tilde{\varphi}_{nm} = -\frac{2}{ar}\frac{\partial}{\partial r}\left(r\tilde{\varphi}_{n-1,m}\right) \qquad \text{in } D_1, \tag{5.13}$$

together with

$$\nabla^2\varphi'_{nm} = \frac{1}{a^2}\varphi'_{n-2,m} \qquad \text{in } D_2, \tag{5.14}$$

$$2\nabla s_1\cdot\nabla\tilde{\psi}_{n0} + \tilde{\psi}_{n0}\triangle s_1 = 0 \qquad \text{in } D_1, \tag{5.15}$$

and, for $m = 1, 2, \ldots,$

$$2\nabla s_1\cdot\nabla\tilde{\psi}_{nm} + \tilde{\psi}_{nm}\triangle s_1 = -a\triangle\tilde{\psi}_{n,m-1} \qquad \text{in } D_1, \tag{5.16}$$

$$2\nabla s_2\cdot\nabla\tilde{\psi'}_{n0} + \tilde{\psi'}_{n0}\triangle s_2 = 0 \qquad \text{in } D_2, \tag{5.17}$$

and, for $m = 1, 2, \ldots,$

$$2\nabla s_2\cdot\nabla\tilde{\psi'}_{nm} + \tilde{\psi'}_{nm}\triangle s_2 = -a\triangle\tilde{\psi'}_{n,m-1} \qquad \text{in } D_2. \tag{5.18}$$

Together with the accompanying sequences for the boundary conditions, the systems of equations (5.13)–(5.18) are well posed. Of course in the above we understand that $\tilde{\varphi}_{-1,m} = 0$, $\varphi'_{-1,m} = \varphi'_{-2,m} = 0$ for all integer values m, and $\tilde{\psi}_{nm} = 0, \tilde{\psi}'_{nm} = 0$ for all integer values n and $m = -1, -2, \ldots$.

5.1. The leading order solutions. The leading order terms in the boundary conditions given in (5.5)–(5.6) and (5.9)–(5.10) are given by

$$iL\tilde{\psi_{n0}} + iL'\tilde{\psi}'_{n0} = 0, \tag{5.19}$$

$$iL\tilde{\psi_{n0}} + iL'\hat{\tau}G'_t\tilde{\psi}'_{n0} = 0, \tag{5.20}$$

and thus, provided that

$$\hat{\tau}G'_t \neq 1, \qquad \text{then} \qquad \tilde{\psi}_{n0} \equiv \tilde{\psi'}_{n0} \equiv 0. \tag{5.21}$$

Similarly, order one terms from (5.5) and $O(La^{-1})$ terms from (5.10) give

$$\tilde{\psi_{n1}} + \tilde{\psi}'_{n1} = 0, \tag{5.22}$$

$$\tilde{\psi_{n1}} + \hat{\tau}G'_t\tilde{\psi}'_{n1} = 0, \tag{5.23}$$

again provided that

$$\hat{\tau}G'_t \neq 1, \qquad \text{then} \qquad \tilde{\psi}_{n1} \equiv \tilde{\psi'}_{n1} \equiv 0, \qquad n = 0, 1, 2, \ldots. \tag{5.24}$$

The remaining order one boundary conditions from (5.5)–(5.6) are

$$\frac{\partial \tilde{\varphi}_{00}}{\partial r} = \frac{\partial \varphi'_{00}}{\partial r}, \tag{5.25}$$

$$1 + \tilde{\varphi}_{00} = \hat{\rho}\varphi'_{00}, \tag{5.26}$$

$$\Rightarrow \varphi'_{00} = \frac{1}{\hat{\rho}}, \qquad \varphi_{00} = 0.$$

In the above we have expanded the incident field as

$$e^{iKz} = \{1 + iKz + \cdots\}$$

and used only the first harmonic. Here we see the decoupling effect of the scaling, as mentioned earlier. The zero order acoustic potentials do not depend upon the thermal potentials. The corresponding transport equations are given by

$$\nabla^2 \tilde{\varphi}_{00} = 0 \qquad \text{in } D_1, \tag{5.27}$$

$$\nabla^2 \varphi'_{00} = 0 \qquad \text{in } D_2, \tag{5.28}$$

$$\tilde{\varphi}_{00} = \frac{A_0^{00} a}{r}, \qquad \varphi'_{00} = C_0^{00}; \tag{5.29}$$

with the associated solutions, (5.25)–(5.26) then yield $A_0^{00} = 0$ and $C_0^{00} = 1/\hat{\rho}$.

In general, from the decomposition of the incident field into spherical harmonics, we can deduce that the associated solutions $\tilde{\varphi}_{nm}, \varphi'_{nm}, \tilde{\psi}_{nm}$, and $\tilde{\psi}'_{nm}$ will contain the first n spherical harmonics. Thus one could expand the incident field into spherical harmonics, given in (2.15), and subsequently deduce that $\tilde{\varphi}$ and φ', which have the associated solutions

$$\tilde{\varphi}_{00} = \sum_{n=0}^{\infty} \frac{A_n^{00} a^{n+1} P_n(\cos\theta)}{r^{n+1}}, \qquad \varphi'_{00} = \sum_{n=0}^{\infty} \frac{C_n^{00} P_n(\cos\theta) r^n}{a^n}, \tag{5.30}$$

yield the solution

$$A_n^{00} = \frac{-I_n n}{n + (n+1)\hat{\rho}}, \qquad C_n^{00} = \frac{I_n (n+1)}{n + (n+1)\hat{\rho}},$$

where $I_n = i^n(2n+1)j_n(Ka)$ and j_n is the nth spherical Bessel function. Thus we have for $n = 0$, $A_0^{00} = 0$ and $C_0^{00} = 1/\hat{\rho}$, as above. In practice, the zero order term dominates the behavior of the solution.

At order (Ka) we have

$$\cos\theta + a\frac{\partial \tilde{\varphi}_{10}}{\partial r} = \frac{K}{K'} a \frac{\partial \varphi'_{10}}{\partial r}, \tag{5.31}$$

$$\cos\theta + \tilde{\varphi}_{10} = \frac{K}{K'} \hat{\rho} \varphi'_{10}, \tag{5.32}$$

together with the associated transport equations

$$\nabla^2 \tilde{\varphi}_{10} = 0 \qquad \text{in } D_1, \tag{5.33}$$

$$\nabla^2 \varphi'_{10} = 0 \qquad \text{in } D_2, \tag{5.34}$$

and solutions

$$\tilde{\varphi}_{10} = \frac{A_0^{10} a}{r} + \frac{A_1^{10} a^2}{r^2} P_1(\cos\theta), \qquad \varphi'_{10} = C_0^{10} + \frac{C_1^{10} r}{a} P_1(\cos\theta), \tag{5.35}$$

where

$$A_0^{10} = C_0^{10} = 0, \qquad A_1^{10} = \frac{\hat{\rho} - 1}{1 + 2\hat{\rho}}, \qquad C_1^{10} = \frac{3}{1 + 2\hat{\rho}},$$

and $P_n(\cos\theta)$ is the nth spherical Legendre polynomial.

To order $((Ka)^2)$ from (5.14)–(5.15), we have the following transport equations:

$$\nabla^2 \tilde{\varphi}_{20} = -\frac{2}{ar} \frac{\partial}{\partial r} (r \tilde{\varphi}_{10}) \qquad \text{in } D_1, \tag{5.36}$$

$$\nabla^2 \varphi'_{20} = \frac{1}{a^2} \varphi'_{00} \qquad \text{in } D_2, \tag{5.37}$$

with solutions of

$$\tilde{\varphi}_{20} = \frac{A_0^{20} a}{r} + \frac{A_1^{20} a^2}{r^2} P_1(\cos\theta) + \frac{A_2^{20} a^3}{r^3} P_2(\cos\theta) - \frac{A_1^{10} a}{r} P_1(\cos\theta),$$

$$\varphi'_{10} = C_0^{20} + \frac{C_1^{20} r}{a} P_1(\cos\theta) + \frac{C_1^{20} r^2}{a^2} P_2(\cos\theta) + \frac{C_0^{00} r^2}{6a^2}, \tag{5.38}$$

where

$$A_0^{20} = \frac{\hat{\rho} - 1}{3\hat{\rho}}, \qquad A_1^{20} = \frac{\hat{\rho} - 1}{1 + 2\hat{\rho}}, \qquad A_2^{20} = \frac{2(\hat{\rho} - 1)}{3(2 + 3\hat{\rho})},$$

$$C_0^{20} = \frac{(3\hat{\rho} - 1)\hat{c}}{6\hat{\rho}^2} - \frac{1}{6\hat{\rho}}, \qquad C_1^{20} = 0, \qquad C_2^{20} = \frac{(5\hat{\rho})}{3(2 + 3\hat{\rho})\hat{\rho}\hat{c}},$$

where $\hat{c} = K'^2/K^2$.

Continuing in this way, we construct the functions $\tilde{\varphi}_{nm}, \varphi'_{nm}, \tilde{\psi}_{nm}$, and $\tilde{\psi}'_{nm}$. At order $(Ka)^2(La)^{-1}$ we observe the first interaction between the thermal and acoustic fields with the following boundary conditions:

$$\tilde{\varphi}_{11} + a \frac{\partial \tilde{\varphi}_{21}}{\partial r} + \tilde{\psi_{22}} = \hat{c}\hat{d} \left\{ a \frac{\partial \varphi'_{21}}{\partial r} - \tilde{\psi'_{22}} \right\}, \tag{5.39}$$

$$\tilde{\varphi}_{21} = \hat{\rho}\, \hat{c}\, \hat{d}\, \varphi'_{21}, \tag{5.40}$$

$$\tilde{\psi}_{22} = -\hat{\tau}\,\hat{c}\,\hat{d}\,G_t'\,\psi_{22}', \tag{5.41}$$

where $\hat{d} = L/L'$. Similar expressions for the order $(Ka)^2(La)^{-2}$ terms together with solutions, developed from the transport equations (5.15)–(5.18), allow determination of the coefficients A_0^{21} and A_0^{22} for the functions

$$\tilde{\varphi}_{21} = \frac{A_0^{21}a}{r} + \frac{A_1^{21}a^2}{r^2}P_1(\cos\theta) + \frac{A_2^{21}a^3}{r^3}P_2(\cos\theta), \tag{5.42}$$

$$\tilde{\varphi}_{22} = \frac{A_0^{22}a}{r} + \frac{A_1^{22}a^2}{r^2}P_1(\cos\theta) + \frac{A_2^{22}a^3}{r^3}P_2(\cos\theta). \tag{5.43}$$

Solutions to the transport equations (5.15) and (5.18) for $n = 2$ and $m = 0, 1, 2$ are of the form

$$\tilde{\psi}_{20} = \frac{D_0^{20}a}{r} + \frac{D_1^{20}a}{r}P_1(\cos\theta) + \frac{D_2^{20}a}{r}P_2(\cos\theta),$$

$$\tilde{\psi}_{21} = \frac{D_0^{22}a}{r} + \frac{D_1^{22}a}{r}P_1(\cos\theta) + \frac{D_2^{22}a}{r}P_2(\cos\theta) - \frac{D_1^{21}a^2}{r^2}P_1(\cos\theta) - \frac{3D_2^{21}a^2}{r^2}P_2(\cos\theta),$$

$$\tilde{\psi}_{22} = \frac{D_0^{22}a}{r} + \frac{D_1^{22}a}{r}P_1(\cos\theta) + \frac{D_2^{22}a}{r}P_2(\cos\theta) - \frac{D_1^{21}a^2}{r^2}P_1(\cos\theta)$$

$$- \frac{3D_2^{21}a^2}{r^2}P_2(\cos\theta) + \frac{3D_2^{20}a^3}{r^3}P_2(\cos\theta),$$

with similar expressions for $\tilde{\psi}'_{2m}$, $m = 0, 1, 2$. However, due to (5.21)–(5.24), considerable simplification is achieved, allowing us to derive the following quantities:

$$A_0^{21} = \left\{\frac{(1-\hat{\tau}G_t')}{\hat{\tau}G_t'}\right\}\left\{\frac{\hat{\tau}(\alpha L^2 G_c' - G_c L'^2\hat{\rho})}{\hat{\rho}L'\,(L'\hat{\tau}+L)}\right\},$$

$$A_0^{22} = \left\{\frac{(L^2 - L'^2\hat{\tau})}{L'(L'\hat{\tau}+L)} - \frac{(\hat{c}-1)\hat{d}}{\hat{\tau}G_t'}\right\}\left\{\frac{\hat{\tau}(\alpha L^2 G_c' - G_c L'^2\hat{\rho})}{\hat{\rho}L'\,(L'\hat{\tau}+L)}\right\}.$$

6. The far field. For the far field we have the following theorem due to Atkinson [15], Barrar and Kay [16], and Wilcox [17].

THEOREM 6.1. *Let u be the radiating solution to the Helmholtz equation,*

$$\bigtriangledown^2 u + k^2 u = 0 \qquad in \quad V_1;$$

then the field scattered from the surface B has an asymptotic form of a spherical wave,

$$u(r) = \frac{e^{ikr}}{r}\sum_{n=0}^{\infty}\frac{f_n(\phi,\Omega)}{r^n}, \tag{6.1}$$

where the series converges absolutely and uniformly for $r > c + \epsilon$, $\epsilon > 0$. Here c is defined as $c = \max_{r\in B} r$. Furthermore, the series may be differentiated term by term

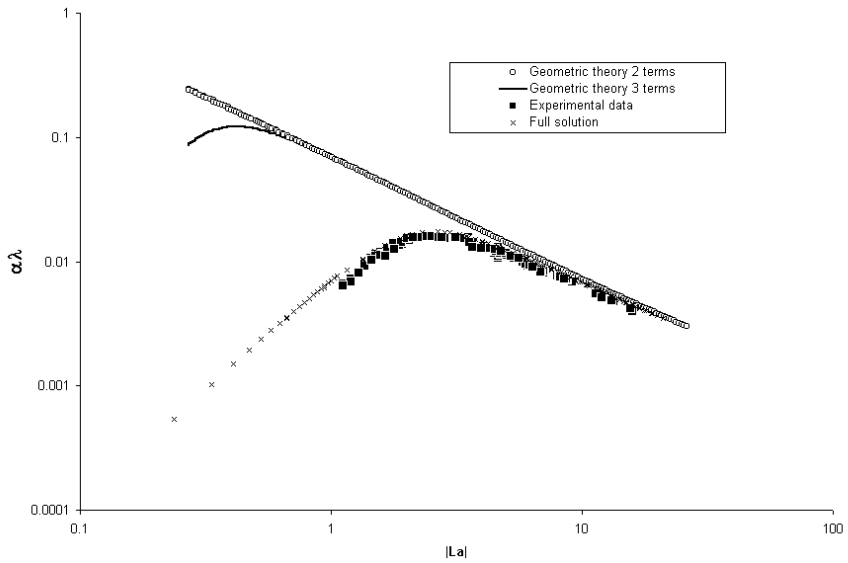


FIG. 6.1. *Plot of $\alpha\lambda$, the attenuation per wavelength, against La, the product of thermal wave number and particle radius, for a silicone (polysiloxane) oil-in-water emulsion. Solid squares denote experimental data for 5% volume concentration of oil, the black line shows the high frequency formulation using the far field pattern in* (6.2)*, and the crosses indicate the ECAH calculation. The open squares show the calculation using only the first two terms in* (6.2)*.*

with r, ϕ, and Ω any number of times, and the resulting series all converge absolutely and uniformly. The functions $f_n(\phi, \Omega)$ are understood to depend on the parameter k, and the function $f_0(\phi)$ is called the far field.

In the above formulation, one can see that the far field pattern for the reflected acoustic field is given by

$$f_0(\phi) = -K^2a^3\left\{A_0^{20} - \frac{iA_0^{21}}{L} - \frac{A_0^{22}}{L^2}\right\} + O(|Ka|^3) \quad \text{as} \quad \mathrm{r} \to \infty, \tag{6.2}$$

where ψ does not contribute due to the nonpropagational character of the thermal wave. The term A_0^{20} contained in the far field pattern derives from the acoustic scattering of the particle, depends upon only the ratio of densities of the two phases, and is independent of the thermal properties. In contrast, the terms A_0^{21} and A_0^{22} are dependent upon the thermal properties of the phases, the former being the dominant term and the coefficient in $(La)^{-1}$. This scaling is seen in Figure 6.1.

6.1. Comparison with experiments. We now compare the high frequency geometric theory of diffraction solution, the ECAH solution, and experimental data. This necessitates the use of multiple particle scattering theory. Lloyd and Berry [14] showed that

$$\left(\frac{B}{K}\right)^2 = 1 + \frac{3\phi f(0)}{K^2a^3} + \frac{9\phi^2 f(0)}{4K^4a^6}\left\{f^2(\pi) - f^2(0) - \int_0^\pi d\theta \frac{1}{\sin(\theta/2)}\frac{d}{d\theta}f^2(\theta)\right\}, \tag{6.3}$$

where B is the wave number of the acoustic wave in the dispersion and $\phi = 4\pi a^3 n_0/3$ is the droplet volume fraction. We have used the above multiple scattering result

TABLE 1
Physical data set for experimental and theoretical emulsions.

	n-Hexadecane	Silicone oil	Aqueous phase
Sound velocity $v(m\ s^{-1})$	1357.9	1004	1482
Density $\rho(kg\ m^{-3})$	773.0	975	998.2
Thermal expansivity $\beta(K^{-1})$	9.1×10^{-4}	9.4×10^{-4}	2.13×10^{-4}
Specific heat $C_p(J\ kg^{-1}K^{-1})$	2215	1460	4182
Thermal conductivity $\tau(W\ m^{-1}K^{-1})$	0.143	0.15	0.591

in what follows, even though we showed in [4] that (6.3) fails to accurately describe multiple scattering in the presence of thermal fields since it assumes that the scatterers are points and neglects the interpenetrating character of the thermal fields. However, we compare our theoretical result with the experimental data only at the lowest concentrations, where the multiple scattering equation (6.3) is most accurate. Figure 6.1 compares experimental data, the theoretical calculations of ECAH, and the above geometric theory of diffraction solution. Since the geometric theory solution is zero order in K, K', one can infer the dominating effect of thermal scattering in acoustic propagation in dispersions, indicating that a significant component of the acoustic attenuation results purely from thermal effects. As can be seen from the figure, the high frequency formulation gives excellent agreement with both experimental results and ECAH for $|La| \gg 1$ and, furthermore, is easy to calculate, providing a very practical solution. The calculations were performed using the parameter values given in Table 1 and compared with experiments performed by Hemar et al. [19]. The data relates to particle sizes in the range 230nm to 760nm for silicone oil-in-water over a range of frequencies from 0.5MHz to 10MHz. For these parameter values the values of Ka ranged from 10^{-4} up to 10^{-2}, and the data satisfy $|Ka| \ll 1$. We present comparisons for the case in which the volume concentration of oil is 5%, since for higher concentrations it is known that multiple scattering of the thermal field makes a significant contribution; see [4].

7. Conclusions. We have obtained an approximation to the ECAH solution for acoustic scattering from a single particle for the acoustic long wavelength regime, when the associated thermal field is in the short wavelength regime. This approximation is well conditioned and well posed and can be computed to arbitrary accuracy. It may be generalized to arbitrary particle shape using boundary-integral and finite-element techniques. This solution for $|La| \gg 1$ complements our previous approximation for $|La| \ll 1$ given in [4]. As noted in [11], since $\sigma\omega/v^2 \ll 1$, $|La| \gg Ka$, and so it is the condition $|La| > 1$ that causes LFPST to fail at higher frequencies even though the acoustic wavelength is still large compared to the particle radius. Together with the LFPST approximation, we now have a set of simple and well-conditioned approximations for the scattering throughout the entire range of low frequency scattering experiments, which provides a further step towards complete understanding of ultrasound propagation through systems of weakly interacting particles in the acoustic long wavelength limit. We have still to address the question raised in [11] concerning the failure of current multiple scattering formulations to adequately deal with thermal scattering.

REFERENCES

[1] A.S. DUKHIN AND P.J. GOETZ, *Acoustic and electroacoustic spectroscopy for characterizing concentrated dispersions and emulsions.*, Adv. Colloid Int. Sci., 92 (2001), pp. 73–132.

[2] P.S. Epstein and R.R. Carhart, *The absorption of sound in suspensions and emulsions* I. *Water fog in air*, J. Acoust. Soc. Amer., 25 (1953), pp. 553–565.
[3] J.R. Allegra and S.A. Hawley, *Attenuation of sound in suspensions and emulsions: Theory and experiments*, J. Acoust. Soc. Amer., 51 (1972), pp. 1545–1564.
[4] O.G. Harlen, M.J. Holmes, M.J.W. Povey, Y. Qiu, and B.D. Sleeman, *A low frequency potential scattering description of acoustic propagation in dispersions*, SIAM J. Appl. Math., 61 (2001), pp. 1906–1931.
[5] J.W. Strutt (Baron Rayleigh), *The Theory of Sound*, 2nd ed., Macmillan, London, 1896.
[6] R.E. Kleinman, *The Dirichlet problem for the Helmholtz equation*, Arch. Ration. Mech. Anal., 18 (1965), pp. 205–229.
[7] D.J. McClements, *Principles of ultrasonic droplet size determination in emulsions*, Langmuir, 12 (1996), pp. 3454–3461.
[8] D.J. McClements and M.J.W. Povey, *Scattering of ultrasound by emulsions*, J. Phys. D Appl. Phys., 22 (1989), pp. 38–47.
[9] M.J.W. Povey, *Particle characterisation by ultrasound*, Pharm. Sci. Tech. Today, 3 (2000), pp. 373–380.
[10] E.R. Pike and P.C. Sabatier *Scattering*, Academic Press, New York, 2001.
[11] O.G. Harlen, Y. Qiu, B.D. Sleeman, and M.J.W. Povey, *Acoustic scattering in dispersions*¡ in Analytical and Computational Methods in Scattering and Applied Mathematics, Chapman & Hall/CRC Res. Notes Math. 417, CRC Press, Boca Raton, FL, 2000, pp. 123–134.
[12] M.J.W. Povey, *Ultrasonic Techniques for Fluids Characterization*, Academic Press, San Diego, 1997.
[13] R.M. Lewis and J.B. Keller, *Asymptotic Methods for Partial Differential Equations: The Reduced Wave Equation and Maxwell's Equation*, Research report EM-194, Courant Institute, New York University, New York, 1964.
[14] P. Lloyd and M.V. Berry, *Wave propagation through an assembly of spheres* IV. *Relations between different multiple scattering theories*, Proc. Phys. Soc., 91 (1967), pp. 678–688.
[15] F.V. Atkinson, *On Sommerfeld's radiation condition*, Philos. Mag. (7), 40 (1949), pp. 645–651.
[16] R.B. Barrar and A.F. Kay, *A Series Development of the Wave Equation in Powers of* $1/r$, internal memorandum, Tech. Res. Group, Inc. EM-194, 1964.
[17] C.H. Wilcox, *Spherical means and radiation conditions*, Arch. Ration. Mech. Anal., 3 (1959), pp. 133–148.
[18] P.M. Morse and H. Feshbach, *Methods of Theoretical Physics*, Vols. I, II, McGraw-Hill, New York, 1953.
[19] Y. Hemar, N. Herrmann, P. Lemaréchal, R. Hocquart, and F. Lequeux, *Effective medium model for ultrasonic attenuation due to the thermo-elastic effect in concentrated emulsions*, J. Phys. II France, 7 (1997), pp. 637–647.

SIAM J. APPL. MATH.
Vol. 63, No. 3, pp. 850–873

CONNECTING A DISCRETE IONIC SIMULATION TO A CONTINUUM*

B. NADLER†, T. NAEH†, AND Z. SCHUSS†

Abstract. An important problem in simulating ions in solution is the connection of the finite simulation region to the surrounding continuum bath. In this paper we consider this connection for a simulation of uncharged independent Brownian particles and discuss the relevance of the results to a simulation of charged particles (ions). We consider a simulation region surrounded by a buffer embedded in a continuum bath. We analyze the time course of the exchange process of particles between the simulation region and the continuum, including re-entrances of particles that left the simulation. We partition the particle population into (i) those that have not yet visited the simulation and (ii) those that have. While the arrival process into the simulation of population (i) is Poissonian with known rate, that of population (ii) is more complex. We identify the ordered set of re-entrance times of population (ii) as a superposition of an infinite number of delayed terminating renewal processes, where the renewal periods may be infinite with positive probability. The ordered entrance times of populations (i) and (ii) form the pooled process of injection times of particles into the simulation. We show that while the pooled process is stationary, it is not Poissonian but rather has infinite memory. Yet, under some conditions on the sizes of the simulation and buffer regions, it can be approximated by a Poisson process. This seems to be the first result on the time course of a discrete simulation of a test volume embedded in a continuum.

Key words. stationary stochastic processes, diffusion, renewal theory, simulation of ions

AMS subject classifications. 60G10, 60G35, 60J60, 60K05

PII. S0036139901393688

1. Introduction.

Computer simulations of ions in electrolytic solutions are a widely used tool in physical chemistry and are becoming increasingly important in molecular biophysics as well [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16]. Since it is impractical to simulate the entire continuum bath, a common approach is to isolate a small finite region of the continuum and simulate only the motion of ions located in this region. The requirements from such a "small" simulation are that the averaged concentrations of the different ionic species in the simulation volume be preserved, the electrostatic forces be correctly reproduced, and the effective measured ionic diffusion coefficients be recovered

Of course, as simulated ions may reach the boundary of the simulation region and nonsimulated bath ions may cross it, the simulation must be *connected* to the surrounding continuum bath. This involves not only the correct computation of the electrostatic field, including the contribution of nonsimulated bath ions, but also the resolution of the two following issues: (i) the imposed boundary behavior on trajectories of simulated ions as they reach the boundary of the simulation region and (ii) the injection scheme (if any) of new ions into the simulation. In this paper we are concerned with these two issues. Specifically, we study the processes of random exit,

*Received by the editors August 10, 2001; accepted for publication July 26, 2002; published electronically January 17, 2003.

http://www.siam.org/journals/siap/63-3/39368.html

†Department of Applied Mathematics, School of Mathematical Sciences, Tel-Aviv University, Tel-Aviv 69978, Israel (boaz.nadler@yale.edu, galor@cs.cornell.edu, schuss@post.tau.ac.il). The research of the first author was partially supported by a research grant from DARPA. The research of the second author was partially supported by a grant from the Israel Ministry of Science and Technology. The research of the third author was partially supported by research grants from the US-Israel Binational Science Foundation, from the Israel Science Foundation, and from DARPA.

entrance, and re-entrance of particles between the simulation region and the continuum bath. We analyze a simulation of *uncharged* particles and discuss the relevance of our results to a simulation with *charged* particles in section 8. The computation of the electrostatic field for a simulation of charged particles will be considered elsewhere.

The total number of simulated ions in a simulation scheme can be either fixed or variable. In simulations with a fixed number of ions, there is no injection scheme of new ions into the simulation, and the imposed boundary conditions on the trajectories of simulated particles are either periodic or reflecting [15], [16]. Simulations with a fixed number of ions, and in particular those with periodic boundary conditions, have serious limitations which have been discussed at length in the literature [17], [18], [19], [20], [21], [22], [23] (and references therein). In particular, density fluctuations are absent in such simulations, and the computation of the electrostatic field is at best problematic.

Density fluctuations are determinants of important properties of an ionic solution [12], [13]. There have been various attempts in the literature to include density fluctuations in simulations with a fixed number of ions. The most common method is the introduction of a *buffer* region between the simulation region and the surrounding continuum bath. The simplest approach, as described in [6], is to run a simulation with a fixed total number of particles in the simulation and buffer region, with reflecting boundary conditions at the outer buffer boundary. In this scheme density fluctuations are of course present in the smaller simulation region, although it is unclear how faithfully they reproduce the actual density fluctuations in the simulation region. Other approaches, as reported in [2] and [8], replace the reflecting boundary conditions at the boundary of the buffer region by "soft" boundary conditions. That is, ions are allowed to leave the buffer region into the bath, but then they are subject to an artificial attracting force, so that they eventually return into the buffer region. In both references, the attracting force was designed to maintain the correct equilibrium density in the simulation region. Once again, while the total number of particles is kept fixed, there are fluctuations in the number of particles in the smaller simulation region. The main problem with these approaches is that the confinement of ions to the simulation by ad hoc artificial attracting forces (or even infinite forces, in the case of reflecting boundaries) imposes unphysical conditions on the simulation and may not necessarily lead to correct time dependent density fluctuations.

Simulations with a variable number of ions also use a buffer region between the simulation and the continuum bath, but replace the reflecting or soft boundary conditions at the boundary of the buffer region by stochastic boundary conditions [7], [8], [9], [10]. These conditions introduce a random exchange mechanism of ions between the simulation and buffer regions with the aim of reproducing the equilibrium density fluctuations. Obviously, different assumptions on the stochastic boundaries lead to different density fluctuations in time and space inside the simulation region. Unfortunately, the stochastic process of equilibrium density fluctuations is unknown in the sense that the joint probability distribution of the number of particles in the simulation volume at different times is unknown. The fluctuation theory proposed by Smoluchowski [11] is valid only for sufficiently long time intervals between observations so that it cannot be applied to a simulation of particles in solution [29]. Yet, these fluctuations affect the physical properties of the solute [12], so proposing a scheme that recovers the correct fluctuations is essential.

In all formulations of stochastic boundaries, the probability laws for the injection times of new particles are *assumed*, rather than derived, from the laws of motion of

ions in solution. The aim of this paper is to *derive* the probability laws of the entrance and re-entrance processes of ions into a finite volume surrounded by a buffer zone as they actually occur in the solution. To derive our results, we make standard general assumptions of physical chemistry about the ionic motion of bath ions.

In our analysis, we consider a finite simulation region surrounded by a buffer region embedded in a practically infinite ionic solution. We assume that all bath ions can be described as *independent* uncharged Brownian particles with an effective diffusion coefficient. This assumption is commonly used in physical chemistry, where ionic solutions are described by an electrochemical potential [12]. This means that, on a large enough time scale, the motions of charged interacting ions in the bath are assumed independent *noninteracting* diffusion processes in a mean field, which reduce to independent Brownian particles for a vanishing mean field.

We consider particle entrances at the boundary of the inner region and their exits at the outer boundary of the buffer zone. In the corresponding simulation, the motion of all particles that enter the inner region is simulated until they cross the outer boundary of the buffer region for the first time. Their motion is simulated once again the next time they enter the inner region, until their next exit at the outer region, and so on. The buffer region in the simulation scheme serves as a separator between the inner simulation region and the surrounding continuum bath, thus avoiding instantaneous re-entrances of Brownian particles at the boundary of the inner region,

To formulate mathematically the problem of introducing particles into the simulation, we divide their entrances into two types: (i) arrivals of "new" particles, which have not visited the simulation region so far, and (ii) arrivals of "returning" particles, which have already visited and exited the simulation region. Obviously, the probability law of the recirculation times is *different* from that of the times between new arrivals, so that particles that leave the simulation at the outer boundary of the buffer zone cannot be returned to the bath on equal footing with particles that have not been in the simulation so far.

In our previous paper [24], we studied the stationary arrival process (i) of new particles. It was shown that in steady state, the interarrival times to an absorbing boundary are exponentially distributed with rate equal to the Smoluchowski flux, rendering the stationary arrival process Poissonian. Apart from its relevance to the problem of connecting a simulation to the surrounding continuum, the study of the arrival problem at an absorbing boundary has many physical applications and a long mathematical history [25], [26].

In this paper, we study the recirculation process (ii) and its role in connecting the simulation to the surrounding bath. We determine stationary probability laws governing the entrance and re-entrance times of all processes (i) and (ii). At any given time in the course of the simulation the particle to be injected next is the one whose arrival time at the inner sphere is the *shortest* among all the particles not currently in the simulation. The candidates for injection are both the new and recirculating particles. In this paper we identify the injection process as a *pooled* process, that is, a superposition of an infinite number of terminating renewal processes and determine some of its statistical properties.

We show that the pooled process converges to a stationary steady state. However, even though in the steady state the process is stationary, its interarrival times are not exponential, not even independent, and have infinite memory. We calculate some of the statistical properties of the pooled process, such as the exact pdf of the residual

first arrival time of the pooled process in steady state, as well as the first and second moments of the pooled process. Our main result is that under some conditions on the size of the simulation and buffer regions, the infinite memory pooled process can be approximated by a Poisson process. This approximation considerably simplifies the simulation. To the best of our knowledge, this work seems to be the first result on the time course of a discrete simulation embedded in a continuum.

The paper is organized as follows. In section 2, we formulate the simulation scheme and identify the entrance process of particles as a pooled process. In section 3, we present a continuum model of the simulation, from which the average flux of the pooled process is calculated. The first two moments of the pooled process are calculated in section 4 by renewal-type considerations. In section 5, we define the entrance times of the pooled process, and in section 6, we calculate the distribution of the residual time till the first particle entrance and the distribution of the subsequent interarrival time. The main result, which asserts that short interarrival times are exponentially distributed and the effective exponential rate is the same as that calculated from the continuum and renewal models, is discussed in section 7. We also present there results of a simulation of the pooled process and discuss its rate of convergence to steady state. Section 8 contains a summary and discussion.

2. Setup of the problem. We consider the following simulation scheme: A practically infinite ionic bath of average density ρ occupies the three dimensional space. Inside this bath, there is a finite simulation region consisting of two concentric spheres of radii a and r_0 ($a < r_0$), centered at the origin (see Figure 1(a)). In the proposed simulation scheme, the motion of all particles that enter the inner sphere is simulated until they cross the outer sphere for the first time. Their motion is simulated once again the next time they enter the inner sphere, until their next exit at the outer sphere, and so on (see Figure 1(b)). The region beyond the outer sphere, $|\boldsymbol{r}| > r_0$, contains no simulated particles and is described by a *continuum* particle density. The annular ring $a < |\boldsymbol{r}| < r_0$ is a *buffer* region that connects the inner region to the surrounding continuum bath in $|\boldsymbol{r}| > r_0$. The buffer region is part continuum and part discrete in the sense that the motions of only some of the particles in it are simulated.

While we do not describe the exact electrostatic interactions between bath ions, we follow the common practice in chemical physics [12] that describes the effective motions of the nonsimulated bath ions as independent diffusions in a mean field. Specifically, for a homogeneous bath with no applied potential, the mean field vanishes, so that exterior of the simulation region can be described as an infinite bath of independent free Brownian particles, with average density ρ. As discussed in the introduction, we assume that the simulation is *self-consistent*. This means that on a large enough time scale the coarse grained motion of simulated ions inside the simulation region can also be described as free diffusion with the same diffusion coefficient as that assumed for the nonsimulated ions in the continuum bath. We further assume that the simulation and buffer regions are large enough so that, for the purpose of calculating the time course of the simulation, all particles, both simulated and nonsimulated, can be described as effectively independent Brownian particles with the above diffusion coefficient. We note that the self-consistency condition is not trivial, and it determines important physical parameters, as discussed in [27].

We introduce the following notation. Particles that have not visited the inner sphere so far are called *blue* particles, and those that have are called *red* particles. The arrival process of blue particles into the simulation is process (i) and the re-

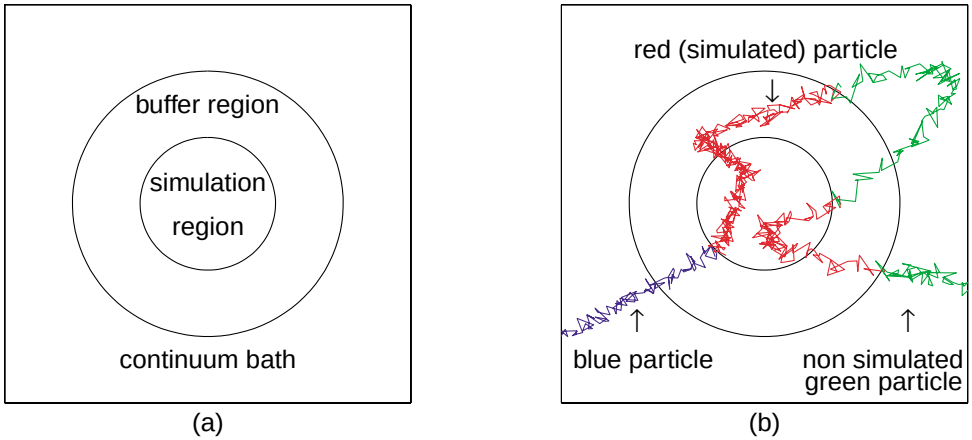


FIG. 1. (a) *The simulation setup.* (b) *The simulated and nonsimulated parts of a typical particle trajectory.*

entrance of the red particles is process (ii).

As shown in [24], in steady state, blue Brownian particles arrive at an absorbing sphere at an exponential rate. A blue particle that reaches the boundary of the inner sphere turns red instantaneously and stays red forever. As long as it is inside the simulation region, its dynamics change from independent Brownian motion, with its effective diffusion coefficient in the ambient solution, to diffusive motion governed by the Langevin equation with electrostatic interactions with the ions in the simulation and with the far field of the ambient solution. The interactions with the nonsimulated ions in the ambient solution outside the outer sphere are replaced by interactions with a mean field, as mentioned above. The assumption of a self-consistent simulation implies that for our purposes the probability distribution of the time a simulated ion spends inside the simulation is identical to that of a free noninteracting Brownian particle.

One way to run this simulation is as follows: Introduce blue particles at exponential interarrival times. Follow the trajectory of each (now red) simulated particle until its first exit at the boundary of the larger sphere. Now, sample a random re-entrance time into the simulation, assuming it performs Brownian motion outside the inner sphere, and store this re-entrance time in a table of all re-entrance times of recirculating particles. In this scheme, the next particle to be injected into the simulation is the one with the *minimal* return time between all particles registered in the table and the next blue particle to be injected into the simulation. There are two main difficulties with this simulation scheme. One is that the table of re-entrance times grows indefinitely with time, because the mean recirculation time of returning particles is infinite (see Proposition A.1 in Appendix A). The other difficulty is that the convergence to steady state of this simulation is extremely slow, as analyzed in section 7. This is due to the fact that as long as the table is finite, all the infinite number of re-entrances of particles that were inside the simulation region *before* the simulation actually started, and are thus not present in this table, are neglected.

This simple example shows that a mechanism to run the simulation in steady state from its start needs to be developed. More specifically, the steady state distribution of return times from this infinite table of recirculated particles has to be calculated.

A key point in this calculation is the well-known feature [28] that for free Brownian motion in three dimensions there is a positive probability that a red particle that just exited the simulation will never return to the inner sphere, so that its recirculation time is infinite. This observation gives rise to the following renewal-type model. The arrivals of blue particles at the inner sphere form a Poisson process [24], [29], as mentioned above. For each arriving particle, its subsequent re-entrance times into the simulation form an independent renewal process. The interarrival times of this process may be infinite with positive probability, thus rendering it a *terminating renewal process* [30]. The renewal processes of different particles start of course at different times, according to their first injection times. A renewal process that starts at a random time with one distribution and is renewed with another is called a *delayed renewal process* [30]. The superposition of all the delayed renewal processes is called the *pooled process*. The steady state of the pooled process is the process of introducing new particles into the simulation, which is the concern of this paper.

3. A continuum model of the simulation. In this section we compute the average *flux* of particle entrances of the pooled process (both blue and red) into the inner simulation region from a continuum model of the above described simulation. To this end, we represent the Brownian particles in the simulation and in the bath as continuum *densities*. Since we consider free Brownian particles, all densities are spherically symmetric and depend only on the radial distance, $r = |\boldsymbol{r}|$, from the center of the simulation spheres. The densities of simulated particles are defined as averages of many different realizations of particle locations of a running simulation.

We start from the radial density of the blue particles, denoted $p_B(r)$. It satisfies the diffusion equation outside the inner sphere [31],

$$\Delta p_B(r) = \frac{d^2 p_B(r)}{dr^2} + \frac{2}{r}\frac{dp_B(r)}{dr} = 0 \qquad \text{for } r > a, \tag{3.1}$$

with absorbing boundary conditions at the boundary of the inner sphere, where blue particles instantaneously turn red,

$$p_B(a) = 0. \tag{3.2}$$

In addition, far away from the simulation region the blue particle density equals the bulk density ρ,

$$\lim_{r\to\infty} p_B(r) = \rho. \tag{3.3}$$

The solution of (3.1)–(3.3) is

$$p_B(r) = \begin{cases} \rho\left(1 - \dfrac{a}{r}\right) & \text{for} \quad r > a, \\ 0 & \text{for} \quad r < a. \end{cases} \tag{3.4}$$

Using (3.4), the continuum flux of blue particles at the inner sphere is given by

$$J_{\text{blue}} = -4\pi a^2 D \frac{d}{dr} p_B(r)\bigg|_{r=a} = 4\pi\rho a D, \tag{3.5}$$

where ρ is the bulk concentration at infinity, and D is the diffusion coefficient of bath particles. Equation (3.5) for the average flux of Brownian particles at an absorbing boundary was already calculated by Smoluchowski in 1917 [25].

Next, we consider the red particle density. According to our assumptions, the total particle density in the bath is uniform and at all locations equals the bulk density ρ. Since the simulation region is an arbitrary region of the bath, the steady state density of the red particles, denoted $p_R(r)$, complements that of the blue particles to the bulk density ρ,

$$p_R(r) = \rho - p_B(r). \tag{3.6}$$

We denote by J_{total} the total flux of particle entrances at the inner sphere. The total flux is the sum of the blue particles flux given by (3.5), and the flux of *returning* red particles, which have exited the simulation through the outer sphere. Since at any given time, only some of the red particles are simulated while others are not, the flux of returning red particles cannot be computed from (3.6). The contribution of red particles to the total incoming flux at the inner sphere comes only from the *nonsimulated* population of red particles. To compute their contribution to the influx, we divide the red particle population into two: simulated red particles, denoted *pink* particles, and nonsimulated red particles, denoted *green* particles. With this notation, the total flux at the inner sphere is given by

$$J_{\text{total}} = J_{\text{blue}} + J_{\text{green}}. \tag{3.7}$$

We denote the steady state densities of the pink and green particles by $p_P(r)$ and $p_G(r)$, respectively. Of course,

$$p_R(r) = p_P(r) + p_G(r).$$

We now calculate the densities of the green and pink particles. Each particle that enters the simulation region at the inner sphere, either new or returning particle, exits it at the outer sphere at some later time with probability one. Once such a particle crosses the outer sphere it immediately becomes green, until its next arrival at the inner sphere, when it becomes pink again. Thus, the green particle density has a *source* at the outer sphere whose strength equals the yet undetermined total absorption flux J_{total} of both blue and returning red particles. That is, the green particle density satisfies the diffusion equation

$$\Delta p_G(r) = \frac{d^2 p_G(r)}{dr^2} + \frac{2}{r}\frac{dp_G(r)}{dr} = J_{\text{total}}\frac{\delta(r-r_0)}{4\pi r_0^2}, \qquad \text{for} \quad r > a, \tag{3.8}$$

with absorbing boundary conditions at the inner sphere,

$$p_G(a) = 0, \tag{3.9}$$

and with the condition that $p_G(r) \to 0$ as $r \to \infty$. The solution for the green particle density, in terms of the yet undetermined parameter J_{total}, is given by

$$p_G(r) = \begin{cases} \dfrac{J_{\text{total}}}{4\pi r_0}\left[\left(1-\dfrac{a}{r}\right) - H(r-r_0)\left(1-\dfrac{r_0}{r}\right)\right], & \text{for} \quad r > a, \\ 0, & \text{for} \quad r < a, \end{cases}$$

where $H(x)$ denotes the Heaviside step function. The averaged densities of the blue, pink, and green particle populations are shown in Figure 2.

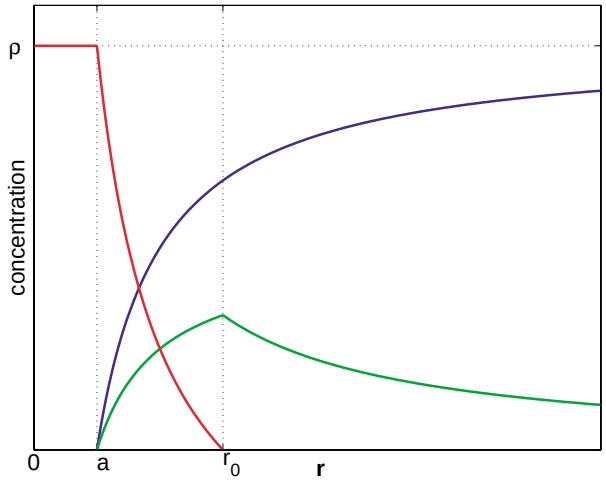


FIG. 2. *The averaged densities of the blue, pink, and green particle population.*

The average influx of green particles at the inner sphere, denoted J_{green}, is thus

$$J_{\text{green}} = 4\pi r^2 \frac{d}{dr} p_G(r)\Big|_{r=a} = \frac{a}{r_0} J_{\text{total}}. \tag{3.10}$$

Note that the flux of green particles into the inner sphere is *smaller* than its source strength J_{total}. This is because there is a positive probability for a green particle that starts its motion at the boundary of the outer sphere to never reach the inner sphere. The ratio between the two fluxes is simply the probability of a free Brownian particle to ever reach the inner sphere from the outer sphere. As shown in section 6, this return probability, denoted p, is given by

$$p = \frac{a}{r_0}. \tag{3.11}$$

The total mean flux of particles into the simulation, denoted Λ, can now be obtained by combining (3.5), (3.7), and (3.10),

$$\Lambda = J_{\text{total}} = J_{\text{blue}} + J_{\text{green}} = \frac{\lambda_B}{1-p}. \tag{3.12}$$

As expected, due to the recirculating red particles, the total flux at the inner sphere is larger than the flux of only the blue particles. For example, in a simulation scheme that inserts only blue particles, absorbs them at the outer sphere and "forgets" about their possible re-entrances, the average particle flux into the simulation region is smaller than it should be. This might have serious effects on the outcome of the simulation.

Finally, note that the flux parameter Λ does not represent a physical quantity, but is rather only a simulation parameter, that depends on the choice of the radii a and r_0 of the simulation spheres. Therefore, all physical parameters that are an outcome of the simulation must not depend on Λ.

4. The mean and variance of the pooled process. In the previous section the total average influx of the steady state pooled process was computed with the

aid of a corresponding continuum model. However, the continuum model is unable to compute other statistical properties of the pooled process, such as the distribution of the interarrival times, or even their variance. Obviously, a simulation scheme should attempt to preserve at least some of these quantities. In this section, we present a statistical renewal model of the pooled process that, in principle, enables the computation of all moments of the pooled process.

For simplicity, we consider a simulation that starts at time $t = 0$ with no particles initially inside the simulation region. Since we are interested in quantities for a simulation that has reached steady state, the results are independent of these initial conditions. We denote by $N^P(t)$ the total number of particle entrances of the pooled process by time t. This includes all entrances of blue particles and re-entrances of returning red particles into the simulation region by time t. We compute the first two moments of $N^P(t)$ and note that the method presented can be applied to compute all higher order moments as well.

THEOREM 4.1. *The average steady state flux of the pooled process is* Λ*, and its variance per unit time is* $\Lambda(1+p)/(1-p)$.

Proof. Let $N^B(t)$ denote the number of arrivals of *blue* particles into the simulation during the time interval $[0,t]$, and let $\{t^i\}_{i=1}^{N^B(t)}$ denote these arrival times. According to [24], the interarrival times of the blue particles are independent identically distributed (i.i.d.) random variables, exponentially distributed with rate λ_B that is equal to the corresponding continuum flux J_{blue} calculated in (3.5). Therefore, the total number of blue arrivals by time t, denoted $N^B(t)$, is a Poisson distributed random variable with parameter $\lambda_B\, t$. For each blue particle we denote by $\xi(t-t^i)$ the (random) number of its re-entrances into the simulation by time t since its first entrance at time t^i. In terms of these random variables, the total number of particle entrances of the pooled process can be written as

$$N^P(t) = \sum_{i=1}^{N^B(t)} \left[1 + \xi(t-t^i)\right]. \tag{4.1}$$

We denote by $\mu_1(t)$ and $\mu_2(t)$ the first two moments of $\xi(t)$,

$$\mu_1(t) = E\left[\xi(t)\right], \qquad \mu_2(t) = E\left[\xi(t)\right]^2. \tag{4.2}$$

Note that $\xi(\infty)$ is the total number of re-entrances of a red particle. Since each particle has probability $p < 1$ of ever returning to the simulation (see (3.11)), the random variable $\xi(\infty)$ follows a *geometric* distribution with parameter p, which gives

$$\mu_1(\infty) = \frac{p}{1-p}, \qquad \mu_2(\infty) = \frac{p}{1-p} + \frac{2p^2}{(1-p)^2}. \tag{4.3}$$

To compute the average of the pooled process, $E[N^P(t)]$, we divide the sum in (4.1) into the first term and the sum of the $N^B(t)-1$ remaining terms,

$$N^P(t) = 1 + \xi(t-t^1) + \sum_{i=2}^{N^B(t)} (1 + \xi(t-t^i)). \tag{4.4}$$

According to our assumptions, the arrival process of blue particles is Poissonian and thus memoryless. In addition, the recirculation processes of different particles are independent and identical, if their starting times are all shifted to an identical initial

time. These two properties imply that given the arrival time of the first blue particle, $t^1 = s$, the sum in (4.4) has the same statistical properties as that of the random variable $N^P(t-s)$. In particular,

$$\Pr\left\{ \sum_{i=2}^{N^B(t)} (1+\xi(t-t^i)) = n \middle| t^1 = s \right\} = \Pr\left\{ \sum_{i=2}^{N^B(t)} [1+\xi\left((t-s)-(t^i-s)\right)] = n \right\}$$

$$= \Pr\left\{ N^P(t-s) = n \right\}. \tag{4.5}$$

Thus, taking expectations in (4.4) and using the exponential distribution of t^1 gives

$$E\left[N^P(t)\right] = \int_0^t E\left[N^P(t)\middle| t^1 = s\right] \lambda_B e^{-\lambda_B s}\, ds$$

$$= \int_0^t \left[1 + \mu_1(t-s) + E \sum_{i=2}^{N^B(t)} [1+\xi((t-s)-(t^i-s))] \right] \lambda_B e^{-\lambda_B s}\, ds,$$

which, according to (4.5), can equivalently be written as a renewal-type integral equation,

$$E\left[N^P(t)\right] = \int_0^t \lambda_B e^{-\lambda_B s} \Big\{ 1 + \mu_1(t-s) + E\Big[N^P(t-s)\Big] \Big\}\, ds, \tag{4.6}$$

along with the initial condition $N^P(0) = 0$. The solution of (4.6) is given by

$$E[N^P(t)] = \int_0^t \lambda_B \left[1 + \mu_1(s)\right]\, ds. \tag{4.7}$$

Therefore, by l′Hôpital′s rule and using (4.3),

$$\lim_{t\to\infty} \frac{E\left[N^P(t)\right]}{t} = \lim_{t\to\infty} \lambda_B \Big[1 + \mu_1(t)\Big] = \lambda_B \Big[1 + \mu_1(\infty)\Big] = \Lambda.$$

As expected, we recover the same total average flux of the pooled process as that computed from the continuum model, (3.12).

Next, we consider the second moment, $E[N^P(t)^2]$. Before computing the expectation, we write $[N^P(t)]^2$ as

$$[N^P(t)]^2 = \sum_{i=1}^{N^B(t)} [1 + 2\xi(t-t^i) + \xi^2(t-t^i)] + 2 \sum_{i=1}^{N^B(t)} [1+\xi(t-t^i)] \sum_{j=i+1}^{N^B(t)} [1+\xi(t-t^j)]$$

$$= H(t) + 2G(t),$$

where

$$H(t) = \sum_{i=1}^{N^B(t)} [1 + 2\xi(t-t^i) + \xi^2(t-t^i)],$$

$$G(t) = \sum_{i=1}^{N^B(t)} [1+\xi(t-t^i)] \sum_{j=i+1}^{N^B(t)} [1+\xi(t-t^j)].$$

The expectations of $H(t)$ and $G(t)$ also satisfy integral equations,

$$E\left[H(t)\right] = \int_0^t \lambda_B e^{-\lambda_B s}\Big\{1 + 2\mu_1(t-s) + \mu_2(t-s) + E\left[H(t-s)\right]\Big\}\, ds,$$
$$E\left[G(t)\right] = \int_0^t \lambda_B e^{-\lambda_B s}\Big\{\big(1 + \mu_1(t-s)\big)E\big[N^P(t-s)\big] + E\left[G(t-s)\right]\Big\}\, ds.$$

The solutions of these equations are

$$E\Big[H(t)\Big] = \int_0^t \lambda_B\Big[1 + 2\mu_1(s) + \mu_2(s)\Big]\, ds,$$
$$E\Big[G(t)\Big] = \int_0^t \lambda_B\Big[1 + \mu_1(s)\Big]E\left[N^P(s)\right]\, ds.$$

The long time behavior of the variance of $N^P(t)$ is found from the identity

$$\lim_{t\to\infty}\frac{\mathrm{Var}[N^P(t)]}{t} = \lim_{t\to\infty}\frac{E\left\{[N^P(t)]^2\right\} - \left\{EN^P(t)\right\}^2}{t}$$
$$= \lim_{t\to\infty}\frac{E[H(t)] + 2E[G(t)] - \left\{E[N^P(t)]\right\}^2}{t}.$$

Inserting the expressions for all quantities, and applying l′Hôpital′s rule, gives

$$\lim_{t\to\infty}\frac{\mathrm{Var}[N^P(t)]}{t} = \Lambda\left[1 + 2\left(\frac{p}{1-p}\right)\right] = \Lambda\,\frac{1+p}{1-p}. \tag{4.8}$$

Note that all moments of the pooled process are independent of the exact distribution of the recirculation time. Rather, they depend only on the return probability p.

Equation (4.8) clearly shows that the pooled process $N^P(t)$ is *not* Poissonian, since the variance per unit time of a Poisson process equals its average rate Λ. The variance of $N^P(t)$ is larger by a factor $(1+p)/(1-p)$, a phenomenon due to the possible re-entrances of exiting particles. In approximating the pooled process by a Poisson process with the same rate, this factor is lost.

5. The entrance times of the pooled process. In the previous section we calculated the first two moments of the pooled process. We now study the actual distribution of the interarrival times of the pooled process, that is, the PDF of the time between the consecutive introductions of particles into the simulation. As discussed above, these entrance times are the ordered union of both entrance times of new blue particles and re-entrance times of red recirculating particles.

Recall that the consecutive arrival times of blue particles were denoted t^j. We denote their interarrival times by $\tau^j = t^j - t^{j-1}$ with the convention that $t^0 = 0$. As shown in [24], the arrival process of blue particles is Poissonian with rate λ_B given by (3.5). Therefore, τ^j are i.i.d. exponential random variables with a common PDF

$$\Pr\{\tau^j \le t\} = F_B(t) = 1 - \exp(-\lambda_B t). \tag{5.1}$$

Next, we consider the recirculation process of red particles. We introduce the following notation for the successive exit and re-entrance times of a red particle. The first entrance time to the inner sphere is denoted by $t_1^j \equiv t^j$. Its first exit time from

the simulation at the outer sphere is denoted θ_1^j, its next re-entrance time t_2^j, and so on. Thus $t_1^j < \theta_1^j < t_2^j < \theta_2^j < \dots$.

The times t_n^j are the consecutive *re-entrance times* of the jth particle into the simulation. We denote $T_1^j = t_1^j$ and set $T_n^j = t_n^j - t_{n-1}^j$ for $n > 1$. The times T_n^j for $n > 1$ are called the *recirculation times* of the jth particle. According to our assumptions, these times are i.i.d. random variables with a positive probability to be infinite. We denote their PDF by

$$F_T(t) = \Pr\{T_n^j \le t\}.$$

Their pdf is given by

$$f_T(t) = f_{\tau_{in}} * f_{\tau_{out}}(t),$$

where τ_{in} is the time a simulated (pink) particle spends in the simulation and τ_{out} is the time a nonsimulated (green) particle spends outside the simulation.

The assumption that the green particles have a positive probability $1-p$ of never returning from the outer sphere to the inner sphere is expressed as

$$\lim_{t\to\infty} F_T(t) = p < 1, \tag{5.2}$$

or, equivalently,

$$\Pr\left\{T_n^j = \infty\right\} = 1 - p > 0 \qquad (n > 1), \tag{5.3}$$

where p is given by (3.11).

Recall that $\xi(t - t^j)$ denoted the number of re-entrances of the jth red particle into the simulation by time t. We denote by $N^j(t)$ its total number of entrances into the simulation by time t, including the first entrance at time t^j,

$$N^j(t) = 1 + \xi(t - t^j).$$

Note that (5.3) implies that $N^j(t)$ is a *terminating renewal process* [30], that is, a renewal process that terminates when an infinite recirculation time occurs.

We now consider the entrance times of the pooled process. By definition, the total number of particle entrances into the simulation by time t, denoted $N^P(t)$, is given by

$$N^P(t) = \sum_{j=1}^{N^B(t)} N^j(t). \tag{5.4}$$

The actual entrance times of the pooled process into the simulation, $\{S_\ell^P\}_{\ell=1}^{N^P(t)}$, are the elements of the set

$$\{S_\ell^P\}_{\ell=1}^{N^P(t)} = \{t_k^j \,|\, 1 \le j \le N^B(t),\, 1 \le k \le N^j(t)\}$$

arranged in *ascending* order.

The times between successive arrivals of the pooled process at the inner sphere, denoted T_ℓ^P, are defined by

$$T_\ell^P = S_{\ell+1}^P - S_\ell^P.$$

With this notation the mathematical problem of simulating the arrivals of ions at the inner sphere is to determine the joint PDF of the times T_ℓ^P (for all ℓ). These times are the interarrival times for introducing new particles into the simulation.

6. The first arrival time of the pooled process. In this section, we calculate the exact PDF of the first arrival time of the pooled process in the steady state. First, we compute the distribution of the *residual* time since start of observation of the simulation till the introduction of the first new particle into the inner sphere. Then, we compute the PDF of the time between consecutive arrivals of the pooled process.

We introduce the following notation. We denote by $\varphi(t)$ the renewal function of the recirculation process of a single particle. It is given by

$$\varphi(t) = \sum_{k=0}^{\infty} f_T^{*k}(t), \tag{6.1}$$

where $f_T^{*k}(t)$ is the k-convolution of the pdf $f_T(t)$ of a single recirculation time, and $f_T^{*0}(t) = \delta(t)$. For future uses, we note that the Laplace transform of φ is

$$\widehat{\varphi}(s) = \sum_{k=0}^{\infty} \widehat{f_T^{*k}}(s) = \sum_{k=0}^{\infty} \left[\widehat{f}_T(s)\right]^k = \frac{1}{1-\widehat{f}_T(s)}, \tag{6.2}$$

and at $s = 0$ we obtain from (5.2)

$$\widehat{\varphi}(0) = \int_0^{\infty} \varphi(t)\, dt = \frac{1}{1-p}. \tag{6.3}$$

Consider a simulation that has been running for an infinite time and is already in steady state. We start to observe the simulation at time $t = 0$, and denote by γ^P the first arrival time of a particle into the simulation after $t = 0$. The first particle to arrive into the simulation may be either a blue particle that has not yet been in the simulation, or a red particle that has visited the simulation in the past and may re-enter the simulation region at the inner sphere after start of observation. Before we compute the exact PDF of γ^P, it is useful to compute the probability that a red particle that initially entered the simulation region at time $-s$ in the past will re-enter the simulation region at time x after start of observation.

LEMMA 6.1. *Let γ_s^R denote the first re-entrance time after $t = 0$ of a red particle that initially entered the simulation at time $-s$. Then*

$$\Pr\left\{\gamma_s^R = x\right\} = \int_0^s \varphi(s-u) f_T(x+u)\, du. \tag{6.4}$$

Proof. Consider a blue particle that entered the simulation at time $t_1 = -s$. In the time interval $[-s, 0]$ this particle may have re-entered the simulation an arbitrary number of times. For a particle that recursed $k-1$ times before time $t = 0$, we denote by t_k its last re-entrance time before $t = 0$ and by t_{k+1} its next re-entrance time after $t = 0$. The event $\left\{\gamma_s^R = x\right\}$ can thus be decomposed into the disjoint union

$$\{\gamma_s^R = x\} = \bigcup_{k=1}^{\infty} \{t_k < 0 \cap t_{k+1} = x \,|\, t_1 = -s\}$$

so that

$$\Pr\left\{\gamma_s^R = x\right\} = \sum_{k=1}^{\infty} \Pr\{t_k < 0 \cap t_{k+1} = x \,|\, t_1 = -s\}. \tag{6.5}$$

The first summand is the probability that the first re-entrance time of the particle occurred at time x. Therefore,

$$\Pr\{t_1 < 0 \cap t_2 = x \,|\, t_1 = -s\} = f_T(s+x). \tag{6.6}$$

The next summand ($k = 2$) is the probability of exactly one recirculation before time $t = 0$ and next re-entrance at time x. To compute this probability, we integrate over all possible times $-u$ for the recirculation time t_2,

$$\begin{aligned}\Pr\{t_2 < 0 \cap t_3 = x \,|\, t_1 = -s\} &= \int_0^s \Pr\{t_2 = -u \cap t_3 = x \,|\, t_1 = -s\}\, du \\ &= \int_0^s f_T(s-u) f_T(u+x)\, du. \end{aligned} \tag{6.7}$$

We now consider the kth term in the sum (6.5). It represents the probability of exactly $k-1$ recirculations before time $t = 0$ and next re-entrance at time x. Let $-u$ denote the last recirculation time prior to time $t = 0$. By assumption, all recirculation times of a particle are i.i.d. random variables with pdf $f_T(t)$. Therefore,

$$\Pr\{t_k = -u \,|\, t_1 = -s\} = f_T^{*(k-1)}(s-u),$$

where f_T^{*k} denotes the kth convolution of the pdf $f_T(t)$. Thus,

$$\Pr\{t_k < 0 \cap t_{k+1} = x \,|\, t_1 = -s\} = \int_0^s f_T^{*(k-1)}(s-u) f_T(u+x)\, du. \tag{6.8}$$

Combining (6.5) with (6.6) and (6.8) and using the definition $f_T^{*0}(t) = \delta(t)$, we obtain that

$$\begin{aligned}\Pr\{\gamma_s^R = x\} &= \int_0^s \sum_{k=0}^{\infty} f_T^{*k}(s-u) f_T(u+x)\, du \\ &= \int_0^s \varphi(s-u) f_T(u+x)\, du, \end{aligned}$$

which concludes the proof of the lemma.

We are now ready to prove the following theorem concerning the PDF of the first arrival time of the pooled process.

THEOREM 6.2. *The stationary PDF of the first arrival time of a steady state pooled process is given by*

$$\Pr\left\{\gamma^P > x\right\} = \exp\{-\Lambda x\} \exp\left\{\Lambda \int_0^x F_T(t)\, dt\right\}. \tag{6.9}$$

Proof. Consider a simulation that has been running for an infinite time which we start to observe at time $t = 0$. The event $\{x < \gamma^P < x + \Delta x\}$ means that no particles arrived into the simulation in the time interval $[0, x]$ and exactly one particle arrived in the short time interval $[x, x + \Delta x]$. This means, of course, that all the remaining particles arrived after time x. The identity of the arriving particle can be either blue or red. Therefore,

$$\begin{aligned}\Pr\{\gamma^P = x\} &= \Pr\left\{\gamma^P > x\right\} \\ &\quad \times [\Pr\{\text{blue arrival at } x\} + \Pr\{\text{red re-entrance at } x\}]. \end{aligned} \tag{6.10}$$

Since the arrivals of blue particles is a memoryless Poisson process with rate λ_B, the probability of the first blue particle to arrive during the time interval $[x, x+\Delta x]$ is approximately $\lambda_B \Delta x$. To compute the re-entrance probability of a red particle at time x, we write

$$\begin{aligned}&\Pr\{\text{red re-entrance at time } x\}\\ &= \int_0^\infty \Pr\{\text{red re-entrance at time } x \,|\, \text{the first entrance time of a red particle}{=}-s\}\\ &\times \Pr\{\text{a red particle first entered the simulation at time } -s\}\, ds.\end{aligned}$$

We consider each term factor in the integral separately. First, we recall that the conditional re-entrance time of this red particle was denoted γ_s^R and its pdf was calculated in the previous lemma. Therefore we write

$$\begin{aligned}&\Pr\{\text{red re-entrance at time } x \,|\, \text{the first entrance time of a red particle}{=}-s\}\\ &= \Pr\{\gamma_s^R = x \,|\, t_1 = -s\}.\end{aligned}$$

Second, by definition,

$$\begin{aligned}&\Pr\{\text{a red particle first entered the simulation at time } -s\}\, ds\\ &= \Pr\{\text{a blue particle entered the simulation at time } -s\}\, ds = \lambda_B\, ds.\end{aligned}$$

Thus, (6.10) can be rewritten as

$$\Pr\left\{\gamma^P = x\right\} = \Pr\left\{\gamma^P > x\right\} \times \lambda_B \left[1 + \int_0^\infty \Pr\{\gamma_s^R = x \,|\, t_1 = -s\}\, ds\right]. \tag{6.11}$$

Inserting (6.4) into (6.11), changing the order of integration in the resulting double integral, and using (6.3) and (5.2) gives

$$\begin{aligned}\Pr\{\gamma^P = x\} &= \lambda_B \Pr\{\gamma^P > x\} \left[1 + \int_0^\infty f_T(x+u)du \int_u^\infty \varphi(s-u)\, ds\right]\\ &= \lambda_B \Pr\{\gamma^P > x\} \left[1 + \frac{F_T(\infty) - F_T(x)}{1-p}\right]\\ &= \lambda_B \Pr\{\gamma^P > x\} \frac{1 - F_T(x)}{1-p}. \qquad (6.12)\end{aligned}$$

Finally, integrating (6.12) with respect to x, we obtain (6.9).

Comment. The fact that the limiting PDF (6.9) is not exponential is yet another manifestation of the non-Poissonian character of the pooled process. The fact that the pooled process is not Poissonian sets this result apart from the known cases of finite mean recurrence times, as analyzed in [33], where the resulting process is Poissonian.

We now show, as mentioned in the introduction, that the pooled process is not a renewal process and has an infinite memory. Therefore its interarrival times are dependent, not identically distributed random variables. Thus, for example, the PDF of the time between the first and the second arrivals after observation begins is not

the same as that of the time between the second and the third arrivals, and so on. The PDF of the interarrival time between the kth and $k+1$th arrivals after observation begins can be calculated as the marginal distribution of the joint distribution of the $k+1$ consecutive interarrival times.

First, we compute the PDF of the time between the first and the second arrivals of the pooled process. Applying considerations similar to the ones in the above computation of the residual first entrance time, it can be shown that

$$\Pr\{\gamma^P = x \cap T_1^P > t\} = \Lambda F_T^c(x) F_T^c(t) \Pr\{\gamma^P > t + x\}.$$

Thus, integrating with respect to x, we obtain

$$\Pr\{T_1^P > t\} = \int_0^\infty \Lambda F_T^c(x) F_T^c(t) \Pr\{\gamma^P > t + x\}\, dx. \tag{6.13}$$

Also, the conditional probability is given by

$$\Pr\{T_1^P > t \,|\, \gamma^P = x\} = \frac{\Lambda F_T^c(x) F_T^c(t) \Pr\{\gamma^P > t + x\}}{\Lambda F_T^c(x) \Pr\{\gamma^P > x\}}. \tag{6.14}$$

A comparison of (6.13) and (6.14) shows that the conditional PDF of T_1^P, given the value of the residual first entrance time γ^P, is *different* than the unconditional PDF of T_1^P. Therefore, the pooled process is not Poissonian, not even a renewal process, but rather has memory. Using similar methods, it is possible to show that the PDF of the kth interarrival time of the pooled process depends on all previous arrival times, which means that the pooled process has infinite memory.

7. Simulation of the pooled process. In this section we present a preliminary statistical analysis of the pooled process and some computer simulation results. First, we estimate the time for convergence of a simulation of the pooled process to steady state. Then, we show both mathematically and numerically that the interarrival times of the pooled process are approximately exponentially distributed. We stress that we study only the pdf of a single interarrival time and not the joint pdf of two or more consecutive interarrival times, nor do we study the time correlations of the pooled process. These issues will be studied in a separate publication.

Consider a computer simulation of the pooled process. As analyzed in section 2, the pooled process is the superposition of many delayed terminating renewal processes. The delayed process is the arrival process of blue particles which is Poissonian with rate λ_B. Thus, the arrival times of the blue particles are easily constructed by sampling their interarrival times from an exponential distribution with rate λ_B. For each blue particle we construct its re-entrance times into the simulation by sampling from the defective distribution F_T. Note that for each blue particle, the total number of its re-entrances follows a geometric distribution with parameter p. Thus, for each blue particle, this sampling procedure results in a *finite* sequence of random re-entrance times. The arrival times of the blue particles are formed by sorting all these entrance and re-entrance times of all particles in increasing order.

7.1. Convergence to steady state. We can estimate the rate of convergence of the pooled process to steady state, starting with no simulated particles inside the inner sphere. Specifically, we determine the minimal time t_S at which

$$\frac{E[N^P(t_S)]}{t_S} = \Lambda(1 - \varepsilon). \tag{7.1}$$

This is a criterion for wide sense convergence to steady state [32]. An explicit expression for $E[N^P(t)]$ was calculated in (4.7) in terms of $\mu_1(s)$, the average number of re-entrances by time s. Since $\mu_1(s)$ is a monotone increasing function, it can be easily seen that $E[N^P(t)]/t$ is also a monotone increasing function. To find a lower bound for t_S, we use the fact that $\mu_1(s)$ satisfies the integral equation [30]

$$\mu_1(s) = \int_0^s f_T(u)[1 + \mu_1(s-u)]du. \tag{7.2}$$

According to (4.3), for all times s,

$$\mu_1(s) < \mu_1(\infty) = \frac{p}{1-p}.$$

Inserting this inequality into (7.2), we obtain a more refined inequality for $\mu_1(s)$,

$$\mu_1(s) < \frac{1}{1-p}F_T(s) < \frac{1}{1-p}\Pr\Big\{\tau_{out} < t\Big\}. \tag{7.3}$$

Combining (A.4) and the inequality

$$\frac{2}{\sqrt{\pi}}\int_x^\infty e^{-u^2}du \le 1 - \frac{2}{\sqrt{\pi}}x \qquad \text{for } x < \frac{1}{\sqrt{2}},$$

we obtain that

$$\Pr\{\tau_{out} < t\} \le \begin{cases} p\left(1 - \dfrac{2}{\sqrt{\pi}}\sqrt{\dfrac{T_b}{t}}\right), & t > 2T_b, \\ \\ p, & t < 2T_b, \end{cases} \tag{7.4}$$

where T_b is defined in (A.5). Inserting (7.4) and (7.3) into (4.7) gives

$$\frac{E[N^P(t)]}{t} \le \Lambda\left(1 - p\frac{4}{\sqrt{\pi}}\sqrt{\frac{T_b}{t}} + p\frac{4\sqrt{2}}{\sqrt{\pi}}\frac{T_b}{t}\right). \tag{7.5}$$

We are now ready to apply the wide sense criteria (7.1). For $\varepsilon \ll 1$, $t_S \gg T_b$, so we can neglect the last term. This gives

$$t_S \ge \frac{16p^2}{\pi}\frac{T_b}{\varepsilon^2}. \tag{7.6}$$

To evaluate whether this time is long or short, consider just the average number of blue particle entrances during this time (neglecting their re-entrances). This number is given by the product $\lambda_B t_S$. Using (3.5) and (7.6) for λ_B and t_S, respectively, gives

$$E[N^B(t_S)] \ge \frac{12(1-p)^2}{\pi}\frac{N_a}{\varepsilon^2}, \tag{7.7}$$

where $N_a = 4/3\pi a^3\rho \gg 1$ denotes the average number of simulated particles in the inner sphere. Therefore, to obtain wide sense convergence up to one percent for a simulation with an average of $N_a = 400$ particles in the inner sphere, the total number of blue particle entrances during this time is of the order of at least $N_a/\varepsilon^2 = 4,000,000$ particles. Note that in a realistic simulation, the time steps of ionic motion are much smaller than the times between consecutive entrances of particles into the simulation. Therefore, (7.7) implies that a simulation that has not started in steady state must be run a prohibitively large number of time steps until convergence to steady state is achieved. Therefore, as discussed in section 2, an algorithm to run the simulation in steady state from the beginning is needed.

7.2. Short interarrival times are approximately exponential. Inserting (6.9) into (6.13) gives the following expression for the distribution of T_1^P:

$$\Pr\{T_1^P > t\} = F_T^c(t) \int_0^\infty \Lambda F_T^c(x) e^{-\Lambda(x+t)} e^{\Lambda \int_0^{t+x} F_T(s) ds} dx. \tag{7.8}$$

We now analyze the short time and long time behavior of this distribution. First we consider short times, $t \ll T_b$. In this case we make a change of variables $x = T_b u$ and $s = T_b w$ in the integrals in (7.8). This gives

$$\Pr\{T_1^P > t\} = F_T^c(t) e^{-\Lambda t} \int_0^\infty \Lambda T_b F_T^c(T_b u) e^{-\Lambda T_b (u - \int_0^{t/T_b + u} F_T(T_b w) dw)}\, du. \tag{7.9}$$

According to our assumptions,

$$\Lambda T_b = \frac{3}{4} \frac{1-p}{p^2} N_a \gg 1.$$

Therefore, applying Laplace's method for the approximation of the integral in (7.9) gives

$$\Pr\{T_1^P > t\} \approx F_T^c(t) e^{-\Lambda t} e^{\Lambda \int_0^t F_T(s)\, ds}.$$

For short times, $t \ll T_b$, $F_T(t) \ll 1$, so that we have the approximation

$$\Pr\{T_1^P > t\} = e^{-\Lambda t}(1 + o(1)) \quad \text{for } t \ll T_b.$$

Next, we consider the long time behavior of $\Pr\{T_1^P > t\}$. For times $t \gg T_b$, we have $F_T(s) \approx p$, and therefore

$$\begin{aligned} \int_0^{x+t} F_T(s)\, ds &= \int_0^x F_T(s)\, ds + \int_x^{x+t} F_T(s)\, ds \\ &\approx \int_0^x F_T(s)\, ds + tp. \end{aligned} \tag{7.10}$$

Inserting (7.10) into (7.8) gives

$$\Pr\{T_1^P > t\} \approx (1-p)\exp(-\lambda_B t) \quad \text{for } t \gg T_b.$$

To conclude, the interarrival time T_1^P is approximately exponential with rate Λ for short times, but due to possible particle recirculations its distribution has a different exponential tail with rate $\lambda_B < \Lambda$. Note, however, that since interarrival times are of the order of $1/\Lambda$, and $1/\Lambda \ll T_b$, all interarrival times of the pooled process are approximately exponentially distributed with rate Λ.

7.3. Simulation results. A simulation of the pooled process for uncharged particles has been run, according to the principles presented at the beginning of this section, with the typical values $a = 50\mathring{A}$, $r_0 = 80\mathring{A}$, $D = 10^{-9}\mathrm{m}^2/\mathrm{sec}$, and $\rho = 0.1\mathrm{M}$. These parameters give a value $p = 0.625$, for the return probability, and a value $N_a = 31$, for the average number of simulated particles inside the inner sphere. The average interarrival time of the pooled process is $1/\Lambda = 10^{-10}\mathrm{sec}$, compared to $T_b = 22.5 \times 10^{-10}\mathrm{sec}$, so that indeed $1/\Lambda \ll T_b$.

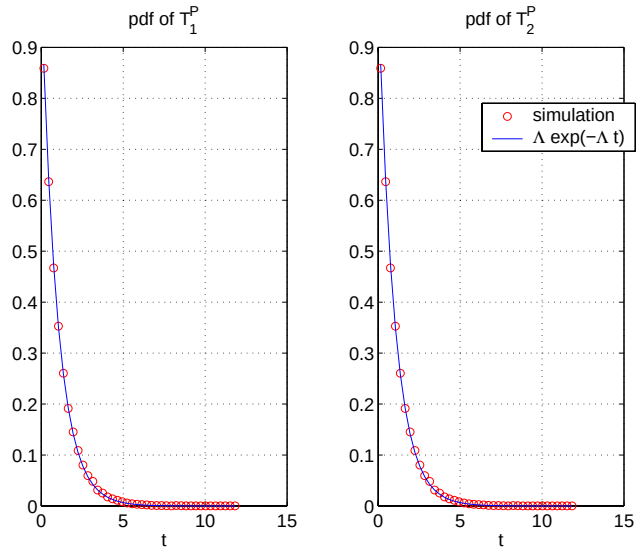


FIG. 3. *The pdfs of T_1^P (left) and T_2^P (right), as computed from the simulation, superimposed on the exponential density with rate Λ.*

In Figure 3 the graphs of the pdfs of T_1^P and T_2^P are superimposed on the exponential pdf with rate Λ. The pdfs of T_1^P and T_2^P are the result of about 250,000 samples of interarrival times of the pooled process. It is apparent that for times shorter than T_b both T_1^P and T_2^P are exponentially distributed with rate Λ. In all of the samples, not even once did an interarrival time longer than T_b occur in the simulation. Therefore, the theoretically predicted exponential tail with rate λ_B cannot be observed in these graphs. Finally, we note that T_1^P and T_2^P are dependent, with a correlation coefficient $r = \mathrm{cov}(T_1^P, T_2^P)/\sigma_1\sigma_2 \approx -0.001$. The fact that the correlation coefficient is negative is not surprising. It reflects the higher probability of recirculation in T_2^P when T_1^P is large. This means that when T_1^P is long there is a higher probability that T_2^P is short. The small correlation coefficient between T_1^P and T_2^P is a consequence of the short interarrival times of the pooled process, $1/\Lambda$, relative to the characteristic time for recirculation, T_b. A detailed analytical and numerical study of the statistical properties of the pooled process will be done elsewhere.

8. Summary and discussion. The time course of the exchange of ions between a test volume embedded in a continuum with a buffer region has been studied. The study of this time course is the basis for a simulation of uncharged and charged particles in a solution. The process of injecting new particles into the simulation has been identified as a stationary pooled process composed of an infinite number of delayed terminating independent renewal processes. While the pooled process converges to a stationary steady state, it is neither a renewal nor a Markov process. We have calculated the first two moments of this process, as well as the probability distribution of its residual time, and the joint distribution of the residual and the next arrival. From these calculations, it is apparent that the pooled process has an infinite memory. Therefore, to run the exact time course of this process in steady state, an infinite record of all past entrances and exits needs to be kept. To avoid this complexity, we have found that under some conditions on the size of the simulation

and buffer regions the pooled process can be approximated by a memoryless Poisson process. This approximation retains the average influx of the pooled process, but underestimates its variance. Our analysis shows how the parameters a and r_0 control the size of the simulation and the accuracy of our proposed approximation.

A closely related mathematical problem is considered in [30, chap. 5, sect. 9], where the renewal periods are assumed to have finite moments, in contrast with the case at hand in which the renewal periods may be infinite with positive probability. In [30], the case of infinitely many uniform sparse renewal processes is considered, and it is shown that under certain conditions on the sparseness, the pooled process becomes Poissonian as the number of processes increases to infinity. As we have seen, in our case the resulting process in not Poissonian.

The application of our results to an actual simulation is different for charged or uncharged particles. The approximations that we derived are not necessary for a simulation of uncharged Brownian particles, though they are necessary for a simulation of charged particles, as discussed below. For uncharged particles the stationary pooled process can be constructed offline to provide the random injection times of particles into the simulation. Such a pooled process has to be constructed for each choice of the parameters λ_B and p, that is, for each set of values for the parameters D, ρ, a, and r_0. With the correct choice of injection times, such a simulation reaches steady state immediately. If a wrong injection time course is adopted, there is a depletion or overcrowding of particles in the simulation region, which renders the simulation not self-consistent.

The simulation of charged particles is completely different from that of uncharged particles. In the uncharged case the mean field is always zero and thus the densities of all species in the bath remain constant throughout the time course of the simulation. In contrast, charged particles cause fluctuations in the net charge inside the simulation volume. The nonzero net charge creates a nonvanishing time dependent electrostatic field outside the simulation region that affects the continuum densities in the bath near the simulation region. Thus, if the net charge in the simulation region is positive, the bath density of positive charges decreases and the density of the negative charges increases in the neighborhood around the simulation region. These changes, in turn, affect the effective entrance rates of the different species into the simulation. Therefore, for charged particles, each configuration of net charge in the simulation region requires the construction of a new table of the pooled process suitable for it. Specifically, each entrance or exit changes the net charge inside the simulation region, which, in turn, changes the entrance rates into the simulation region.

Under the assumption of a *fast bath,*[1] the entrance law of a new ion into the simulation is that of the residual of the pooled process (6.9) that corresponds to the instantaneous concentrations due to the momentary net charge inside the simulation. These conditions change every time an ion enters or exits the simulation. Our analytical expression for the PDF of the residual time eliminates the need to run a simulation of the pooled process (as described in section 7) every time conditions change. In this way, a small portion of the bulk solution can be studied without the need to resort to ad hoc assumptions, such as artificial periodic boundary conditions.

In our model and analysis of a simulation of interacting ions (e.g., charged ions or ions with finite volume), we have used implicitly the concept of a *self-consistent* simulation. This notion is concerned with the detailed laws of ionic motion in the

[1] This means that the time to equilibrate the densities in the bath is shorter than the time between changes in the net charge inside the simulation region.

simulation region and the effective motion of ions in the surrounding continuum. The effective diffusive motion of ions is the result of their thermal interaction with the surrounding solvent and their interactions with each other. Thus, the motion inside the simulation volume is governed by both an unknown diffusion coefficient of ions in infinitely dilute solution and the interionic forces (computed by the simulation). On a sufficiently coarse time scale, the resulting motion of a simulated ion can be viewed as an effective diffusion with an effective diffusion coefficient that can be calculated from statistics of simulated trajectories. This calculated effective diffusion coefficient *must* be equal to the assumed (experimentally measured) diffusion coefficient in the bulk solution. This is a self-consistency requirement from the simulation. In addition, the average concentrations of the ionic species inside the simulation volume must be the same as those assumed in the bulk solution, as mentioned above. This is another self-consistency requirement. Still another self-consistency requirement is concerned with the notion of chemical activity. The concentrations of ions in the presence of an electrostatic field is different than that in the absence of a field, as can be readily seen from the Poisson–Boltzmann theory [12]. To compensate for the replacement of charged particles with independent uncharged particles the notion of activity has been introduced in physical chemistry [12]. More precisely, the change in the chemical potential as a function of particle density is assumed to take a simple form derived from the theory of gases, which replaces the physical density with a larger effective density. The ratio of the two densities is the activity factor. It is a directly measurable physical parameter. In a self-consistent simulation of charged particles, the unknown activity and diffusion coefficients in an infinitely dilute solution have to be chosen in such a way that all the above-mentioned self-consistency conditions are met. Finally, in addition to these self-consistency conditions, the electrostatic field has to be calculated in a self-consistent way at each time step of the simulation. A detailed description of a self-consistent simulation of charged particles will be presented in a separate paper. If the correct time course of the simulation is not followed it may be difficult to meet some of these self-consistency requirements.

Simulations of ions in solution have a wide range of applications. An important one is the theoretical study of permeation of uncharged molecules and ions through protein channels of biological membranes [34], [35]. Protein channels are small natural nano-devices of length in the range of 20–100$\mathring{A}$, and 5–20$\mathring{A}$ diameter. A computer simulation of a channel involves the simulation of the mobile ions both inside the channel and around it, in a volume comparable to the channel size. In the spirit of the theory discussed in this paper, such a simulation must be connected to its surrounding continuum. This leads to a small simulation with large time dependent density and potential fluctuations. The results of this paper are directly applicable to a simulation of permeation of uncharged molecules through protein channels, such as maltoporins that conduct sugar. Although the motion of the sugar molecules inside the maltoporin channel cannot be assumed a diffusion process and has to be simulated by molecular dynamics, the connection of the simulation to the continuum is described by the present work. For channels that conduct ions or other charged molecules, there are additional new elements in their simulation, namely, the presence of an impermeable membrane and a permanent charge profile of the channel itself, inside the simulation volume. The injection process of new particles into the simulation region on both sides of the membrane is similar to that described above. There are many differences, though, between a simulation with and without a channel, that are a subject for a separate study.

Appendix A. The time distribution outside the simulation. We now compute the distribution of τ_{out}, the time a nonsimulated (green) particle spends outside the simulation until its next re-entrance into the simulation. To this end, we denote by $p(r,t\,|\,r_0)$ the (radial) conditional probability density of the particle's location at time t, given that it has exited the simulation at the outer sphere at time $t=0$ and has not yet returned to the inner sphere. Obviously, in terms of this distribution,

$$\Pr\{\tau_{out} > t\} = \int_a^\infty 4\pi r^2 p_{out}(r,t\,|\,r_0)dr. \tag{A.1}$$

According to our assumptions, nonsimulated green particles perform independent free Brownian motion with diffusion coefficient D. Thus, the pdf $p_{out}(r,t\,|\,r_0)$ is the solution of the Fokker–Planck equation [31]

$$\begin{aligned} &\frac{\partial}{\partial t}p_{out}(r,t\,|\,r_0) = D\Delta p(r,t\,|\,r_0), \qquad a<r<\infty,\\ &p_{out}(a,t\,|\,r_0) = 0, \qquad t>0, \\ &p_{out}(r,0\,|\,r_0) = \frac{\delta(r-r_0)}{4\pi r_0^2}, \qquad a<r<\infty. \end{aligned} \tag{A.2}$$

The solution of (A.2) is given by

$$p_{out}\,(r,t\,|\,r_0) = \frac{1}{(4\pi Dt)^{1/2}}\frac{1}{4\pi r_0}\frac{1}{r}\left\{e^{-(r-r_0)^2/4D\,t} - e^{-(r+r_0-2a)^2/4D\,t}\right\}. \tag{A.3}$$

Inserting (A.3) into (A.1) gives

$$\Pr\{\tau_{out} \le t\,|\,r_0\} = p\frac{2}{\sqrt{\pi}}\int_{\sqrt{T_b/t}}^\infty e^{-u^2}\,du, \tag{A.4}$$

where

$$T_b = \frac{(r_0-a)^2}{4D} \tag{A.5}$$

is a characteristic time for the motion of a particle from the outer sphere to reach the inner sphere.

Equation (A.4) shows that τ_{out} can be infinite with probability $1-p>0$, that is, it has a defective probability distribution. It follows that it has an infinite mean value. The pdf of its defective distribution, given by

$$f_{\tau_{out}}(t) = \frac{d}{dt}\Pr\{\tau_{out}\le t\,|\,r_0\} = p\frac{1}{\sqrt{\pi}}\frac{\sqrt{T_b}}{t^{3/2}}\exp\left\{-\frac{T_b}{t}\right\},$$

gives

$$\int_0^\infty t f_{\tau_{out}}(t)\,dt = \infty. \tag{A.6}$$

Moreover, we have the following proposition.

PROPOSITION A.1. *The first moment of τ_{out}, conditioned on $\{\tau_{out} < \infty\}$, is infinite.*

Proof. Since

$$\Pr\{\tau_{out} < t \,|\, \tau_{out} < \infty\} = \frac{\Pr\{\tau_{out} < t,\, \tau_{out} < \infty\}}{\Pr\{\tau_{out} < \infty\}} = \frac{\Pr\{\tau_{out} < t\}}{p},$$

we obtain, in view of (A.6), that

$$\int_0^\infty t d \Pr\{\tau_{out} < t \,|\, \tau_{out} < \infty\} = \frac{1}{\sqrt{\pi}} \int_0^\infty \frac{\sqrt{T_b}}{t^{1/2}} \exp\left\{-\frac{T_b}{t}\right\} dt = \infty.$$

REFERENCES

[1] J. P. VALLEAU AND S. G. WHITTINGTON, *A guide to Monte Carlo for statistical mechanics:* 1. *Highways*, in Statistical Mechanics: Part A: Equilibrium Techniques, Modern Theoretical Chemistry 5, B. J. Berbe, ed., Plenum Press, New York, 1977, p. 15.

[2] G. KING AND A. WARSHEL, *A surface constraint all atom solvent model for effective simulations of polar solutions*, J. Chem. Phys., 91 (1989), p. 3647.

[3] J. M. CAILLOL, D. LEVESQUE, AND J. J. WEISS, *Electrical properties of polarizable ionic solutions* II: *Computer simulation results*, J. Chem. Phys., 91 (1989), pp. 5555–5566.

[4] J. M. CAILLOL, *A new potential for the numerical simulations of electrolyte solutions on a hyper-sphere*, J. Chem. Phys., 99 (1993), pp. 8953–8963.

[5] Y.-Y. SHAMM AND A. WARSHEL, *The surface constraint all atom model provides size independent results in calculations hydration free energies*, J. Chem. Phys., 109 (1998), pp. 7940–7944.

[6] G. CICCOTTI AND A. TENENBAUM, *Canonical ensemble and nonequilibrium states by molecular dynamics*, J. Statist. Phys., 23 (1980), p. 767.

[7] M. BERKOWITZ AND J. A. MCCAMMON, *Molecular dynamics with stochastic boundary conditions*, Chem. Phys. Lett., 90 (1982), pp. 215–217.

[8] C. L BROOKS III AND M. KARPLUS, *Deformable stochastic boundaries in molecular dynamics*, J. Chem. Phys., 79 (1983), p. 6312.

[9] A. C. BELCH AND M. BERKOWITZ, *Molecular dynamics simulations of TIPS*2 *water restricted by a spherical hydrophobic boundary*, Chem. Phys. Lett., 113 (1985), pp. 278–282.

[10] I. M. WONPIL, S. SEEFELD, AND B. ROUX, *A grand canonical Monte Carlo-Brownian dynamics algorithm for simulating ion channel*, Biophys. J., 79 (2000), p. 788–801.

[11] S. CHANDRASEKHAR, *Stochastic problems in physics and astronomy*, Rev. Mod. Phys., 15 (1943), pp. 1–89.

[12] R. S. BERRY, S. RICE, AND J. ROSS, *Physical Chemistry*, 2nd ed., Oxford University Press, Oxford, 2000.

[13] S. G. BRUSH, *The Kind of Motion We Call Heat*, Vol. I, North–Holland, Amsterdam, 1986.

[14] S. G. BRUSH, *The Kind of Motion We Call Heat*, Vol. II, North–Holland, Amsterdam, 1986.

[15] M. P. ALLEN AND D. J. TILDESLEY, *Computer Simulation of Liquids*, Oxford University Press, Oxford, 1991.

[16] B. CORRY, S. KUYUCAK, AND S. CHUNG, *Tests of continuum theories as models of ion channels.* II. *Poisson-Nernst-Planck theory versus Brownian dynamics*, Biophys. J., 78 (2000), pp. 2364–2381.

[17] F. L. ROMAN, J. A. WHITE, AND S. VELASCO, *Fluctuations in an equilibrium hard-disk fluid: Explicit size effects*, J. Chem. Phys., 107 (1997), pp. 4635–4641.

[18] F. L. ROMAN, J. A. WHITE, AND S. VELASCO, *Fluctuations in the number of particles of the ideal gas: A simple example of explicit finite-size effects*, Amer. J. Phys., 67 (1999), pp. 1149–1151.

[19] J. HORBACH, W. KOB, K. BINDER, AND C. A. ANGELL, *Finite size effects in simulations of glass dynamics*, Phys. Rev. E., 54 (1996), pp. 5897–5900.

[20] T. M. NYMAND AND P. LINSE, *Ewald summation and reaction field methods for potentials with atomic charges, dipoles, and polarizabilities*, J. Chem. Phys., 112 (2000), pp. 6152–6160.

[21] J. M. CAILLOL, *A Monte Carlo study of the dielectric constant of the restricted primitive model of electrolytes on the vapor branch of the coexistence line*, J. Chem. Phys., 102 (1995), pp. 5471–5479.

[22] R. A. Friedman and M. Mezei, *The potentials of mean force of sodium chloride and sodium dimethylphosphate in water: An application of adaptive umbrella sampling*, J. Chem. Phys., 102 (1995), pp. 419–426.
[23] M. P. Allen and D. J. Tildesley, eds., *Computer Simulation in Chemical Physics*, NATO ASI Ser. C Math. Phys. Sci. 397, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1992.
[24] B. Nadler, T. Naeh, and Z. Schuss, *The stationary arrival process of independent diffusers from a continuum to an absorbing boundary is Poissonian*, SIAM J. Appl. Math., 62 (2001), pp. 433–447.
[25] R. von Smoluchowski, *Versuch einer mathematischen Theorie der Koagulationskinetic koloider Lösungen*, Zeits. Phys. Chem., 92 (1917), p. 129.
[26] P. Bordewijk, *Defect-diffusion models of dielectric relaxation*, Chem. Phys. Lett., 32 (1975), pp. 592–596.
[27] T. Naeh-Galor, *Simulation of Ionic Solution*, Ph.D. dissertation, Tel-Aviv University, Tel Aviv, Israel, 2001.
[28] W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 2, John Wiley, New York, 1970.
[29] B. Nadler, *Density Fluctuations*, M.Sc. dissertation, Tel-Aviv University, Tel Aviv, Israel, 1994.
[30] S. Karlin and H. M. Taylor, *A First Course in Stochastic Processes*, Academic Press, New York, 1975.
[31] Z. Schuss, *Theory and Application of Stochastic Differential Equation*, John Wiley, New York, 1980.
[32] A. Papoulis, *Probability, Random Variables and Stochastic Processes*, 3rd ed., McGraw-Hill, New York, Toronto, London, 1991.
[33] D. R. Cox and P. A. W. Lewis, *The Statistical Analysis of Series of Events*, Methuen, London, 1968.
[34] B. Hille, *Ionic Channels of Excitable Membranes*, 2nd ed., Sinauer Associates, Sutherland, MA, 1992.
[35] R. S. Eisenberg, *From structure to function in open ionic channels*, J. Mem. Biol., 171 (1999), pp. 1–24.

SIAM J. APPL. MATH.
Vol. 63, No. 3, pp. 874–888

DECAY OF KDV SOLITONS*

HIÊÚ D. NGUYÊÑ†

Abstract. In this paper we develop a linear eigenvalue decomposition for N-soliton solutions of the Korteweg–de Vries equation and use it to obtain a new mathematical explanation of two-soliton interaction in terms of particle decay. We discover that the two-soliton "particles" or pulses which appear in each solution exchange identities upon collision and emit a dual "ghost" particle pair in order to conserve mass and momentum.

Key words. solitons, Korteweg–de Vries (KdV) equation, particle decay, ghost particles

AMS subject classifications. 35Q51, 35Q53

PII. S0036139902402695

1. Introduction.

It is well known that the Korteweg–de Vries (KdV) equation,

$$u_t - 6uu_x + u_{xxx} = 0,$$

is a model for many wave related phenomena and admits a special family of localized solutions called N-solitons corresponding to reflectionless potentials (cf. [M]). Here, N denotes the number of solitons, i.e., the number of pulses or potential wells, that appear in each solution. One-solitons or solitary waves were first observed by J. Scott Russell along the Union Canal at Edinburgh in 1834 (cf. [M]). Then in 1895, Korteweg and de Vries [KV] published their (KdV) equation as a model for these waves. However, it would require another seventy years before two-soliton interaction was observed by Zabusky and Kruskal [ZK] through numerical calculation; they reported that "solitons 'pass through' one another without losing their identity." The exact interaction of two solitons was then determined numerically by Zabusky [Z] and soon thereafter Lax [La] gave a mathematical proof.

The idea that perhaps solitons actually bounce off each other upon collision dates back to Bowtell and Stuart [BS]. The exchange of mass that occurs then between the two colliding soliton particles allows them to exchange identities. More recent work advocating this viewpoint can be found in [Le] and [MC]. Now, to mathematically investigate such behavior, it is desirable to isolate each particle in any given N-soliton solution. This can be achieved, say, by decomposing the solution into a linear sum, even though the KdV equation itself is nonlinear, so that the superposition principle fails to hold. To this end, various such decompositions can be found in the literature (cf. [GGKM], [HM], [S], [MC]). We shall discuss some of these decompositions in relation to ours at the end of this paper.

In this paper, we develop a linear eigenvalue decomposition of N-soliton solutions for the Korteweg–de Vries equation and use it to obtain a new mathematical explanation of two-soliton interaction in terms of particle decay. This decomposition is obtained through a diagonalization procedure that is applied to the corresponding soliton matrix and has the effect of isolating the decay of each soliton "particle." For two-solitons, the interaction described by Theorem 3.3 suggests a decay phenomenon that occurs frequently in elementary particle physics: the two-soliton particles split

*Received by the editors February 19, 2002; accepted for publication (in revised form) June 24, 2002; published electronically January 23, 2003.
http://www.siam.org/journals/siap/63-3/40269.html
†Department of Mathematics, Rowan University, Glassboro, NJ 08028 (nguyen@rowan.edu).

upon collision, resulting in an exchange of identities and the emission of a dual "ghost" particle pair (cf. Figure 1). Theorem 3.4 then shows that each decay process conserves mass and momentum and supports our particle decay interpretation of soliton interaction. Interesting properties of our dual ghost particles are then described in Theorem 3.6. In fact, we like to view each ghost particle as a nonlinear "difference" between two given soliton particles. Lastly, an explicit example is given in section 4 to illustrate our results (cf. Figures 2–4).

2. Soliton particles. Let N be a positive integer and assume that the initial scattering data for $u(x,0)$, obtained through the time-independent Schrodinger equation

$$\psi_{xx} - [\lambda - u(x,0)]\psi = 0, \tag{1}$$

has only a discrete energy spectrum. This means that λ takes on a discrete set of N negative energy eigenvalues $\{\lambda_1 < \lambda_2 < \cdots < \lambda_N < 0\}$ with corresponding eigenfunctions $\{\psi_1, \psi_2, \ldots, \psi_N\}$. It is standard that we normalize these eigenfunctions and compute their normalized factors c_n, commonly referred to as "phase shifts":

$$\int_{-\infty}^{\infty} \psi_n^2 dx = 1, \qquad c_n = \lim_{x \to -\infty} e^{k_n x}\psi_n. \tag{2}$$

The initial scattering data is then used to produce the N-soliton solution of the KdV equation through the determinant formula

$$u(x,t) = -2\frac{\partial^2}{\partial x^2} \log \det(I + A). \tag{3}$$

Here, the $N \times N$ *soliton matrix* A has entries defined by

$$A = (a_{mn}), \qquad a_{mn} = \frac{c_m c_n}{k_m + k_n} e^{(k_m + k_n)x - 4(k_m^3 + k_n^3)t}, \tag{4}$$

where the spectral parameter $k_n > 0$ is defined via the relation $\lambda_n = -k_n^2$. This solution was obtained independently in the early 1970s by Gardner et al. [GGKM] and Wadati and Toda [WT], both groups by means of the inverse scattering method, and by Hirota [H] through his direct method.

We now turn to developing our working definition of a soliton particle. It is well known that A is symmetric and positive definite (cf. [KM], [GGKM], [WT]). This allows us to diagonalize it so that

$$B^{-1}AB = D = \begin{pmatrix} \mu_1(x,t) & 0 & \ldots & 0 \\ 0 & \cdots & & \\ \cdots & & & \\ 0 & & & \mu_N(x,t) \end{pmatrix}. \tag{5}$$

Here, $\{\mu_1 > \cdots > \mu_N\}$ is the (ordered) set of real positive eigenvalues of A and B is the orthogonal matrix consisting of an orthonormal basis of eigenvectors of A. It follows that we can write $u(x,t)$ in terms of $\{\mu_n\}$, which we shall refer to as *decay eigenvalues*:

$$u(x,t) = -2\frac{\partial^2}{\partial x^2} \log \det(I + A) \tag{6}$$

$$= -2\frac{\partial^2}{\partial x^2}\log\det[B^{-1}(I+A)B] \tag{7}$$

$$= -2\frac{\partial^2}{\partial x^2}\log\det(I+D) \tag{8}$$

$$= -2\frac{\partial^2}{\partial x^2}\log\prod_{n=1}^{N}[1+\mu_n(x,t)] \tag{9}$$

$$= \sum_{n=1}^{N} -2\frac{\partial^2}{\partial x^2}\log[1+\mu_n(x,t)]. \tag{10}$$

DEFINITION 2.1. *Define*

$$s_n(\nu_n) \equiv -2k_n^2\operatorname{sech}^2(k_n\nu_n), \qquad n=1,\dots,N, \tag{11}$$

to be the nth soliton particle *of u, where* $\nu_n = x - 4k_n^2 t$ *is the nth moving frame. Then we shall refer to*

$$u_n(x,t) \equiv -2\frac{\partial^2}{\partial x^2}\log[1+\mu_n(x,t)] \tag{12}$$

as the decay function *of* s_n *and to the sum* $u = \sum_{n=1}^{N} u_n$ *as derived in* (10) *as the* decay decomposition *of u. The results of the next section will justify our use of terminology.*

3. Decay of two-solitons. In this section we assume $N = 2$ and investigate the asymptotic behavior of the decay functions u_1 and u_2 as a means of understanding soliton interaction. We begin by writing the matrix A explicitly in terms of the two moving frames ν_1 and ν_2:

$$A = \begin{pmatrix} \frac{c_1^2}{2k_1}e^{2k_1\nu_1} & \frac{c_1c_2}{k_1+k_2}e^{k_1\nu_1+k_2\nu_2} \\ \frac{c_1c_2}{k_1+k_2}e^{k_1\nu_1+k_2\nu_2} & \frac{c_2^2}{2k_2}e^{2k_2\nu_2} \end{pmatrix}. \tag{13}$$

Denoting by $p = \operatorname{Tr}(A)$ and $q = \det(A)$, it follows that the two eigenvalues of A are given by

$$\mu_1 = \frac{1}{2}\left(p + \sqrt{p^2-4q}\right), \tag{14}$$

$$\mu_2 = \frac{1}{2}\left(p - \sqrt{p^2-4q}\right). \tag{15}$$

DEFINITION 3.1. *We define*

$$A_g = \begin{pmatrix} \frac{c_1^2}{2k_1}e^{2k_1\nu_g} & \frac{c_1c_2}{k_1+k_2}e^{(k_1+k_2)\nu_g} \\ \frac{c_1c_2}{k_1+k_2}e^{(k_1+k_2)\nu_g} & \frac{c_2^2}{2k_2}e^{2k_2\nu_g} \end{pmatrix} \tag{16}$$

to be the ghost matrix *of A where* $\nu_g = x - 4k_g^2 t$ *and* $k_g = (k_1^2 + k_1k_2 + k_2^2)^{1/2}$. *In addition, if* γ_1 *and* γ_2 *denote the eigenvalues of* A_g *corresponding to* μ_1 *and* μ_2, *respectively, then we shall refer to*

$$g(\nu_g) \equiv -2\frac{\partial^2}{\partial \nu_g^2}\log[\gamma_1(\nu_g)] \tag{17}$$

as the ghost particle *and*

$$\bar{g}(\nu_g) \equiv -2\frac{\partial^2}{\partial \nu_g^2}\log[\gamma_2(\nu_g)] \tag{18}$$

as the antighost particle *corresponding to the pair* $\{u_1, u_2\}$.

Note that ν_g represents the moving frame of both g and $\bar{g}$ and that $4k_g^2$ represents their velocity and exceeds that of the two-soliton particles. The following lemma assures us that the above-mentioned correspondence between the two sets of eigenvalues is well defined.

LEMMA 3.2. *Denote by* $\hat{k}^2 = k_1^2 k_2 + k_1 k_2^2$. *Then*

$$A = e^{8\hat{k}^2 t} A_g.$$

Moreover, $\mu_n = e^{8\hat{k}^2 t}\gamma_n$ *for* $n = 1, 2$.

Proof. It suffices to prove that every coefficient of A has $e^{8\hat{k}^2 t}$ as a common factor when rewritten in terms of ν_g. This quickly follows from the relation

$$\begin{aligned} e^{k_n(x-4k_n^2 t)} &= e^{k_n(\nu_g + 4k_g^2 t - 4k_n^2 t)} \\ &= e^{k_n\nu_g + 4(k_1^2 k_2 + k_1 k_2^2)t} \\ &= e^{4\hat{k}^2 t} e^{k_n \nu_g}. \end{aligned}$$

The fact that $\mu_n = e^{8\hat{k}^2 t}\gamma_n$ also follows from this relation and can be easily checked by the reader. □

We are now ready to present our theorem describing particle decay of two-solitons. This will justify our use of the terms "particle" and "decay" in referring to s_n and u_n, respectively.

THEOREM 3.3. *The following asymptotic relations hold for* u_1 *and* u_2*:*

(i)

$$\begin{aligned} &u_1 \sim s_1(\nu_1 + \delta_1), && \text{as } t \to -\infty, \\ &u_1 \sim s_2(\nu_2 + \delta_2) + g(\nu_g), && \text{as } t \to \infty, \end{aligned}$$

in the sense that

$$\lim_{\substack{\nu_1 \text{ fixed} \\ t \to -\infty}} u_1 = s_1(\nu_1 + \delta_1), \qquad \lim_{\substack{\nu_2 \text{ fixed} \\ t \to \infty}} u_1 = s_2(\nu_2 + \delta_2), \qquad \lim_{\substack{\nu_g \text{ fixed} \\ t \to \infty}} u_1 = g(\nu_g).$$

Here, the relative phase shifts δ_1 *and* δ_2 *are defined by*

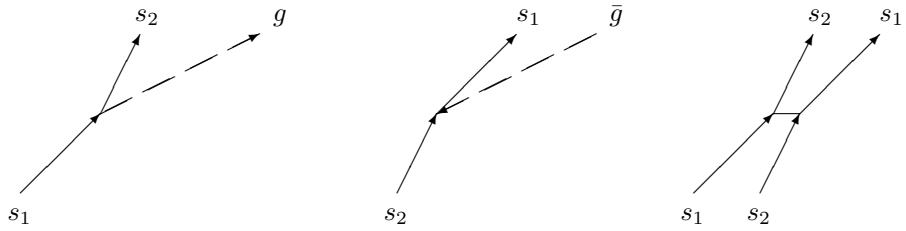
$$e^{2k_1\delta_1} = \frac{c_1^2}{2k_1}, \qquad e^{2k_2\delta_2} = \frac{c_2^2}{2k_2}.$$

(ii)

$$\begin{aligned} &u_2 \sim s_2(\nu_2 + \delta_2 + \Delta), && \text{as } t \to -\infty, \\ &u_2 \sim s_1(\nu_1 + \delta_1 + \Delta) + \bar{g}(\nu_g), && \text{as } t \to \infty. \end{aligned}$$

Here, Δ *is defined by*

$$e^{2k_2\Delta} = \frac{(k_1 - k_2)^2}{(k_1 + k_2)^2}.$$



(a) Diagram of u_1 (b) Diagram of u_2 (c) Diagram of u

FIG. 1. *Space-time plots of two-soliton decay.*

Following the physics literature we shall summarize the decay described by u_1 and u_2 as follows:

$$\begin{aligned} u_1: &\quad s_1 \to s_2 + g, \\ u_2: &\quad s_2 \to s_1 + \bar{g}. \end{aligned}$$

The corresponding space-time plots are drawn in Figure 1. Notice that they describe the exchange of identities between s_1 and s_2 and the fact that the emitted ghost particles (represented by the dashed lines) have velocities greater than both soliton particles.

Proof of Theorem 3.3. (i) Our approach is to analyze u_1 from the perspective of the three moving frames corresponding to the velocities $\nu_1,\ \nu_2$, and ν_g and to treat each as a separate case.

Case I. Assume ν_1 is fixed. We rewrite the trace and determinant of A as

$$\begin{aligned} p &= \mathrm{Tr}(A) \\ &= \tfrac{c_1^2}{2k_1} e^{2k_1\nu_1} + \tfrac{c_2^2}{2k_2} e^{2k_2\nu_2} \\ &= e^{2k_1\nu_1} \left(\tfrac{c_1^2}{2k_1} + \tfrac{c_2^2}{2k_2} e^{2k_2\nu_2 - 2k_1\nu_1} \right) \\ &= e^{2k_1\nu_1} \left(\tfrac{c_1^2}{2k_1} + \tfrac{c_2^2}{2k_2} e^{2(k_2-k_1)\nu_1 + 8k_2(k_1^2-k_2^2)t} \right) \end{aligned}$$

and

$$\begin{aligned} q &= \det(A) \\ &= \left(\tfrac{k_1-k_2}{k_1+k_2} \right)^2 \tfrac{c_1^2 c_2^2}{4k_1k_2} e^{2(k_1\nu_1+k_2\nu_2)} \\ &= e^{2k_1\nu_1} \left(\tfrac{k_1-k_2}{k_1+k_2} \right)^2 \tfrac{c_1^2 c_2^2}{4k_1k_2} e^{2k_2\nu_1 + 8(k_1^2-k_2^2)t} \end{aligned}$$

so that

$$\lim_{\substack{\nu_1 \text{ fixed} \\ t\to-\infty}} p = \frac{c_1^2}{2k_1} e^{2k_1\nu_1}, \qquad \lim_{\substack{\nu_1 \text{ fixed} \\ t\to-\infty}} q = 0.$$

This forces

$$\begin{aligned}\lim_{\substack{\nu_1 \text{ fixed}\\ t\to-\infty}}(1+\mu_1) &= \lim_{\substack{\nu_1 \text{ fixed}\\ t\to-\infty}}\left[1+\frac{1}{2}\left(p+\sqrt{p^2-4q}\right)\right]\\ &= 1+\tfrac{c_1^2}{2k_1}e^{2k_1\nu_1}\end{aligned}$$

and implies

$$\begin{aligned}\lim_{\substack{\nu_1 \text{ fixed}\\ t\to-\infty}}u_1 &= \lim_{\substack{\nu_1 \text{ fixed}\\ t\to-\infty}}\left\{-2\frac{\partial^2}{\partial x^2}\log(1+\mu_1)\right\}\\ &= -2\tfrac{\partial^2}{\partial x^2}\log(1+\tfrac{c_1^2}{2k_1}e^{2k_1\nu_1})\\ &= \frac{-8k_1c_1^2e^{2k_1\nu_1}}{\left(1+\frac{c_1^2}{2k_1}e^{2k_1\nu_1}\right)^2}\\ &= s_1(\nu_1+\delta_1),\end{aligned}$$

where δ_1 is defined by $e^{2k_1\delta_1} = \frac{c_1^2}{2k_1}$. Note that we have implicitly used the fact $\frac{\partial}{\partial x} = \frac{\partial}{\partial \nu_1}$.

Case II. Assume that ν_2 is fixed. We proceed in the same manner as Case I but factor $e^{2k_2\nu_2}$ instead of $e^{2k_1\nu_1}$ from p and q. It is then a straightforward exercise to show that

$$\lim_{\substack{\nu_2 \text{ fixed}\\ t\to\infty}}u_1 = s_2(\nu_2+\delta_2),$$

where this time δ_2 is defined by $e^{2k_2\delta_2} = \frac{c_2^2}{2k_2}$.

Case III. Assume ν_g is fixed. Applying Lemma 3.2, we obtain, for $t \to -\infty$,

$$\begin{aligned}\lim_{\substack{\nu_g \text{ fixed}\\ t\to-\infty}}u_1 &= \lim_{\substack{\nu_g \text{ fixed}\\ t\to-\infty}}\left\{-2\frac{\partial^2}{\partial x^2}\log(1+\mu_1)\right\}\\ &= \lim_{\substack{\nu_g \text{ fixed}\\ t\to-\infty}}\left\{-2\frac{\partial^2}{\partial x^2}\log(1+e^{4\hat{k}^2t}\gamma_1)\right\}\\ &= 0.\end{aligned}$$

On the other hand, for $t \to \infty$,

$$\begin{aligned}\lim_{\substack{\nu_g \text{ fixed}\\ t\to\infty}}u_1 &= \lim_{\substack{\nu_g \text{ fixed}\\ t\to\infty}}\left\{-2\frac{\partial^2}{\partial x^2}\log(1+e^{4\hat{k}^2t}\gamma_1)\right\}\\ &= \lim_{\substack{\nu_g \text{ fixed}\\ t\to\infty}}\left\{-2\frac{\partial^2}{\partial x^2}\log e^{4\hat{k}^2t}+\log(e^{-4\hat{k}^2t}+\gamma_1)\right\}\\ &= \lim_{\substack{\nu_g \text{ fixed}\\ t\to\infty}}\left\{-2\frac{\partial^2}{\partial x^2}\log(e^{-4\hat{k}^2t}+\gamma_1)\right\}\\ &= -2\tfrac{\partial^2}{\partial \nu_g^2}\log\gamma_1\\ &= g.\end{aligned}$$

This completes the proof of part (i).

(ii) We apply a similar analysis to u_2 by again considering three separate cases.

Case I. Assume ν_2 is fixed. We rewrite p and q as

$$p = e^{8k_1(k_2^2-k_1^2)t}\left(\frac{c_1^2}{2k_1}e^{2k_1\nu_2} + \frac{c_2^2}{2k_2}e^{2k_2\nu_2-8k_1(k_2^2-k_1^2)t}\right),$$
$$q = e^{8k_1(k_2^2-k_1^2)t}\frac{(k_1-k_2)^2}{(k_1+k_2)^2}\frac{c_1^2c_2^2}{4k_1k_2}e^{2(k_1+k_2)\nu_2}.$$

The relations

$$\lim_{\substack{\nu_2 \text{ fixed}\\ t\to-\infty}}\frac{q}{p} = \frac{(k_1-k_2)^2}{(k_1+k_2)^2}\frac{c_2^2}{2k_2}e^{2k_2\nu_2}, \qquad \lim_{\substack{\nu_2 \text{ fixed}\\ t\to-\infty}}\frac{q}{p^2} = 0$$

now tell us how μ_2 behaves in the limit once we rationalize it:

$$\begin{aligned}\lim_{\substack{\nu_2 \text{ fixed}\\ t\to-\infty}}\mu_2 &= \lim_{\substack{\nu_2 \text{ fixed}\\ t\to-\infty}}\left\{\frac{1}{2}\left(p-\sqrt{p^2-4q}\right)\frac{p+\sqrt{p^2-4q}}{p-\sqrt{p^2-4q}}\right\}\\ &= \lim_{\substack{\nu_2 \text{ fixed}\\ t\to-\infty}}\left\{\frac{\frac{2q}{p}}{1+\sqrt{1-\frac{4q}{p^2}}}\right\}\\ &= \frac{(k_1-k_2)^2}{(k_1+k_2)^2}\frac{c_2^2}{2k_2}e^{2k_2\nu_2}.\end{aligned}$$

Hence,

$$\begin{aligned}\lim_{\substack{\nu_2 \text{ fixed}\\ t\to-\infty}}u_2 &= \lim_{\substack{\nu_2 \text{ fixed}\\ t\to-\infty}}\left\{-2\frac{\partial^2}{\partial x^2}\log(1+\mu_2)\right\}\\ &= \lim_{\substack{\nu_2 \text{ fixed}\\ t\to-\infty}}\left\{-2\frac{\partial^2}{\partial x^2}\log\left[1+\frac{(k_1-k_2)^2}{(k_1+k_2)^2}\frac{c_2^2}{2k_2}e^{2k_2\nu_2}\right]\right\}\\ &= s_2(\nu_2+\delta_2+\Delta),\end{aligned}$$

where Δ is defined by $e^{2k_2\Delta} = \frac{(k_1-k_2)^2}{(k_1+k_2)^2}$.

Case II. Assume that ν_1 is fixed. As the line of argument here is the same as that for Case I with ν_2 fixed, we leave it for the reader to verify that

$$\lim_{\substack{\nu_1 \text{ fixed}\\ t\to\infty}}u_2 = s_1(\nu_1+\delta_1+\Delta).$$

Case III. Assume that ν_g is fixed. The argument establishing

$$\lim_{\substack{\nu_g \text{ fixed}\\ t\to\infty}}u_2 = \bar{g}$$

is exactly the same as that for Case III in (i) and will be left for the reader. This completes the proof of our theorem. □

The following result provides evidence to support our theory of soliton decay.

THEOREM 3.4. (i) *Conservation of mass.*

$$\int_{-\infty}^{\infty}u_n(x,t)dx = -4k_n, \qquad n=1,2.$$

(ii) *Conservation of momentum.*

$$\frac{d}{dt}\int_{-\infty}^{\infty}xu_n(x,t)dx = -16k_n^3, \qquad n=1,2.$$

Proof. (i) For u_1, we have

$$\begin{aligned}\int_{-\infty}^{\infty} u_1(x,t)dx &= \int_{-\infty}^{\infty}\left[-2\tfrac{\partial^2}{\partial x^2}\log(1+\mu_1)\right]dx\\ &= \left[-2\tfrac{\partial}{\partial x}\log(1+\mu_1)\right]_{-\infty}^{\infty}\\ &= -2\left[\tfrac{\mu_1'}{1+\mu_1}\right]_{-\infty}^{\infty}\\ &= -4k_1.\end{aligned}$$

A similar argument applied to u_2 (after first rationalizing μ_2) shows that

$$\int_{-\infty}^{\infty} u_2(x,t)dx = -4k_2.$$

(ii) Integration by parts yields

$$\begin{aligned}\int_{-\infty}^{L} xu_n(x,t)dx &= \left[-2x\tfrac{\partial}{\partial x}\log(1+\mu_n)\right]_{-\infty}^{L} - \int_{-\infty}^{L}\left[-2\tfrac{\partial}{\partial x}\log(1+\mu_n)\right]dx\\ &= -2L\tfrac{\mu_n'(L)}{(1+\mu_n(L)}+2\log(1+\mu_n(L))\\ &\sim -4k_nL+4k_n(L-4k_n^2t+\delta_n)\end{aligned}$$

as $L\to\infty$. It follows that

$$\frac{d}{dt}\int_{-\infty}^{\infty} xu_n(x,t)dx = -16k_n^3, \qquad n=1,2.$$

This completes the proof. □

For $n=1,2$, we define the *center of mass* of u_n to be

$$x_n(t) \equiv \frac{\int_{-\infty}^{\infty} xu_n(x,t)dx}{\int_{-\infty}^{\infty} u_n(x,t)dx}. \tag{19}$$

The next corollary follows immediately from Theorem 3.4.

COROLLARY 3.5. *The center of mass $x_n(t)$ as defined by* (19) *moves with constant velocity $4k_n^2$, i.e.,*

$$\frac{dx_n}{dt} = 4k_n^2, \qquad n=1,2.$$

Let us now investigate our ghost particles a little more closely. We begin with the following theorem which justifies our use of the terms "ghost" and "antighost" for g and $\bar{g}$ as they do not appear in u due to cancellation.

THEOREM 3.6. *The ghost particles g and $\bar{g}$ enjoy the following properties:*

(i) $g+\bar{g}=0$.

(ii) $\int_{-\infty}^{\infty} g(\nu_g)d\nu_g = 4(k_1-k_2)$.

(iii) $g = -32k_1k_2(\frac{p_gq_g}{r_g^{3/2}}) < 0$, *where* $p_g = \mathrm{Tr}(A_g)$, $q_g=\det(A)$, *and* $r_g = p_g^2-4q_g$.

(iv) $g(\nu_g) = O(\mathrm{sech}^2[(k_1-k_2)(\nu_g+\delta_g)])$ *as* $\nu_g\to\pm\infty$, *where δ_g is defined by the relation* $e^{2(k_1-k_2)\delta_g} = \frac{c_1^2k_2}{c_2^2k_1}$.

(v) $|g(\nu_g)| \le (k_1-k_2)^2(k_1+k_2)/\sqrt{k_1k_2}$ *with equality holding precisely when* $\nu_g = -\delta_g$.

Proof. (i) If one recalls that

$$\begin{aligned}\gamma_1\gamma_2 &= \det(A_g)\\ &= \tfrac{(k_1-k_2)^2}{(k_1+k_2)^2}\tfrac{c_1^2c_2^2}{4k_1k_2}e^{2(k_1+k_2)\nu_g},\end{aligned}$$

then it directly follows that

$$\begin{aligned}g+\bar{g} &= -2\tfrac{\partial^2}{\partial\nu_g^2}\log(\gamma_1\gamma_2)\\ &= 0.\end{aligned}$$

(ii) We have

$$\begin{aligned}\int_{-\infty}^{\infty} g(\nu_g)d\nu_g &= \int_{-\infty}^{\infty}\left[-2\tfrac{\partial^2}{\partial\nu_g^2}\log\gamma_1\right]d\nu_g\\ &= \left[-2\tfrac{\gamma_1'}{\gamma_1}\right]_{-\infty}^{\infty}.\end{aligned} \tag{20}$$

Substituting the relations

$$\lim_{\nu_g\to-\infty}\frac{\gamma_1'}{\gamma_1}=2k_2, \qquad \lim_{\nu_g\to\infty}\frac{\gamma_1'}{\gamma_1}=2k_1.$$

into (20) then yields the desired result:

$$\int_{-\infty}^{\infty} g(\nu_g)d\nu_g = 4(k_2-k_1).$$

We note that this result also follows directly from Theorem 3.4 due to conservation of mass of u_1.

(iii) First write γ_1 in the form

$$\gamma_1 = \frac{1}{2}\left(p_g+\sqrt{r_g}\right), \tag{21}$$

where

$$p_g = \mathrm{Tr}(A_g) = \frac{c_1^2}{2k_1}e^{2k_1\nu_g}+\frac{c_2^2}{2k_2}e^{2k_2\nu_g}, \tag{22}$$

$$q_g = \det(A_g) = \frac{(k_1-k_2)^2}{(k_1+k_2)^2}\frac{c_1^2c_2^2}{4k_1k_2}e^{2(k_1+k_2)\nu_g}, \tag{23}$$

$$r_g = p_g^2-4q_g. \tag{24}$$

Then we can express γ_1 in terms of an appropriate hyperbolic cosine function by introducing the identity

$$p_g = \frac{c_1c_2}{\sqrt{k_1k_2}}e^{(k_1+k_2)\nu_g}\cosh[(k_1-k_2)(\nu_g+\delta_g)], \tag{25}$$

where δ_g is defined by the relation $e^{2(k_1-k_2)\delta_g}=\frac{c_1^2k_2}{c_2^2k_1}$. It follows that

$$\begin{aligned}\gamma_1 = \frac{c_1c_2}{\sqrt{k_1k_2}}e^{(k_1+k_2)\nu_g}\Bigg(&\cosh\left[(k_1-k_2)(\nu_g+\delta_g)\right]\\ &+\sqrt{\cosh^2\left[(k_1-k_2)(\nu_g+\delta_g)\right]-\frac{(k_1-k_2)^2}{(k_1+k_2)^2}}\Bigg)\end{aligned} \tag{26}$$

$$= \frac{(k_1-k_2)}{(k_1+k_2)}\frac{c_1c_2}{\sqrt{k_1k_2}}e^{(k_1+k_2)\nu_g}\left(z+\sqrt{z^2-1}\right), \tag{27}$$

where $z = \frac{(k_1+k_2)}{(k_1-k_2)} \cosh[(k_1-k_2)(\nu_g+\delta_g)]$. Therefore,

$$g = -2\frac{\partial^2}{\partial \nu_g^2} \log \gamma_1 \tag{28}$$

$$= -2\frac{\partial^2}{\partial \nu_g^2} \left[\log\left(\frac{(k_1k_2)}{(k_1+k_2)} \frac{c_1c_2}{\sqrt{k_1k_2}} e^{(k_1+k_2)\nu_g} \right) + \log(z+\sqrt{z^2-1}) \right] \tag{29}$$

$$= -2\frac{\partial^2}{\partial \nu_g^2} \cosh^{-1} z \tag{30}$$

$$= -8k_1k_2 \frac{z}{(z^2-1)^{3/2}} \tag{31}$$

$$= -8k_1k_2 \frac{\frac{(k_1+k_2)}{(k_1-k_2)} \cosh[(k_1-k_2)(\nu_g+\delta_g)]}{\left[\frac{(k_1+k_2)^2}{(k_1-k_2)^2} \cosh^2[(k_1-k_2)(\nu_g+\delta_g)] - 1 \right]^{3/2}} \tag{32}$$

$$= -32k_1k_2 \frac{p_g q_g}{r_g^{3/2}}, \tag{33}$$

as desired. Moreover, g is negative because the quantities p_g, q_g, and r_g are all positive.

(iv) It is now easy to deduce from (32) that

$$g(\nu_g) = \mathrm{O}(\mathrm{sech}^2[(k_1-k_2)(\nu_g+\delta_g)])$$

as $\nu_g \to \pm\infty$.

(v) Using (31), we find that $g(\nu_g)$ has the derivative

$$\frac{dg}{d\nu_g} = 8k_1k_2 \frac{2z^2+1}{(z^2-1)^{3/2}} \left(\frac{dz}{d\nu_g} \right). \tag{34}$$

Since $z^2 - 1 > 0$, it follows that $\frac{dg}{d\nu_g}$ is zero precisely when

$$\frac{dz}{d\nu_g} = \frac{(k_1+k_2)^2}{(k_1-k_2)} \sinh[(k_1-k_2)(\nu_g+\delta_g)] \tag{35}$$

is zero, or equivalently, when $\nu_g = -\delta_g$. We can therefore conclude that g has an absolute minimum of

$$g(-\delta_g) = -\frac{(k_1-k_2)^2(k_1+k_2)}{\sqrt{k_1k_2}}$$

at this critical point because of (iv). This completes the proof of Theorem 3.6. □

Remark. Property (iv) of Theorem 3.6 shows that in some sense g can be viewed as a nonlinear difference between the soliton particles s_1 and s_2 as defined by (11). Moreover, $g(\nu_g) \to 0$ as $k_2 \to k_1$ and $g(\nu_g) \to -4k_1\delta(\nu_g)$ as $k_2 \to 0$, where $\delta(\nu_g)$ is the Dirac delta function.

Next, we show that each decay function itself can be decomposed as a sum of a "soliton" term and a "ghost" term:

$$u_n(x,t) = -2\frac{\partial^2}{\partial x^2} \log(1+\mu_n) \tag{36}$$

$$= -2\left[\frac{(1+\mu_n)\mu_n'' - (\mu_n')^2}{(1+\mu_n)^2}\right] \tag{37}$$

$$= -2\frac{\mu_n''}{(1+\mu_n)^2} - 2\left[\frac{\mu_n\mu_n'' - (\mu_n')^2}{\mu_n^2}\right]\left(\frac{\mu_n}{1+\mu_n}\right)^2 \tag{38}$$

$$= -2\frac{\mu_n''}{(1+\mu_n)^2} - 2\left(\frac{\partial^2}{\partial x^2}\log\mu_n\right)\left(\frac{\mu_n}{1+\mu_n}\right)^2 \tag{39}$$

$$= u_n^s + u_n^g. \tag{40}$$

DEFINITION 3.7. *We shall call*

$$u_n^s = -2\frac{\mu_n''}{(1+\mu_n)^2} \tag{41}$$

the soliton component of u_n *and*

$$u_n^g = -2\left(\frac{\partial^2}{\partial x^2}\log\mu_n\right)\left(\frac{\mu_n}{1+\mu_n}\right)^2 \tag{42}$$

the ghost component of u_n*. Moreover, we shall refer to the decomposition given by* (40) *as the splitting decomposition of* u_n*.*

For two-solitons, the following lemma holds.

LEMMA 3.8.

$$u_n^g(x,t) = (-1)^{n-1}g(x-4k_g^2t)\left(\frac{\mu_n(x,t)}{1+\mu_n(x,t)}\right)^2, \quad n=1,2. \tag{43}$$

Remark. The decomposition described in (40) reveals the time-asymmetry of soliton decay in that ghost particles are born at $t=\infty$ and is essentially due to the identity matrix appearing in the N-soliton formula. In particular, the behavior of $\mu_n/(1+\mu_n)\to 0$ as $t\to-\infty$ and $\mu_n/(1+\mu_n)\to 1$ as $t\to\infty$ in (43) indicates that the ghost component u_n^g represents creation of the ghost particle $g(x-4k_g^2t)$ at $t=\infty$. This implies that there is actually interaction between solitons even before "collision" occurs; however, this interaction is insignificant until then. Lastly, it is straightforward to verify that each soliton component u_n^s asymptotically describes an exchange of identities between the two-soliton particles.

4. Two-soliton example. We end our paper with a concrete example to illustrate our results. Let $k_1=c_1=2$ and $k_2=c_2=1$ be the given scattering data. Our soliton matrix A then takes the form

$$A = \begin{pmatrix} e^{4x-64t} & \frac{2}{3}e^{3x-36t} \\ \frac{2}{3}e^{3x-36t} & \frac{1}{2}e^{2x-8t} \end{pmatrix} \tag{44}$$

and has eigenvalues

$$\mu_1 = \frac{1}{12}\left(3e^{2x-8t} + 6e^{4x-64t} + e^{2x-8t}\sqrt{9+28e^{2x-56t}+36e^{4x-112t}}\right), \tag{45}$$

$$\mu_2 = \frac{1}{12}\left(3e^{2x-8t} + 6e^{4x-64t} - e^{2x-8t}\sqrt{9+28e^{2x-56t}+36e^{4x-112t}}\right). \tag{46}$$

The decay functions u_1 and u_2 can now of course be computed through the formula

$$u_n = -2\frac{\partial^2}{\partial x^2}\log(1+\mu_n), \quad n=1,2,$$

but we shall avoid doing this here due to their complicated expressions. The ghost matrix

$$A_g = \begin{pmatrix} e^{4\nu_g} & \frac{2}{3}e^{3\nu_g} \\ \frac{2}{3}e^{3\nu_g} & \frac{1}{2}e^{2\nu_g} \end{pmatrix} \tag{47}$$

has eigenvalues

$$\gamma_1 = \frac{1}{12}\left(3e^{2\nu_g} + 6e^{4\nu_g} + e^{2\nu_g}\sqrt{9 + 28e^{2\nu_g} + 36e^{4\nu_g}}\right), \tag{48}$$

$$\gamma_2 = \frac{1}{12}\left(3e^{2\nu_g} + 6e^{4\nu_g} - e^{2\nu_g}\sqrt{9 + 28e^{2\nu_g} + 36e^{4\nu_g}}\right). \tag{49}$$

Therefore,

$$g = -32k_1k_2\left(\frac{p_g q_g}{r_g^{3/2}}\right) \tag{50}$$

$$= -\frac{384e^{2\nu_g}(1 + 2e^{2\nu_g})}{(9 + 28e^{2\nu_g} + 36e^{4\nu_g})^{3/2}} \tag{51}$$

$$= -\frac{48\cosh(\nu_g + \log\sqrt{2})}{\left[9\cosh^2(\nu_g + \log\sqrt{2}) - 1\right]^{3/2}} \tag{52}$$

and the ghost moving frame is given by $\nu_g = x - 28t$. Of course, we also have $\bar{g} = -g$. Figures 2–4 illustrate the motions of $-u(x,t)$, $-u_1(x,t)$, and $-u_2(x,t)$, respectively, over time through a sequence of six frames corresponding to time steps $t = -0.4, -0.2, \ldots, 0.6$. The soliton particles s_1 and s_2 have amplitudes of 8 and 2, respectively, and velocities of 16 and 4, respectively. The ghost particle g has an amplitude of $3/\sqrt{2} \approx 2.12$ and a velocity of 28. Splitting occurs in the fourth frame at $t = 0.2$ for both u_1 and u_2 as seen in Figures 3 and 4, respectively.

5. Concluding remarks. Our work raises interesting questions some of which deserve comment.

Q1. What happens during collisions of more than two-solitons? Are more ghost particles generated? Can ghost particles from different pairs interact?

A1. It is found that each collision between any two-soliton particles produces a ghost particle pair with the same properties as those described by Theorem 3.6. On the other hand, each collision between two ghost particles where each comes from a different pair will result in their fusion. Because of duality, there is an accompanying fission process which is interpreted as the same fusion process but reversed in time. Moreover, the final states of all ghost particles created is independent of their order of collision (modulo phase shifts). A mathematical theory formulating the creation and interaction of ghost particles will be described in a forthcoming paper.

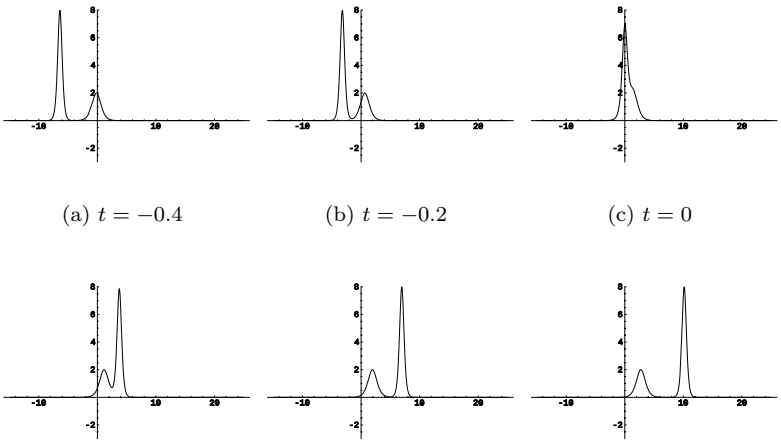
Q2. Do the decay functions $\{u_n\}$ satisfy any partial differential equations?

A2. This is not presently known as we have been unsuccessful at finding such equations. On the other hand, it is known that the eigenvalues $\{\mu_n\}$ of the soliton matrix A which defines $\{u_n\}$ satisfy ordinary differential equations of the form

$$\frac{d\mu_n}{dx} = (E^T \cdot X_n)^2, \quad n = 1, \ldots, N. \tag{53}$$

Here, X_n is a normalized eigenvector of A corresponding to μ_n and

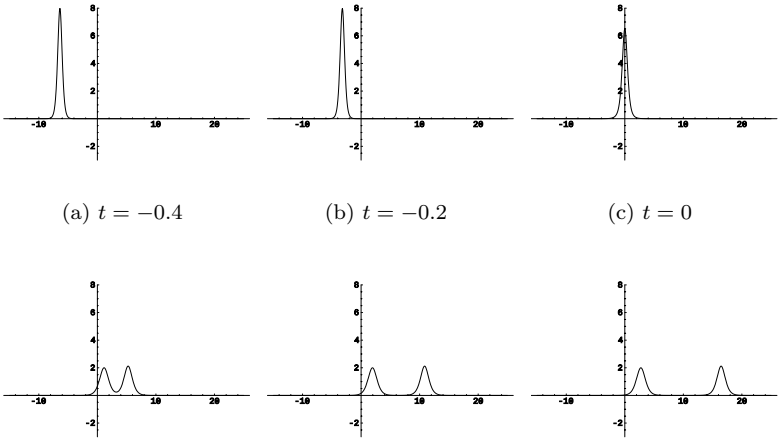
$$E^T = (c_1 e^{k_1\nu_1}, c_2 e^{k_2\nu_2}, \ldots, c_N e^{k_1\nu_1})^T.$$



(a) $t = -0.4$ (b) $t = -0.2$ (c) $t = 0$

(d) $t = 0.2$ (e) $t = 0.4$ (f) $t = 0.6$

FIG. 2. *Plots of* $-u(x,t)$.



(a) $t = -0.4$ (b) $t = -0.2$ (c) $t = 0$

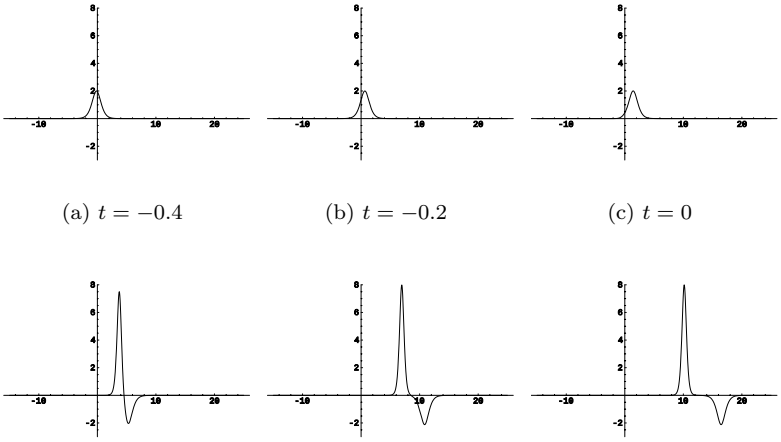
(d) $t = 0.2$ (e) $t = 0.4$ (f) $t = 0.6$

FIG. 3. *Plots of* $-u_1(x,t)$*:* $s_1 \to s_2 + g$.

These differential equations can be easily derived from the symmetry and positive definiteness of A. However, their usefulness is unclear as they do not make direct use of the KdV equation.

Q3. How is the linear eigenvalue decomposition described in this paper related to others in the literature, e.g., Hodnett and Moloney [HM] and Miller and Christiansen [MC]?

A3. Hodnett and Moloney's work in [HM] involves using the Hirota formalism to decompose each N-soliton solution into a linear sum of squares of hyperbolic secant functions having time-dependent amplitudes and phase shifts (a Lie-theoretic gener-



(a) $t = -0.4$ (b) $t = -0.2$ (c) $t = 0$

(d) $t = 0.2$ (e) $t = 0.4$ (f) $t = 0.6$

FIG. 4. *Plots of* $-u_2(x,t)$*:* $s_2 \to s_1 + \bar{g}$.

alization of this decomposition is given by Fuchssteiner in [F]). For two-solitons, this decomposition takes the form

$$u = u_1 + u_2, \tag{54}$$

where

$$u_1 = 2a_1^2 H(\theta_2)\operatorname{sech}^2[\theta_1 + G(\theta_2)], \tag{55}$$

$$u_2 = 2a_2^2 H(\theta_1)\operatorname{sech}^2[\theta_2 + G(\theta_1)]. \tag{56}$$

Here, a_i and θ_i are the spectral parameters and moving frames, respectively. Exact formulas for $H(\nu_1)$ and $G(\nu_2)$ can then be derived by requiring u_1 and u_2 to conserve mass for all times as in Theorem 3.4. In essence, this approach by Hodnett and Moloney views the hyperbolic secant function as the building block for a soliton particle whereas our approach views the eigenvalues of the soliton matrix A as playing this role. As a result, the decomposition of Hodnett and Moloney seems to asymptotically describe only an exchange of soliton identities and not soliton decay as revealed by our decomposition.

As for Miller and Christiansen [MC], they considered soliton solutions of the coupled system

$$\frac{\partial u_k}{\partial t} + \frac{\partial}{\partial x}\left[\frac{u_k}{2}\sum_{j=1}^{N} u_j + \frac{\partial^2 u_k}{\partial x^2}\right] = 0, \qquad k = 1, \ldots, N. \tag{57}$$

This system can be viewed as a multicomponent generalization of the KdV equation and is derived by requiring symmetry and conservation of mass principles. For $N = 2$, numerical solutions for u_1 and u_2 were obtained which indicated an exchange of mass between two given soliton particles after collision. However, there is no prediction of ghost particles, which again is in contrast to our decomposition. In short, we believe our model of soliton interaction to be one that is most consistent with the laws of classical mechanics.

Acknowledgments. The author wishes to sincerely thank Eduardo Flores from Rowan University for his helpful comments and clear explanations of the physical concepts discussed in this paper. The author also wishes to thank the referee for making useful suggestions about this paper and raising the questions addressed above.

REFERENCES

[BS] G. Bowtell and A. E. G. Stuart, *A particle representation for Korteweg-de Vries solitons*, J. Math. Phys., 24 (1983), pp. 969–981.

[F] B. Fuchssteiner, *The interaction equation*, Phys. A, 288 (1996), pp. 189–211.

[GGKM] C. S. Gardner, J. M. Greene, M. D. Kruskal, and M. R. Miura, *Korteweg-deVries equation and generalizations.* VI. *Methods for exact solution*, Comm. Pure Appl. Math., 27 (1974), pp. 97–133.

[H] R. Hirota, *Exact solutions of the Korteweg-deVries equation for multiple collisions of solitons*, Phys. Rev. Lett., 27 (1971), pp. 1192–1194.

[HM] P. F. Hodnett and T. P. Moloney, *On the structure during interaction of the two-soliton solution of the Korteweg–de Vries equation*, SIAM J. Appl. Math, 49 (1989), pp. 1174–1187.

[KM] I. Kay and H. E. Moses, *Reflectionless transmission through dielectrics and scattering potentials*, J. Applied Physics., 27 (1956), pp. 1503–1508.

[KV] D. J. Korteweg and G. de Vries, *On the change of long waves advancing in a rectangular canal, and on a new type of long stationary waves*, Phil. Mag., 39 (1895), pp. 422–443.

[La] P. D. Lax, *Integrals of nonlinear equations of evolution and solitary waves*, Comm. Pure Appl. Math., 21 (1968), pp. 467–490.

[Le] R. J. LeVeque, *On the interaction of nearly equal solitons in the KdV equation*, SIAM J. Appl. Math, 47 (1987), pp. 254–262.

[MC] P. D. Miller and P. L. Christiansen, *A coupled Korteweg-de Vries system and mass exchanges among solitons*, Phys. Scripta, 61 (2000), pp. 518–525.

[M] R. M. Miura, *The Korteweg–deVries equation: A survey of results*, SIAM Rev., 18 (1976), pp. 412–459.

[S] A. Shabat, *The infinite-dimensional dressing dynamical system*, Inverse Problems, 8 (1992), pp. 303–308.

[WT] M. Wadati and M. Toda, *The exact N-soliton solution of the Korteweg-deVries equation*, J. Phys. Soc. Japan, 32 (1972), pp. 1403–1411.

[Z] N. J. Zabusky, *A synergetic approach to problems of nonlinear dispersive wave propagation and interaction*, in Proceedings of the Symposium on Nonlinear Partial Differential Equations, W. F. Ames, ed., Academic Press, New York, 1967, pp. 223–258.

[ZK] N. J. Zabusky and M. D. Kruskal, *Interaction of "solitons" in a collisionless plasma and the recurrence of initial states*, Phys. Rev. Lett., 15 (1965), pp. 240–243.

SIAM J. APPL. MATH.
Vol. 63, No. 3, pp. 889–904

THE BIFURCATION STRUCTURE OF THE HOLLING–TANNER MODEL FOR PREDATOR-PREY INTERACTIONS USING TWO-TIMING*

PETER A. BRAZA†

Abstract. The Holling–Tanner model for predator-prey systems has two Hopf bifurcation points for certain parameters. The dependence of the environmental parameters on the underlying bifurcation structure is uncovered using two-timing. Emphasis is on how the bifurcation diagram changes as the Hopf bifurcation points separate. Two degenerate cases require a modification of conventional two-timing. When the two Hopf bifurcation points nearly coalesce, the two stable periodic solution branches are shown to be connected. As a ratio of linear growth rates varies, the Hopf bifurcation points separate further and one limit cycle becomes unstable. This situation can correspond to an outbreak in populations. The modified two-timing analysis analytically captures the unstable and stable limit cycles of the new branch.

Key words. Holling–Tanner, predator-prey, degenerate Hopf bifurcations, limit cycles, two-timing, outbreaks

AMS subject classifications. 34C25, 37G15, 92D25

PII. S0036139901393494

1. Introduction.

Predator-prey dynamics continue to be of interest to both applied mathematicians and ecologists. The early Lotka–Volterra model has given way to more sophisticated models from both a mathematical and biological point of view. Although useful for developing some basic intuition about predator-prey systems, the Lotka–Volterra model has the ecologically offensive property of a neutrally stable equilibrium point giving rise to a family of periodic solutions in which initial conditions, rather than environmental parameters, determine long term behavior [1], [2], [3]. As far back as Kolmogorov [4] in 1936, it was suggested that a stable equilibrium or a stable limit cycle should be the result of a wide class of single predator, single prey mathematical models [1]. From an ecological perspective, the fairly regular oscillations seen with the snowshoe hare and lynx in Canada support the isolated periodic solution conclusion (limit cycle) rather then the family of periodic solutions [5].

Robert May developed a model in which he incorporated Holling's rate [6], [7] at which predators remove prey and Leslie's somewhat unconventional equation for predator dynamics [1], [8]. This model, also known as the Holling–Tanner model [9], has been studied both for its mathematical properties and its efficacy for describing real ecological systems such as mite/spider mite, lynx/hare, sparrow/sparrow hawk, etc. by Tanner [10] and Wollkind, Collings, and Logan [11].

The May or Holling–Tanner model for predator-prey interaction is

$$\begin{aligned} \frac{dx}{dt} &= r_1 x\left(1-\frac{x}{k}\right)-\frac{qxy}{x+a}, \\ \frac{dy}{dt} &= r_2 y\left(1-\frac{y}{\gamma x}\right) \end{aligned} \tag{1.1}$$

*Received by the editors December 6, 2001; accepted for publication (in revised form) June 13, 2002; published electronically January 23, 2003.
http://www.siam.org/journals/siap/63-3/39349.html
†Department of Mathematics and Statistics, University of Northern Florida, 4567 St. John's Bluff Road N., Jacksonville, FL 32224-2666 (pbraza@unf.edu).

[1], [9], [10], [11].

The variables $x(t)$ and $y(t)$ denote the prey and predator, respectively. The parameters r_1 and r_2 are the intrinsic growth rates. The value k is the carrying capacity of the prey and γx takes on the role of a prey-dependent carrying capacity for the predator. The parameter γ is a measure of the quality of the prey as food for the predator. The rate at which predators remove the prey, $qx/(x+a)$, is known as a Holling type 2 predator response [1], [6], [9], [12]. The parameter q is the maximum number of prey that can be eaten per predator per time and the parameter a is a saturation value; it corresponds to the number of prey necessary to achieve one half the maximum rate q.

The dynamics of the Holling–Tanner model have proven quite interesting. Thus far the mathematical analysis of the dynamics has generally been either qualitative or numerical. May and Tanner used phase plane stability analysis to find criteria for stable limit cycles [1], [2], [6]. Hsu and Huang [12] discuss global stability questions of the equilibrium point. Sáez and Gonzáles-Olivares [9] develop curves in parameter space that delineate regions of two limit cycles, semistable limit cycles, etc. and discuss bifurcations from an existence viewpoint. Numerical work with AUTO [13] was used in two papers by Collings [14] and Wollkind, Collings, and Logan [11] to illustrate the bifurcations in mite/spider mite systems.

This body of work spurred this author to try to understand the dynamics from a complementary perspective. The point in this paper is to bridge the results on the existence of limit cycles and the numerical work by using techniques to *analytically construct* solutions, applicable to a broad class of predator-prey problems, that describe the underlying bifurcations. A focus is to understand the bifurcation diagram as the two Hopf bifurcation points separate. Two-timing methods will be used to construct the periodic solutions that arise from the Hopf bifurcation points in terms of environmental parameters. Of particular interest will be the ways the two-timing method must be modified to handle the degenerate cases discussed by the authors above. The two branches of periodic solutions are shown to be connected when the Hopf bifurcation points are close. Also, the entire subcritical branch is captured analytically along with conditions for having a semistable limit cycle. The subcritical case is important ecologically since it can correspond to large outbreaks in the populations.

2. Analysis. At the beginning of this section, linear stability results give conditions on the parameters when there will be two Hopf bifurcation points. Then the periodic solutions bifurcating from the two Hopf bifurcation points are found using two-timing. Based on the slow time amplitude equations, it will be shown that one branch is always supercritical while the other one changes from supercritical to subcritical as the parameters are varied.

As a first step in the bifurcation analysis, equations (1.1) are nondimensionalized by defining new predator and prey variables as

$$prey\ \ U = \frac{x}{k}, \qquad predator\ \ V = \frac{y}{\gamma k},$$

along with a new time

$$t_{new} = r_1 t_{old},$$

and parameters

$$r = \frac{r_2}{r_1}, \qquad b = \frac{a}{k}, \qquad c = \frac{q\gamma}{r_1}.$$

Since k is the carrying capacity of the prey and the parameter a is the saturation value of the prey, it only makes sense biologically to consider $b < 1$ [11]. The parameter $r > 0$ is the ratio of linear growth rates of predator to prey and will be an important parameter in the bifurcation analysis to follow. Tanner [10] has field data with estimated intrinsic growth rates for various predator-prey pairs and he notes that cyclic population dynamics only occur when r is small enough relative to other parameters.

With the nondimensional variables and parameters, equations (1.1) become

$$\begin{aligned} \frac{dU}{dt} &= U(1-U) - \frac{cUV}{U+b}, \\ \frac{dV}{dt} &= rV\left(1 - \frac{V}{U}\right). \end{aligned} \tag{2.1}$$

There are two equilibrium solutions. One equilibrium point, given by $U = 1$, $V = 0$, is not of interest since it corresponds to the prey being at its carrying capacity with no predators. In the (U, V) phase plane it is a saddle. The other equilibrium point depends on the parameters b and c and is given implicitly by

$$U^2 + (b + c - 1)U - b = 0 \quad \text{and} \quad V = U. \tag{2.2}$$

A necessary condition for a Hopf bifurcation is that the trace of system (2.1) linearized about the steady state (2.2) be equal to zero. This condition is represented by the equation

$$r = 1 - 2U - \frac{bcU}{(b+U)^2}. \tag{2.3}$$

Of interest is the fact the Hopf bifurcation equation (2.3) depends on the growth rate parameter r whereas the steady state equation (2.2) does not. This will be utilized later when the bifurcations are unfolded.

When (2.2) and (2.3) are solved simultaneously, we get the values of $U = u$ and c where Hopf bifurcations occur:

$$\begin{aligned} u &= \frac{1}{4}\left(1 - b - r \pm \sqrt{(1-b-r)^2 - 8br}\right) \quad \text{and} \\ c &= \frac{1}{4r}\left(2 + r + r^2 - b\,(b+3r) \mp (2+r)\sqrt{(1-b-r)^2 - 8br}\right). \end{aligned} \tag{2.4}$$

The feature to focus on is that there will be two Hopf bifurcation points when $(1-b-r)^2 - 8br > 0$ and none when $(1-b-r)^2 - 8br < 0$. These regions in the (r, b) plane are shown in Figure 1. The degenerate case when $0 < (1-b-r)^2 - 8br \ll 1$ (nearly coalescing Hopf bifurcation points) will be considered in section 3.

The region of Hopf bifurcation points in Figure 1 is in accord with cyclic behavior seen in real ecological systems. May [1] observed that stable limit cycles are likely with weak self-limitation of the prey (large k) and $r_2/r_1 < 1$. The dynamics switch from an attractive fixed point to a stable limit cycle as "life gets better," meaning r_1 or k increases. This corresponds to a smaller r or b. He argues that the work of Baltensweiler [15] on the eight year cyclic behavior of the larch bud moth in Switzerland, which has been observed since 1855, elegantly illustrates this. Tanner [10] and May [1] suggested that the oscillating population of snowshoe hare in subarctic regions as studied by Dolbeer [16] may be due to poor cover since more southerly

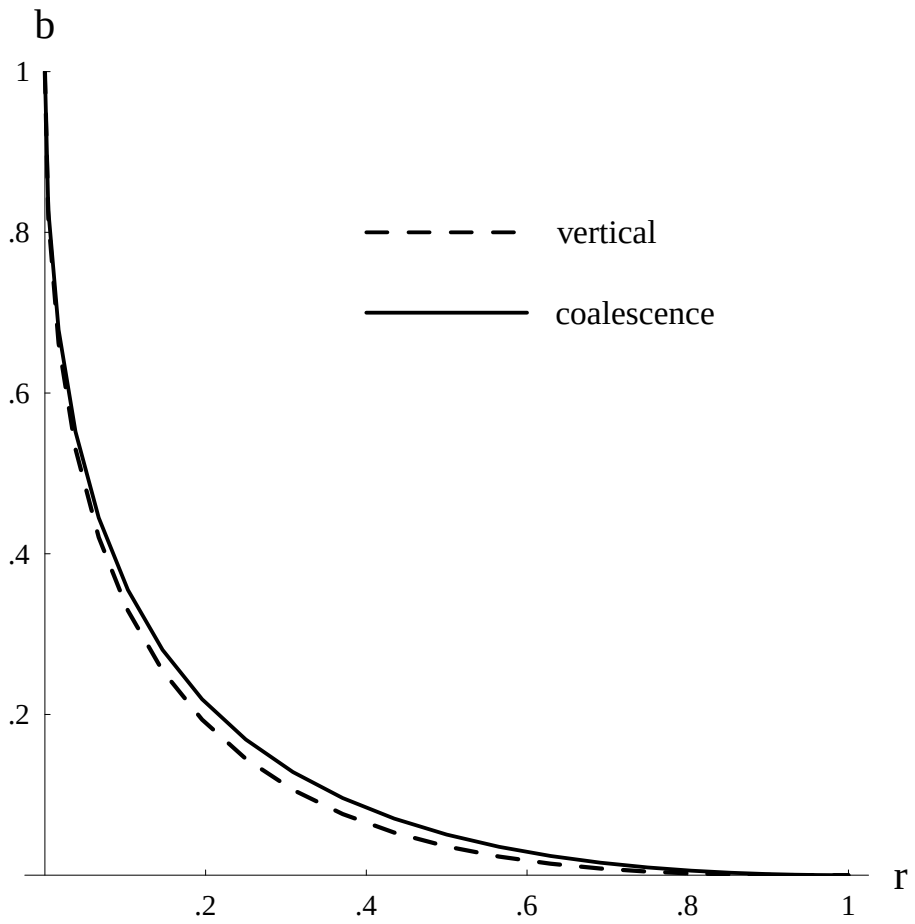


FIG. 1. *The Hopf bifurcation points coalesce for parameters r and b on the solid curve. There are no Hopf bifurcation points in the region above this curve and two Hopf points below it. The dashed curve shows the values of r and b for which the branch of periodic solutions from the right Hopf bifurcation point is vertical.*

populations of snowshoe hare are not cyclic but have more protective cover in the spruce-fir forests of Colorado. Less protective environments mean a smaller value of the parameter $b = a/k$ since a is directly proportional to the time for the predator to search and find the prey [10]. In the same vein, Tanner argues that laboratory experiments by Luckinbill [17] of Paramecium aurelia and its predator Didinium nasutum demonstrate the dependence of population dynamics on the parameter a. He writes, "Cultures of these two species normally exhibited increasing oscillation until one or the other become extinct; an increase in the viscosity of the medium (higher a) strengthened the system's stability, possibly by increasing the predator's searching time." On a different note, Tanner [10] observed that the intrinsic growth rate of the North American mountain lion is larger than its prey, the mule deer (implying $r = r_2/r_1 > 1$), and that these populations are generally near some equilibrium except for natural stochastic variations.

The interrelated influence of ecological parameters on the behavior of a predator-prey system can be complicated to discern. Two-timing will help reveal the depen-

dence of the solutions on the parameters in a fairly explicit way.

In standard two-timing analysis, the bifurcating periodic solutions are found by expanding the variables and bifurcation parameter about the Hopf bifurcation values (2.4). The choice of $c = q\gamma/r_1$ as the primary bifurcation parameter stems from it being a combination of three ecological parameters. Rather than use the values of u and c in (2.4), the analysis will be algebraically simpler if we use the equivalent expressions

$$c_h = \frac{(1-u)^2}{r+u} \quad \text{and} \quad b_h = \frac{u(1-r-2u)}{r+u}, \tag{2.5}$$

also found by solving the steady state (2.2) and Hopf bifurcation (2.3) equations simultaneously, and consider u and r as free parameters. An added benefit is that the local analysis about both Hopf bifurcation points can be done simultaneously by using (2.5) rather than separate analyses using (2.4).

To find the periodic solutions locally about the Hopf bifurcation points, we begin by defining a small parameter ε as a measure of the deviation from the Hopf value of c by

$$c = c_h + P\,\varepsilon^2 \tag{2.6}$$

with $P = \pm 1$ determining the direction of the bifurcation. A slow time is defined by

$$\tau = \varepsilon^2 t, \tag{2.7}$$

and the variables are expanded in a power series in ε as

$$\begin{aligned} U(t,\tau) &= u + \varepsilon u_1(t,\tau) + \varepsilon^2 u_2(t,\tau) + \cdots, \\ V(t,\tau) &= u + \varepsilon v_1(t,\tau) + \varepsilon^2 v_2(t,\tau) + \cdots. \end{aligned} \tag{2.8}$$

When the variables in (2.8), slow time τ, and expressions (2.5) and (2.6) are substituted into the main equations (2.1), we get a sequence of equations at each order of ε.

The $O(\varepsilon)$ equations are

$$\begin{bmatrix} u_{1t} \\ v_{1t} \end{bmatrix} = \begin{bmatrix} r & u-1 \\ r & -r \end{bmatrix} \begin{bmatrix} u_1 \\ v_1 \end{bmatrix}. \tag{2.9}$$

The periodic solutions, in polar form, are

$$\begin{bmatrix} u_1(t,\tau) \\ v_1(t,\tau) \end{bmatrix} = R(\tau)e^{i\theta(\tau)} \begin{bmatrix} r+i\lambda \\ r \end{bmatrix} e^{i\lambda t} + R(\tau)e^{-i\theta(\tau)} \begin{bmatrix} r-i\lambda \\ r \end{bmatrix} e^{-i\lambda t}. \tag{2.10}$$

The fast time frequency is $\lambda = \sqrt{r(1-u-r)}$. (It is assumed from here on that $1-u-r > 0$.) The amplitude is proportional to $R(\tau)$, which depends on the slow time.

After the $O(\varepsilon^2)$ equations are solved for $u_2(t,\tau)$ and $v_2(t,\tau)$, the differential equations for $R(\tau)$ and $\theta(\tau)$ are obtained by considering the $O(\varepsilon^3)$ equations which are of the form

$$\begin{bmatrix} u_{3t} \\ v_{3t} \end{bmatrix} - \begin{bmatrix} r & u-1 \\ r & -r \end{bmatrix} \begin{bmatrix} u_3 \\ v_3 \end{bmatrix} = \begin{bmatrix} f_3(u_1,u_2,v_1,v_2) - u_{1\tau} \\ g_3(u_1,u_2,v_1,v_2) - v_{1\tau} \end{bmatrix}. \tag{2.11}$$

For periodic solutions to exist, the right-hand side of (2.11) must satisfy a solvability condition:

$$\int_0^{2\pi/\lambda} \begin{bmatrix} f_3 - u_{1\tau} \\ g_3 - v_{1\tau} \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} dt = 0, \tag{2.12}$$

where $\left[\begin{smallmatrix} x \\ y \end{smallmatrix}\right]$ is any solution of the adjoint homogeneous problem of (2.9),

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} -r & -r \\ 1-u & r \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}. \tag{2.13}$$

The resulting differential equation for $R(\tau)$ is

$$R_\tau = \alpha(r,u)PR + \beta(r,u)R^3. \tag{2.14}$$

The coefficients $\alpha(r,u)$ and $\beta(r,u)$ are given by

$$\begin{aligned} \alpha(r,u) &= \frac{(r+u)(r(r-1)+4ru+2u^2)}{2(1-u)^2(1-r-u)}, \\ \beta(r,u) &= \frac{r(1-r-2u)((1-r)^2r + r(4r-7)u + (5r-4)u^2 + 2u^3)}{2u^2(1-u)(1-r-u)}. \end{aligned} \tag{2.15}$$

The algebra was quite messy so Mathematica [18] was enlisted to help out. There is also an equation for the slow time phase $\theta(\tau)$ but it is not relevant for the purposes here.

The equation (2.14) is the typical cubic form for a Hopf bifurcation. There are two steady states:

$$R = 0 \quad \text{and} \quad R = \sqrt{\frac{-\alpha(r,u)P}{\beta(r,u)}}. \tag{2.16}$$

The value $R = 0$ corresponds to the steady state of (2.1) whereas $R = \sqrt{\frac{-\alpha(r,u)P}{\beta(r,u)}}$ is the amplitude of the bifurcating periodic solutions. It is important to note that there are two values of u from (2.4) which give the nontrivial amplitude, so (2.16) captures the amplitude of the bifurcating periodic solutions from *both* Hopf bifurcation points. The stability of the bifurcating periodic solutions is determined by the signs of $\alpha(r,u)$ and $\beta(r,u)$. Figure 2 delineates regions in the (r,u) plane showing the signs of $\alpha(r,u)$ and $\beta(r,u)$ and the consequent stability.

When there are two Hopf bifurcation points, one arises from the parameters r and u of region A in Figure 2 and the other one comes from region B_1 or B_2. Since $\alpha(r,u) > 0$, $\beta(r,u) < 0$, and $P = 1$ in region A, the bifurcating periodic solution is supercritical (stable). The signs of $\alpha(r,u)$ and $\beta(r,u)$ also determine that the other branch will be stable if it comes from B_1 but that a branch from B_2 will be locally unstable (subcritical). The analysis in section 4 shows that this latter branch becomes stable. In all cases, the branch from either B_1 or B_2 joins the branch from region A. In general, the use of AUTO is required to establish this connection, but the connection of the periodic solution branches from regions A and B_1 is constructed analytically in section 3 in the case of nearly coalescing Hopf bifurcation points. The construction reveals more about the underlying dependence of the bifurcation diagram on the parameters than previous work in which other authors used AUTO exclusively [11], [14].

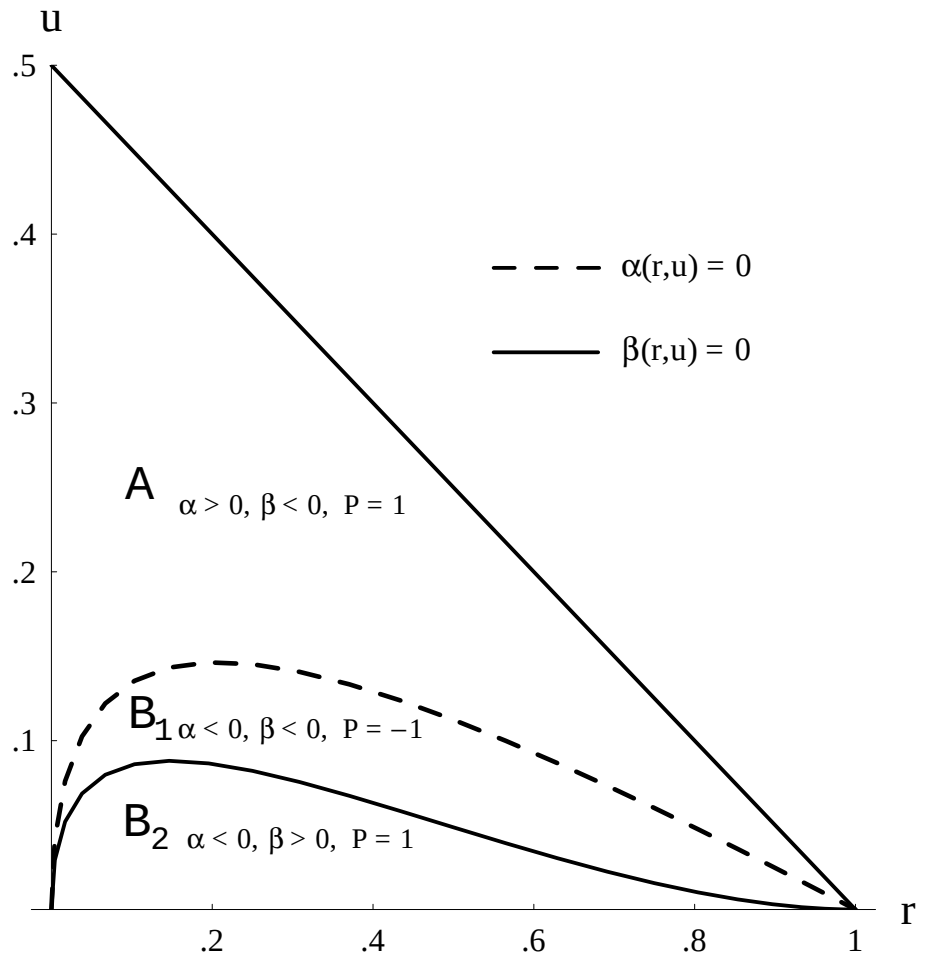


FIG. 2. *The signs of the coefficients $\alpha(r,u)$ and $\beta(r,u)$ of the amplitude equation* (2.15) *are indicated in the three regions of the (r,u) plane. The leftmost Hopf bifurcation point comes from the parameters in region A and the rightmost Hopf point comes from region B_1 or B_2. The resulting branches of periodic solutions are stable from regions A or B_1 and unstable from region B_2.*

In Figure 3, a sequence of pictures was created using AUTO to illustrate how the bifurcation diagrams change as the Hopf bifurcation points separate. The separation of the Hopf bifurcation points increases as ratio of linear growth rate parameter r decreases. In Figure 3, the left Hopf point comes from region A of Figure 2 whereas the right Hopf point comes from region B_1 in Figure 3(a), region B_2 in Figure 3(c), or on the border curve $\beta(r,u) = 0$ in Figure 3(b). The branch of periodic solutions from the left Hopf point is always supercritical but the branch from the right Hopf point is supercritical in Figure 3(a), "vertical" in Figure 3(b), and subcritical in Figure 3(c). The value of R from (2.16) accurately captures the periodic solutions locally. Similar bifurcation diagrams have been found by using AUTO in papers by Collings [14] and Wollkind, Collings, and Logan [11] in regards to mite/spider mite interactions with temperature dependence.

A nice feature about the amplitude equation (2.14) is that the parameters u and r

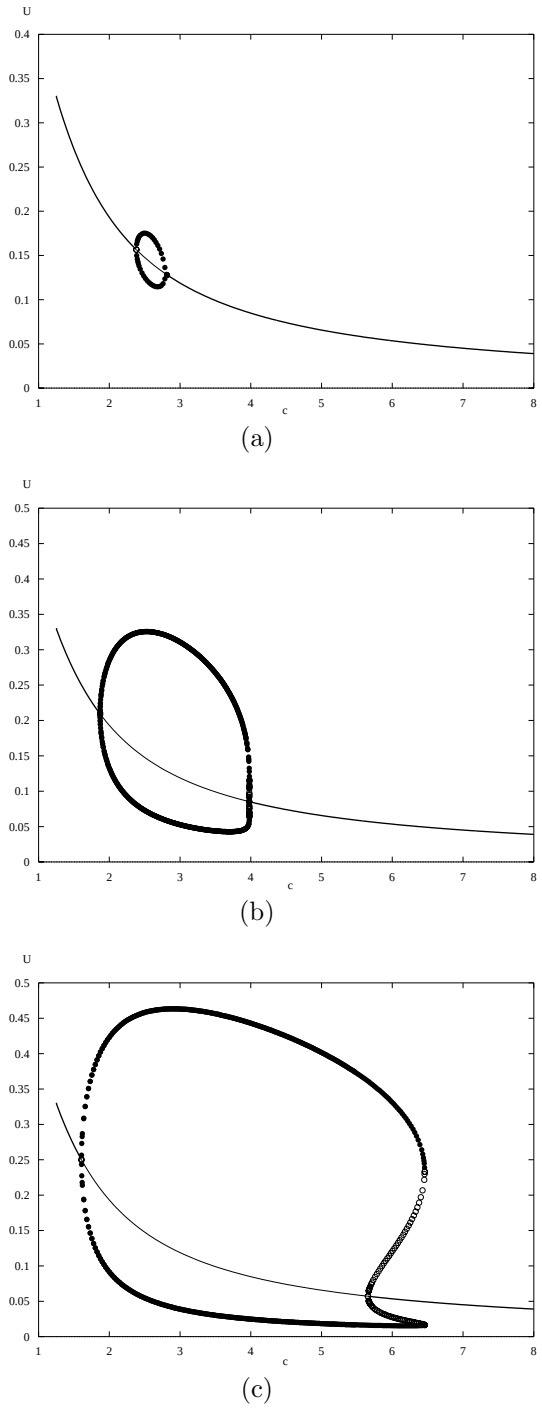


FIG. 3. (a)–(c) *show the bifurcation diagrams (computed with AUTO) of equations* (2.1) *as the Hopf bifurcation points separate. The light line is the unstable steady state and the slightly darker line is the stable steady state. The open circles denote unstable periodic solutions and the solid circles (often appearing as a thick solid curve) are the stable periodic solutions. The parameter b is fixed at* $b = 2/7$ *but* $r = .142$ *in Figure* 3(a), $r = .125$ *in Figure* 3(b), *and* $r = .10$ *in Figure* 3(c). *The coalescence value is* $r = 1/7$.

are not fixed so degeneracies can be readily identified by setting either coefficient $\alpha(r,u)$ or $\beta(r,u)$ of (2.15) equal to zero. The case $\alpha(r,u)=0$, which is equivalent to $(1-b-r)^2-8br=0$, corresponds to the two Hopf bifurcation points coalescing and will be analyzed in detail in the next section. The vertical Hopf bifurcation case occurring when $\beta(r,u)=0$ will be considered in section 4.

3. Nearly coalescing Hopf bifurcation points. Two-timing results generally give only local information about bifurcations. It would be nice to extend the results to show that the periodic solution branches from the two Hopf bifurcation points are connected. This can be done in the case of nearly coalescing Hopf points if the two-timing analysis is appropriately modified. Golubitsky and Langford [19] give the universal unfolding normal form for this case as well as many other degenerate Hopf bifurcations, but that work is not directly of use here.

From (2.4) the Hopf points will coalesce if

$$(1-b-r)^2-8br=0. \tag{3.1}$$

We note from (2.4) and Figure 3 that for a fixed b, as r decreases, the Hopf bifurcation points separate further and the amplitude of the periodic solution increases. This eventually can lead to a destabilization of the system since reductions in minimum population densities can increase the chance of extinction as a result of demographic or environmental stochastic influence [1].

If (3.1) holds, then the linear coefficient, $\alpha(r,u)$, in the amplitude equation (2.14) is zero so the steady state amplitude, $R=\sqrt{\frac{-\alpha(r,u)P}{\beta(r,u)}}$ from (2.16), becomes zero, causing the standard two-timing to break down. To rectify this situation, we begin by splitting the Hopf bifurcation points by defining the small parameter ε in terms of the growth rate parameter by

$$r=\rho-\varepsilon^2 \tag{3.2}$$

and fixing b at the coalescence value

$$b_{coal}=1+3\rho-2\sqrt{2\rho}\sqrt{1+\rho} \tag{3.3}$$

obtained from (3.1) with $\varepsilon=0$. Please note that $\varepsilon=\varepsilon(r)=\sqrt{\rho-r}$ is a measure of how far the growth rate parameter r is from its coalescence value ρ. It is not defined in the conventional way as a deviation from the Hopf value of c as in (2.6). The parameter c will be defined by

$$c=c_{coal}+\varepsilon c_1 \quad \text{with} \quad c_{coal}=-2-2\rho+(2+3\rho)\sqrt{\frac{1+\rho}{2\rho}} \tag{3.4}$$

being the coalescence value of c found by using $r=\rho$ and b_{coal} in expression (2.4). As will be seen, the amplitude and period of the periodic solution will depend on c_1, which measures how far c is from c_{coal}. The parameter c_1 is allowed to vary from 0 to the two Hopf bifurcation values

$$\pm\frac{2+\rho}{2\rho}\sqrt{2+4\rho-3\sqrt{2\rho(1+\rho)}}. \tag{3.5}$$

(These values are the $O(\varepsilon)$ terms in the expansion of c in (2.4) using $r=\rho-\varepsilon^2$ and b_{coal}.) The slow time $\tau=\varepsilon^2 t$ is defined as before. What is not apparent at this

stage is that a slow time frequency shift must be introduced at $O(\varepsilon)$ to account for the near coalescence degeneracy, so a new fast time is introduced as

$$T = (1 + \omega_1 \varepsilon + \cdots)t. \tag{3.6}$$

The variables U and V are expanded about their steady state values, u_s, found by using c from (3.4) and b_{coal} and solving the steady state equation (2.2). Thus we let

$$\begin{aligned} U(T,\tau) &= u_s + \varepsilon u_1(T,\tau) + \varepsilon^2 u_2(T,\tau) + \cdots, \\ V(T,\tau) &= u_s + \varepsilon v_1(T,\tau) + \varepsilon^2 v_2(T,\tau) + \cdots, \end{aligned} \tag{3.7}$$

where

$$u_s = \left(\sqrt{\frac{\rho(1+\rho)}{2}} - \rho\right) - \frac{c_1\rho}{2+\rho}\varepsilon + \frac{2\rho(2\rho + \sqrt{2\rho(1+\rho)}\,)c_1^2}{(2+\rho)^3(1-\rho)}\varepsilon^2 + O(\varepsilon^3). \tag{3.8}$$

The analysis proceeds by substituting the expressions in the above paragraph into the main equations (2.1) to get a sequence of equations at each order of ε. The $O(\varepsilon)$ equations are

$$\begin{bmatrix} u_{1T} \\ v_{1T} \end{bmatrix} = \begin{bmatrix} \rho & \sqrt{\frac{\rho(1+\rho)}{2}} - \rho - 1 \\ \rho & -\rho \end{bmatrix} \begin{bmatrix} u_1 \\ v_1 \end{bmatrix}. \tag{3.9}$$

The solutions are given by

$$\begin{bmatrix} u_1(T,\tau) \\ v_1(T,\tau) \end{bmatrix} = R(\tau)e^{i\theta(\tau)} \begin{bmatrix} \rho + i\lambda \\ \rho \end{bmatrix} e^{i\lambda T} + R(\tau)e^{-i\theta(\tau)} \begin{bmatrix} \rho - i\lambda \\ \rho \end{bmatrix} e^{-i\lambda T} \tag{3.10}$$

and the fast time frequency is $\lambda = \sqrt{\rho(2 - \sqrt{2\rho(1+\rho)}\,)/2}$.

There are secular terms in the $O(\varepsilon^2)$ equations. This impediment is removed by using a solvability condition similar to (2.12) giving the frequency correction that depends on c_1,

$$\omega_1 = \frac{c_1\rho}{\lambda(2+\rho)(2 - \sqrt{2\rho(1+\rho)}\,)}. \tag{3.11}$$

The solvability condition in this case requires that the right-hand side of the $O(\varepsilon^2)$ equations be orthogonal to the homogeneous solutions of the adjoint problem of (3.9).

As in the general case in the previous section, the equations for the slow time amplitude $R(\tau)$ and phase $\theta(\tau)$ are obtained from the $O(\varepsilon^3)$ equations:

$$\begin{bmatrix} u_{3T} \\ v_{3T} \end{bmatrix} - \begin{bmatrix} \rho & \sqrt{\frac{\rho(1+\rho)}{2}} - \rho - 1 \\ \rho & -\rho \end{bmatrix} \begin{bmatrix} u_3 \\ v_3 \end{bmatrix} = \begin{bmatrix} f_3(u_1,u_2,v_1,v_2) - u_{1\tau} \\ g_3(u_1,u_2,v_1,v_2) - v_{1\tau} \end{bmatrix}. \tag{3.12}$$

With the solvability condition applied to the right-hand side, we get the equation for $R(\tau)$:

$$R_\tau = k_1 R + k_3 R^3 \tag{3.13}$$

with coefficients

$$k_1 = \frac{1}{2} - \frac{2\rho^2 c_1^2}{(2+\rho)^2(2+4\rho-3\sqrt{2\rho(1+\rho)}\,)}, \tag{3.14}$$

$$k_3 = -\frac{\rho(1+\rho+\sqrt{2\rho(1+\rho)}\,)}{1-\rho}.$$

The nontrivial steady state amplitude satisfies

$$R^2 = -\frac{k_1}{k_3} = \frac{(2+\rho)^2(2+4\rho-3\sqrt{2\rho(1+\rho)}\,) - 4\rho^2 c_1^2}{2\rho(2+\rho)^2(2+2\rho-\sqrt{2\rho(1+\rho)}\,)}. \tag{3.15}$$

The coefficient k_3 is always negative since $0 < \rho < 1$. However, $k_1 > 0$ for c_1 satisfying

$$-\frac{2+\rho}{2\rho}\sqrt{2+4\rho-3\sqrt{2\rho(1+\rho)}} < c_1 < \frac{2+\rho}{2\rho}\sqrt{2+4\rho-3\sqrt{2\rho(1+\rho)}}. \tag{3.16}$$

The amplitude R is zero ($k_1 = 0$) at the two outer values of c_1 thus indicating these values correspond to the Hopf bifurcation points. This is in agreement with (3.5). Due to the signs of the coefficients, the analysis shows that the slow time steady state R in (3.15) is stable so the bifurcating periodic solution is stable. It is important to note that, since c_1 varies from both Hopf points, the expression R of (3.15) shows that the periodic solution branches from the two Hopf points are connected. Also, the amplitude is a function of the deviation c_1 from the coalescence value of c. A bifurcation diagram reflecting the analysis in this section is shown in Figure 4.

4. Resolving the vertical branch—the onset of outbreaks. From an ecological viewpoint, the case when the branch of periodic solutions from the right Hopf point is subcritical, as in Figure 3(c), is important since it can correspond to a large cyclic outbreak in the populations. Wollkind, Collings, and Logan [11], in their study of predacious mites and spider mite pests, suggest that events such as reducing the prey population via pesticides may effectively shift the initial conditions so that the system moves from a stable equilibrium to a high amplitude periodic orbit, thus enabling the pest to reach intolerable levels. As a possible control on the populations when the parameters are in a subcritical regime, they discuss artificially increasing r_2 (which is equivalent to increasing r) to stabilize the system. In terms of Figure 1 and (2.4), a larger r value moves the Hopf bifurcation points closer or may remove them entirely. In the former case the system would have smaller amplitude periodic solutions as seen in Figure 3(a) or Figure 4 and, in the latter case, there would only be stable equilibrium populations. Either case would be ecologically preferable to large outbreaks of the spider mite pest that are possible in the subcritical parameter regime.

The ecological ramifications of outbreaks make studying the bifurcation structure in the subcritical case important. In this section the subcritical branch will be analyzed when the parameters correspond to the onset of outbreaks. The amplitude of both the stable and unstable periodic solutions of the subcritical branch are analytically determined and the size of the outbreaks is found as a function of the deviation of the growth rate parameter r from a critical value.

When the R^3 coefficient, $\beta(r,u)$, is zero in the amplitude equation (2.14), the branch of periodic solutions from the rightmost Hopf bifurcation point is "vertical" (on the border between supercritical and subcritical) as in Figure 3(b). The curve

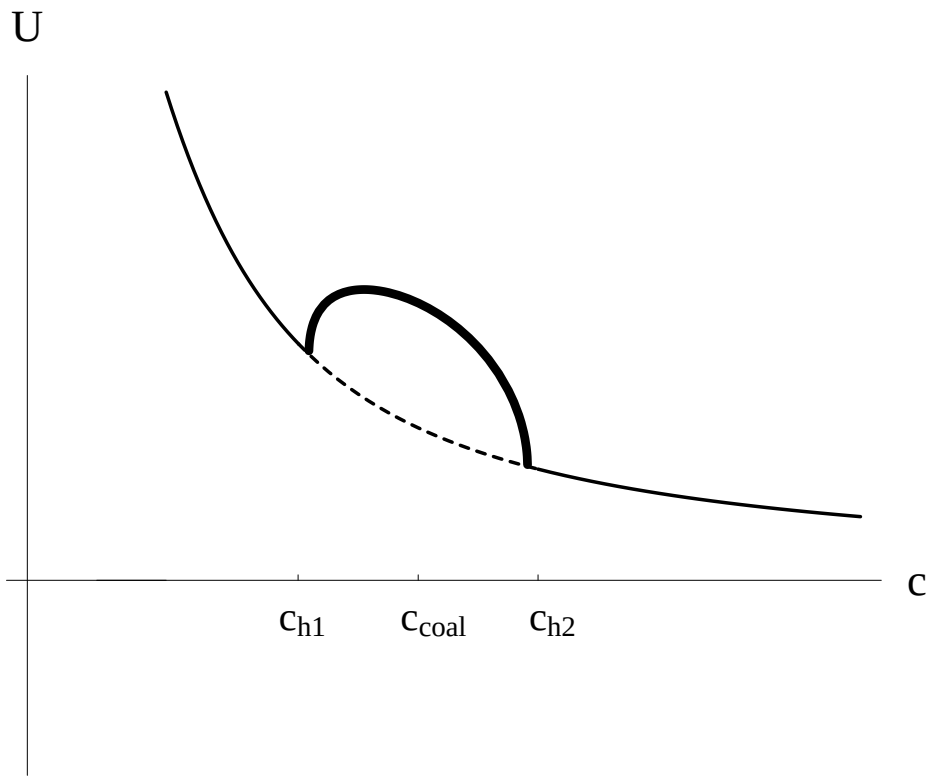


FIG. 4. *The bifurcation diagram in the case of nearly coalescing Hopf bifurcation points, denoted by c_{h1} and c_{h2}, calculated with the modified two-timing results of section* 3. *The thick solid curve shows the maximum amplitude of the stable branch of periodic solutions using the amplitude expression* (3.15). *The dashed (solid) curve is the unstable (stable) steady state. The parameters used were $b = 2/7$ and $r = .13$. The coalescence value is $r = \rho = 1/7$.*

$\beta(r,u) = 0$ is seen in Figure 2 and the corresponding curve in the (r,b) plane is in Figure 1. The two-timing analysis must be modified to address this degeneracy. The steady state amplitude $R = \sqrt{\frac{-\alpha(r,u)P}{\beta(r,u)}}$ from (2.16) suggests a rescaling is necessary in the case when $0 < |\beta(r,u)| \ll 1$.

We proceed by defining the small parameter ε by

$$r = \rho + r_2\,\varepsilon^2, \tag{4.1}$$

where $r_2 = \pm 1$ will determine whether the branch is supercritical or subcritical and ρ is a free parameter. The parameter b is fixed at the value

$$b_v = u_v(1 - \rho - 2u_v)/(\rho + u_v) \tag{4.2}$$

by using (2.5) with u_v satisfying $\beta(\rho, u_v) = 0$ (the subscript v signifies vertical). The parameter c is expanded about the rightmost Hopf bifurcation point of (2.4) as

$$c = c_+(b_v, r) + \varepsilon^4 c_4 \tag{4.3}$$

with c_4 allowed to vary. The amplitude of the periodic solutions will be shown to depend on c_4. The "size" of the outbreaks is defined as the distance from $c_+(b_v, r)$ to

the turning point value of c of the semistable limit cycle. By (4.3) and (4.1), the size is proportional to $(r-\rho)^2$, the square of the difference of the growth rate parameter r from its vertical value ρ.

The variables U and V are expanded about the rightmost Hopf point of (2.4) as

$$\begin{aligned} U(T,\tau) &= u_-(b_v,r) + \varepsilon u_1(T,\tau) + \varepsilon^2 u_2(T,\tau) + \cdots, \\ V(T,\tau) &= u_-(b_v,r) + \varepsilon v_1(T,\tau) + \varepsilon^2 v_2(T,\tau) + \cdots. \end{aligned} \tag{4.4}$$

Note that the expansions in (4.3) and (4.4) will rescale the bifurcating branch to be quartic rather than quadratic as found with the usual scaling (2.6) and (2.8). The quartic branch appropriately accounts for the vertical branch since it is flatter locally. A frequency correction is required in a new fast time, $T = (1+\varepsilon^2\omega_2 + \cdots)t$ and the slow time must be defined by

$$\tau = \varepsilon^4 t \tag{4.5}$$

rather than $\tau = \varepsilon^2 t$, found in the general case of section 2.

As before, with (4.1)–(4.5) substituted into the main equations (2.1), we get a sequence of equations at each power of ε. Unlike the previous analysis, the slow time amplitude equations are found by considering a solvability condition at $O(\varepsilon^5)$ rather than $O(\varepsilon^3)$. Rather than show the general case (the algebra is unwieldy), the method will be illustrated by using specific numerical values for the parameters. The values $\rho = 1/12$ and $u_v = 1/12$ satisfy the vertical condition $\beta(\rho, u_v) = 0$ and give $b_v = 3/8$. With these parameters, the rightmost Hopf bifurcation point occurs at

$$\begin{aligned} u_- &= \frac{1}{12} + \frac{11\,r_2}{5}\varepsilon^2 + \frac{7128}{125}\varepsilon^4 + O(\varepsilon^6) \qquad \text{and} \\ c_+ &= \frac{121}{24} - 121\,r_2\varepsilon^2 + O(\varepsilon^6). \end{aligned} \tag{4.6}$$

The $O(\varepsilon)$ equations are

$$\begin{bmatrix} u_{1T} \\ v_{1T} \end{bmatrix} = \begin{bmatrix} 1/12 & -11/12 \\ 1/12 & -1/12 \end{bmatrix} \begin{bmatrix} u_1 \\ v_1 \end{bmatrix} \tag{4.7}$$

and the solutions are given by

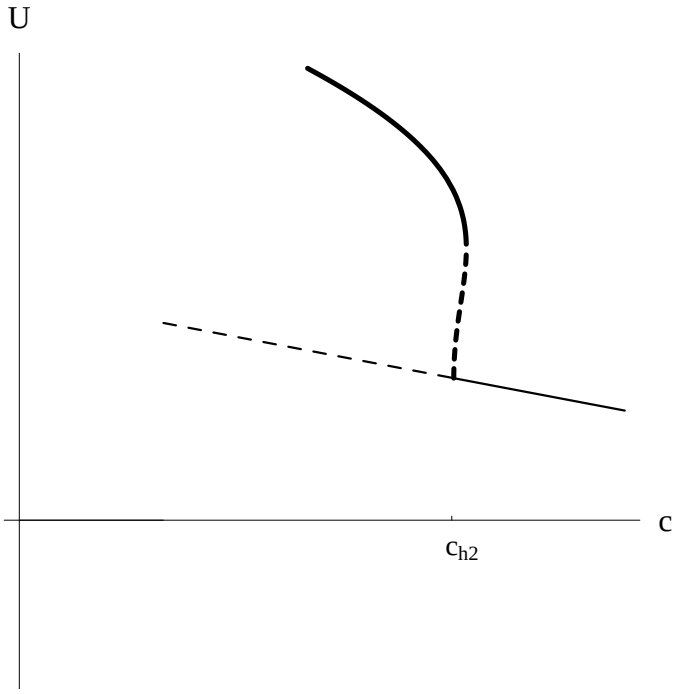
$$\begin{bmatrix} u_1(T,\tau) \\ v_1(T,\tau) \end{bmatrix} = R(\tau)e^{i\theta(\tau)} \begin{bmatrix} 1+i\sqrt{10} \\ 1 \end{bmatrix} e^{i\frac{\sqrt{10}}{12}T} + R(\tau)e^{-i\theta(\tau)} \begin{bmatrix} 1+i\sqrt{10} \\ 1 \end{bmatrix} e^{-i\frac{\sqrt{10}}{12}T}. \tag{4.8}$$

A solvability condition applied to the $O(\varepsilon^3)$ equations gives the amplitude-dependent frequency correction, $\omega_2 = \frac{72}{25\sqrt{10}}(17\,r_2 - 1270R^2)$. Finally, the slow time amplitude equation for $R(\tau)$ is obtained by considering the $O(\varepsilon^5)$ equations. The solvability condition gives the fifth order equation

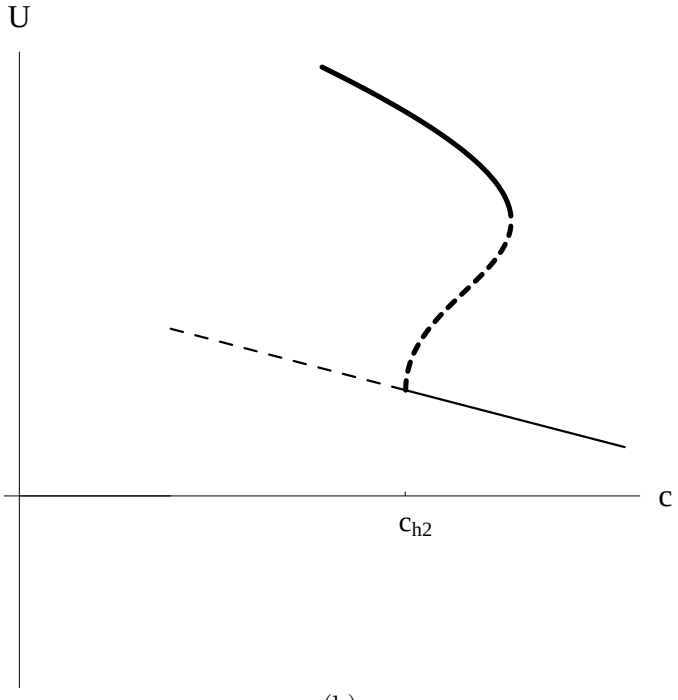
$$R_\tau = \frac{-9797760}{121}\,R^5 - \frac{104976\,r_2}{55}\,R^3 - \frac{c_4}{242}\,R. \tag{4.9}$$

The relevant feature is that there are either one or two nonzero steady states depending on the parameters c_4 and r_2 in addition to the zero steady state. The nonzero steady state(s) satisfy

$$R^2 = \frac{-33r_2 \pm \sqrt{33^2 - \frac{875}{2187}c_4}}{2800}. \tag{4.10}$$



(a)



(b)

FIG. 5. *The subcritical branch of periodic solutions as calculated using the expressions in section* 4 *with* $\varepsilon = .02$ *in Figure* 5(a) *and* $\varepsilon = .04$ *in Figure* 5(b). *The thick solid (thick dashed) curve shows the maximum amplitude of the stable (unstable) branch of periodic solutions. The parameters used were* $b = 3/8$ *and* $\rho = 1/12$.

There are different cases to consider. If $r_2 = 1$, then (4.10) gives one real R when $c_4 < 0$ and the branch is supercritical (stable), but there are no real R if $r_2 = 1$ and $c_4 > 0$. If $r_2 = -1$, the branch is subcritical and (4.10) gives two real R when $0 < c_4 < 33^2(2187)/875$. The lesser R value is unstable whereas the greater R value is stable. The branch has a turning point at $c_4^* = 33^2(2187)/875$. This value of c_4 corresponds to the coalescence of the stable and unstable limit cycles forming a semistable limit cycle. For $r_2 = -1$ and $c_4 \leq 0$ the single value of R from (4.10) is stable and represents a continuation of the subcritical branch.

Figure 5 shows the bifurcation diagram with the subcritical branch of periodic solutions using the results of this section. The upper branch of stable periodic solutions corresponds to onset of outbreaks since the populations can achieve a relatively high amplitude oscillation. For the parameters of this example, the outbreak size is $\varepsilon^4 c_4^* = (\rho - r)^2 c_4^* = (1/12 - r)^2 c_4^*$. The analysis in this section is valid when $\varepsilon \ll 1$ so it can only be said to describe the onset of outbreaks. However, as is often the case, the analysis agrees quite well with numerical results when the condition $\varepsilon \ll 1$ is relaxed somewhat. Hence the parametric features of outbreaks (i.e., the size is proportional to $(\rho - r)^2$, etc.) can be extended beyond the onset case.

5. Further comments. The presence of two Hopf bifurcation points in the Holling–Tanner model is ubiquitous. Actual ecological dynamics in the field observed when physical parameters are varied agree qualitatively with features, such as outbreaks, predicted by the model. The two-timing techniques give an organized way to characterize the changes in the bifurcation diagram as the parameters are varied. The modifications required in the degenerate cases have broad applicability to other systems in which two Hopf bifurcations are present.

REFERENCES

[1] R. M. May, *Stability and Complexity in Model Ecosystems*, Princeton University Press, Princeton, NJ, 1973.

[2] R. M. May, *Limit cycles in predator-prey communities*, Science, 177 (1972), pp. 900–902.

[3] W. S. C. Gurney and R. M. Nisbet, *Ecological Dynamics*, Oxford University Press, New York, 1998.

[4] A. N. Kolmogorov, *Sulla, Teoria di Volterra della Lotta per l'Esisttenza*, Giorna. Instituto Ital. Attuarri, 7 (1936), pp. 74–80.

[5] W. E. Hutchinson, *An Introduction to Population Ecology*, Yale University Press, New Haven, CT, 1978.

[6] C. S. Holling, *The functional response of invertebrate predators to prey density*, Mem. Ent. Soc. Can., 45 (1965), pp. 3–60.

[7] M. P. Hassell, *The Dynamics of Arthropod Predator-Prey Systems*, Princeton University Press, Princeton, NJ, 1978.

[8] P. H. Leslie, *Some Further Notes on the Use of Matrices in Population Mathematics*, Biometrica, 35 (1948), pp. 213–245.

[9] E. Sáez and E. González-Olivares, *Dynamics of a predator-prey model*, SIAM J. Appl. Math., 59 (1999), pp. 1867–1878.

[10] J. T. Tanner, *The stability and the intrinsic growth rates of prey and predator populations*, Ecology, 56 (1975), pp. 855–867.

[11] D. J. Wollkind, J. B. Collings, and J. A. Logan, *Metastability in a temperature-dependent model system for predator-prey mite outbreak interactions on fruit flies*, Bull. Math. Biol., 50 (1988), pp. 379–409.

[12] S.-B. Hsu and T.-W. Huang, *Global stability for a class of predator-prey systems*, SIAM J. Appl. Math., 55 (1995), pp. 763–783.

[13] E. J. Doedel and J. P. Kernévez, *AUTO: Software for Continuation and Bifurcation Problems in Ordinary Differential Equations*, Technical report, California Institute of Technology, Pasadena, CA, 1986.

[14] J. B. Collings, *Bifurcation and stability analysis of a temperature-dependent mite predator-*

prey interaction model incorporating a prey refuge, Bull. Math. Biol., 57 (1995), pp. 63–76.
[15] W. Baltensweiler, *The relevance of changes in the composition of larch bud moth populations for the dynamics of its numbers*, in Dynamics of Populations, P. J. den Boer and G. R. Gradwell, eds., Center for Agricultural Publication and Documentation, Wageningen, The Netherlands, 1971, pp. 208–219.
[16] R. A. Dolbeer, *Population Dynamics of the Snowshoe Hare in Colorado*, Ph.D. Thesis, Colorado State University, Fort Collins, CO, 1972.
[17] L. S. Luckinbill, *Coexistence in laboratory populations of Paramecium aurelia and its predator Didinium nasutum*, Ecology, 54 (1973), pp. 1320–1327.
[18] S. Wolfram, *Mathematica* 4.0, Wolfram Research Inc., Champaign, IL, 1999.
[19] M. Golubitsky and W. F. Langford, *Classification and unfoldings of degenerate Hopf bifurcations*, J. Differential Equations, 41 (1981), pp. 375-415.

SIAM J. APPL. MATH.
Vol. 63, No. 3, pp. 905–922

A TERRAIN-FOLLOWING BOUSSINESQ SYSTEM*

ANDRÉ NACHBIN†

Abstract. A long wave model is derived asymptotically from the nonlinear potential theory equations. The flow regime of interest is incompressible, irrotational, and inviscid. Asymptotic analysis leads to a weakly nonlinear, weakly dispersive (variable coefficient) Boussinesq system valid for a wide class of topographies. The mild slope hypothesis is not required and rapidly varying topographies are also considered. In analogy with atmospheric models we use a terrain-following coordinate system. The novelty is that this coordinate system naturally suggests the weighted averaging of terrain-following velocity components, as opposed to the depth-average of horizontal velocity components found in standard shallow water formulations. Furthermore, a Schwarz–Christoffel toolbox is used to provide additional insight on these new results. Regarding applications, the proposed model can be used for studying solitary waves interacting with fine scale inhomogeneities, a theme of great interest. The terrain-following model also presents potential numerical advantages for Boussinesq solvers.

Key words. dispersive waves, inhomogeneous media, asymptotic theory

AMS subject classifications. 76B07, 76B15, 35Q

PII. S0036139901397583

1. Introduction.

Shallow water (long wave) models have been known and studied for many years. Among several important partial differential equations that fall into this class are weakly dispersive, weakly nonlinear models such as the KdV equation or the Boussinesq equations [7, 16]. The flow regime considered is incompressible, irrotational, and inviscid. These long wave models can be derived asymptotically from the nonlinear potential theory equations when the leading order terms, in the nonlinearity and dispersion parameters, are retained. Their analysis has received a great deal of attention from both the mathematical theory and applications viewpoints.

Of great research interest are problems involving these models in the presence of heterogeneous media, which can be of different nature depending on the application. Shallow water models are not only of interest in ocean applications but also very common in meteorology [5], where the shallow propagation medium is the atmosphere. Literature is abundant on the mathematical analysis of long wave models, their numerical discretization and behavior, as well as the implementation of oceanic and atmospheric simulators. Reference to some of this work will not be attempted due to the richness of the subject. It is bound to be injudicious.

This work is concerned with the derivation of a long wave model valid in the presence of a wide class of bottom profiles, including rapidly varying topographies. Most of the well-established long wave models are derived under the hypothesis of having mild slope topographies. Among others, we mention three interesting recent works which improve classical shallow water models and include topographic effects. Camassa, Holm, and Levermore [2] derive shallow water equations that have a Hamiltonian principle formulation and which model long-time effects of slowly varying topographies. Milewski [8] considers the propagation of long waves on the surface of a three-

*Received by the editors November 6, 2001; accepted for publication (in revised form) July 22, 2002; published electronically January 23, 2003. This research was supported by CNPq/Brazil under grant 300368/96-8 and NEC/CEMAT–IMPA under grant DMS 2.419/98-00.
http://www.siam.org/journals/siap/63-3/39758.html
†Instituto de Matemática Pura e Aplicada, Est. D. Castorina 110, Jardim Botânico, Rio de Janeiro, RJ 22460-320, Brazil (nachbin@impa.br).

dimensional fluid domain bounded below by a slowly varying topography. Weakly nonlinear wave-topography interaction is studied in two limits giving rise to a variable coefficient KdV equation or a variable coefficient KP (Kadomtsev–Petviashvili) equation. We also mention the work of Schäffer and Madsen [15], who derive a new set of Boussinesq equations which provide improvements for the linear dispersion relation and shoaling characteristics. For rapidly varying topographies it is worth mentioning the earlier work of Rosales and Papanicolaou [14], where effective KdV equations were derived for rapidly varying periodic and (small amplitude) random topographies. But in both cases back-scattering is negligible. Finally we point out the work of Hamilton [6], which is the starting point for the present paper. Hamilton used a conformal mapping technique to derive long wave models on a fluid of rapidly varying depth. In particular, a Boussinesq system is presented [6, equations (A14)–(A15)] in the form of a second order in time 2×2 system. The dependent variables considered are the depth-averaged potential and wave elevation. We derive a first order 2×2 Boussinesq system in terms of the wave elevation and the averaged terrain-following velocity components. For flat channels our system reduces to the standard Boussinesq system found in the literature [7, 16]. Hamilton's conformal mapping technique has also been used by the author to study the reflection-transmission problem over disordered (random) topographies [11, 12, 13].

Moreover, our main result is the derivation of a weakly nonlinear, weakly dispersive (variable coefficient) Boussinesq system valid for a wide class of topographies and multiple scattering problems. The mild slope hypothesis is not required. In analogy with some atmospheric models [3] we use a terrain-following coordinate system for the independent variables. This curvilinear coordinate system is obtained from the theory of conformal mappings as suggested in Hamilton [6]. Hamilton does not use dimensionless variables and performs the asymptotic simplification of equations arguing through chosen length scales. As mentioned above, his formulation of a nonlinear Boussinesq system is briefly outlined in Appendix A [6]. Our approach is to perform the asymptotic simplification of the nonlinear potential theory equations in a systematic way (as in Whitham [16]) by using dimensionless variables and the conformal mapping setting of Hamilton. All steps in the derivation are presented. The curvilinear coordinate system naturally suggests the weighted averaging of terrain-following velocity components, as opposed to the depth-average of horizontal components found in standard shallow water formulations. For rapidly varying topographies the standard depth-average strategy breaks down. As pointed out by Hamilton [6] the truncation term in the near-bottom series expansion will dominate the leading order terms when the bottom is rapidly varying (top of page 292 in [6]). In the dimensionless formulation (given below in section 2) the large truncation term will be caused by the β/γ term in the Neumann condition, when γ is small. In contrast, the asymptotic expansion, for the free surface conditions in the curvilinear coordinates, generates truncation terms that are negligible compared to the leading order ones. The Neumann condition is trivial in this framework, as will be shown in section 4.

Furthermore, our use of dimensionless variables enables a better understanding of the impact of the conformal mapping technique on the asymptotic analysis. Using Driscoll's [4] MATLAB Schwarz–Christoffel toolbox (SC-Toolbox), we provide additional insight on why the terrain-following Boussinesq system is a good model even in the presence of rapidly varying topographies. Using the SC-Toolbox, we graph the curvilinear coordinate system for both slowly and rapidly varying topographies. The average of terrain-following velocity components is clearly seen as a weighted average

along special curves, orthogonal to the topography and to the undisturbed free surface. This weighted averaging naturally adjusts to either slowly varying or rapidly varying topographies, as will be shown graphically with the SC-Toolbox.

Another result of interest is that the nonlinear shallow water equations can be viewed as an $O(\alpha, \sqrt{\beta})$ approximation of the full potential theory equations. The standard notation is used: α is the nonlinearity parameter and β the dispersion parameter. The nonlinear shallow water system is (of course) hyperbolic, but the square root of the dispersion parameter plays a role only on the smoothing of sharp features of the topography. This had been pointed out by Hamilton [6], but the rate of smoothing was not explicitly identified, nor the nonlinear shallow water equation presented.

The proposed model can be used for studying solitary waves over disordered (random) topographies, a theme of great interest [9, 10]. This could not have been done with existing Boussinesq models. We are currently investigating the numerical advantages of working with this model. A potential advantage resides in the fact that within this new framework, the variable ("topographic") coefficient moves away from the dispersive term (a third order derivative) and places itself at first order terms. We have results that indicate the improved performance of Boussinesq solvers [10].

The paper is organized as follows. In section 2 we present the standard scaling [16] and the (dimensionless) nonlinear potential theory equations in Cartesian coordinates. Section 3 describes the conformal mapping theory in full detail along with the calculation of the variable free surface coefficient's leading order term. The (dimensionless) nonlinear potential theory equations, in curvilinear coordinates, are presented in section 4 and the weakly nonlinear, weakly dispersive asymptotic theory in section 5. The applications and conclusions are given in section 6. Appendix A contrasts the terrain-following Boussinesq system with standard Boussinesq formulations.

2. Formulation and scaling. Let variables with physical dimensions be denoted with a tilde. We introduce the length scales σ (a typical pulse width or wavelength), h_0 (a typical depth), a (a typical wave amplitude), l_b (the horizontal length scale for bottom irregularities), and L (the total length of the rough region or the total propagation distance). The acceleration due to gravity is denoted by g and the reference shallow water speed is $c_0 = \sqrt{gh_0}$. Dimensionless variables are then defined in a standard fashion [14, 16] by having

$$\tilde{x} \;=\; \sigma x, \qquad \tilde{y} \;=\; h_0\, y, \qquad \tilde{t} \;=\; \left(\frac{\sigma}{c_0}\right) t,$$

$$\tilde{\eta} \;=\; a\,\eta, \qquad \tilde{\phi} \;=\; \left(\frac{g\sigma a}{c_0}\right)\phi, \qquad \tilde{h} \;=\; h_0\, H\left(\frac{\tilde{x}}{l_b}\right).$$

The velocity potential $\phi(x, y, t)$ and wave elevation $\eta(x, t)$ satisfy the dimensionless equations [16]:

$$\beta\, \phi_{xx} + \phi_{yy} \;=\; 0 \quad \text{for} \quad -H(x/\gamma) < y < \alpha\eta(x, t),$$

with the nonlinear free surface conditions

$$\eta_t + \alpha\phi_x\eta_x - \frac{1}{\beta}\phi_y = 0,$$

$$\eta + \phi_t + \frac{\alpha}{2}\left(\phi_x^2 + \frac{1}{\beta}\phi_y^2\right) = 0$$

at $y = \alpha\eta(x,t)$. The Neumann condition at the impermeable bottom is

$$\phi_y + \frac{\beta}{\gamma}\, H'(x/\gamma)\phi_x = 0.$$

The bottom topography is described by $y = -H(x/\gamma)$, where

$$H(x/\gamma) \;=\; \begin{cases} 1 + n(x/\gamma) & \text{when } \; 0 < x < L, \\ 1 & \text{when } \; x \leq 0 \;\text{ or }\; x \geq L. \end{cases}$$

The bottom profile is described by the (possibly rapidly varying) function $-n(x/\gamma)$. The topography is rapidly varying when $\gamma \ll 1$. The undisturbed depth is given by $y = -1$ and the topography can be of large amplitude provided that $|n| < 1$. We do not need to assume that the fluctuations n are small, nor continuous, nor slowly varying.

The following dimensionless parameters arise:

$\alpha = a/h_0$ (nonlinearity parameter)

$\beta = h_0^2/\sigma^2$ (dispersion parameter)

$\gamma = l_b/\sigma$ (bottom irregularities compared to the wavescale).

Before proceeding with the asymptotic analysis at the level of equations, we first change the underlying Cartesian coordinate system as follows.

3. Conformal mapping. A mapping from a uniform strip onto the fluid domain at rest is constructed analytically. Let the former (computational) domain be defined in the complex w-plane and the rough undisturbed channel (physical domain) be defined in the complex z-plane. Properties of the $z(w)$ mapping will be calculated below. As will be shown, working with a symmetric domain is very convenient for the asymptotic analysis to be performed. This is the reason why we solve the (harmonic) conformal mapping problem in such a symmetric configuration (cf. Figure 3.1). Moreover, instead of referring to the w-plane as our computational domain, we will (equivalently) interpret our formulation as working with orthogonal curvilinear coordinates in the physical domain (the w-plane horizontal and vertical level curves). In Figure 3.1 we superimpose the symmetric domain in the complex z-plane with the curvilinear level curves from the w-plane coordinate system. The polygonal line at the bottom of Figure 3.1 is a schematic representation of the topography.

Following Hamilton [6] we define a symmetric flow domain by reflecting the original one about the undisturbed free surface (cf. Figure 3.1). We denote this domain by Ω_z, where $z = x + i\sqrt{\beta}y$, and consider it as the conformal image of the strip Ω_w, where $w = \xi + i\tilde{\zeta}$ with $|\tilde{\zeta}| \leq \sqrt{\beta}$. Then $z = x(\xi,\tilde{\zeta}) + i\sqrt{\beta}y(\xi,\tilde{\zeta}) = x(\xi,\tilde{\zeta}) + i\tilde{y}(\xi,\tilde{\zeta})$ with x and $\tilde{y}$ a pair of harmonic functions on Ω_w. As will become clear in the following sections (in particular section 5), working with x and y is convenient for long wave (shallow water) asymptotics. The depth is kept fixed in this case. On the other hand, working with x and $\tilde{y}$ is convenient for computing harmonic functions because the parameter β drops from the Laplacian. These are the reasons why we switch from one scaling to the other.

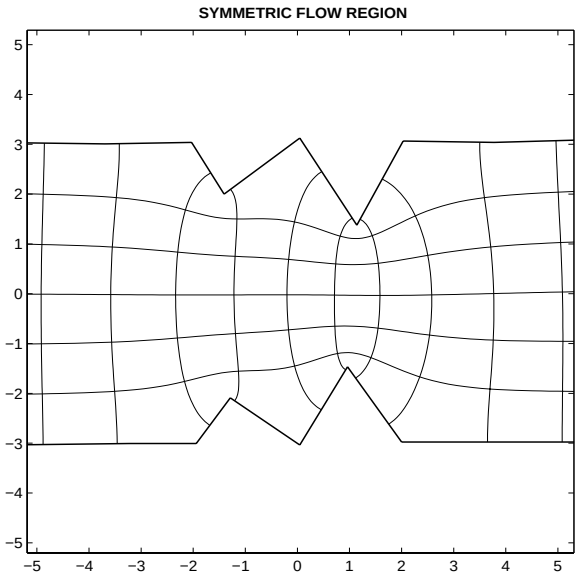


FIG. 3.1. *The symmetric domain in the complex z-plane, where* $z = x(\xi, \tilde{\zeta}) + i\tilde{y}(\xi, \tilde{\zeta})$. *The lower half* ($x \in [-5, 5]$, $y \in [-3, 0]$) *is the physical channel with* $y = \tilde{\zeta} = 0$ *indicating the undisturbed free surface. Superimposed in this complex z-plane domain are the (curvilinear) coordinate level curves from the w-plane system* $\xi\tilde{\zeta}$. *The polygonal line at the bottom of the figure is a schematic representation of the topography (where* $\tilde{\zeta} = \pm\sqrt{\beta}$). *This figure was generated using the SC-Toolbox* [4].

As will become clear, our main goal is to calculate $\tilde{y}_{\tilde{\zeta}}$ along the undisturbed free surface $\tilde{\zeta} = \tilde{y} = 0$. This coefficient is calculated analytically by solving Laplace's equation for the imaginary part of the conformal change of variables. Once this coefficient is obtained we proceed with the derivation of the full potential theory equations in the curvilinear $(\xi, \tilde{\zeta})$ coordinate system.

The imaginary part of the conformal map is the harmonic function that satisfies [6, 11]:

$$\Delta\, \tilde{y}(\xi, \tilde{\zeta}) \;=\; 0 \qquad \text{in } \Omega_w \tag{3.1}$$

with Dirichlet boundary conditions

$$\tilde{y}(\xi, \pm\sqrt{\beta}) \;=\; \pm\, h(x(\xi)) \equiv \pm\sqrt{\beta} H\left(\frac{x(\xi, \pm\sqrt{\beta})}{\gamma}\right). \tag{3.2}$$

Its harmonic conjugate is $x(\xi, \tilde{\zeta})$. The Green's function for problem (3.1)–(3.2) is slightly different from Hamilton's because we introduce dimensionless variables and keep depth effects through the parameter $\sqrt{\beta}$. The Green's function, vanishing along the lines $\tilde{\zeta} = \pm\sqrt{\beta}$, is given by

$$G(w; w_0) = \mathbf{Re}\, \log\, \left((e^{\pi w/2\sqrt{\beta}} - e^{\pi w_0/2\sqrt{\beta}})/(e^{\pi w/2\sqrt{\beta}} + e^{\pi \overline{w_0}/2\sqrt{\beta}})\right), \tag{3.3}$$

where **Re** stands for the real part and the overbar denotes complex conjugation. Near a source point w_0

$$G(w; w_0) \sim\; \mathbf{Re}\, \log(w - w_0),$$

meaning that it behaves like the free-space Green's function. At the solid boundaries $\tilde{\zeta} = \pm\sqrt{\beta}$,

$$G \equiv 0.$$

Additional asymptotic properties are [6]

$$G, \;\; G_\xi \to 0 \;\; \text{as} \;\; \xi \to \pm\infty \;\; \text{in} \;\; -\sqrt{\beta} \leq \tilde{\zeta} \leq \sqrt{\beta}.$$

Using Green's third identity, we have that

$$2\pi \; \tilde{y}(\xi_0, \tilde{\zeta}_0) \;=\; \oint_{\partial\Omega_w} \tilde{y}(\xi, \tilde{\zeta}) \; \frac{dG}{dn}(w; w_0) \; ds.$$

By the conditions above this is the same as

$$2\pi \; \tilde{y}(\xi_0, \tilde{\zeta}_0) \;=\; \int_{-\infty}^{\infty} h(x(\xi)) \; \left(G^+_{\tilde{\zeta}} \;+\; G^-_{\tilde{\zeta}} \right) \; d\xi,$$

where

$$G^+_{\tilde{\zeta}} = \frac{\partial G}{\partial \tilde{\zeta}}(\xi, +\sqrt{\beta}; \xi_0, \tilde{\zeta}_0) \quad \text{and} \quad G^-_{\tilde{\zeta}} = \frac{\partial G}{\partial \tilde{\zeta}}(\xi, -\sqrt{\beta}; \xi_0, \tilde{\zeta}_0).$$

Differentiating this identity with respect to $\tilde{\zeta}_0$ and evaluating at $\tilde{\zeta}_0 = 0$ we get our quantity of interest. Namely we have that at the undisturbed free surface

$$\tilde{y}_{\tilde{\zeta}_0}(\xi_0, 0) \;=\; \frac{1}{2\pi} \int_{-\infty}^{\infty} h(x(\xi)) \left(G^+_{\tilde{\zeta}\tilde{\zeta}_0} \;+\; G^-_{\tilde{\zeta}\tilde{\zeta}_0} \right) d\xi. \tag{3.4}$$

The kernel comes from evaluating

$$G_{\tilde{\zeta}\tilde{\zeta}_0}(\xi, -\sqrt{\beta}; \xi_0, 0) \;+\; G_{\tilde{\zeta}\tilde{\zeta}_0}(\xi, \sqrt{\beta}; \xi_0, 0) \;=\; \frac{2\pi^2}{\beta} \;\; \frac{e^{\frac{\pi}{\sqrt{\beta}}(\xi + \xi_0)}}{(e^{\frac{\pi}{\sqrt{\beta}}\xi} + e^{\frac{\pi}{\sqrt{\beta}}\xi_0})^2} \tag{3.5}$$

$$= \; \frac{\pi^2/\beta}{2 \; \cosh^2 \frac{\pi}{2\sqrt{\beta}}(\xi - \xi_0)},$$

and finally we obtain

$$\tilde{y}_{\tilde{\zeta}_0}(\xi_0, 0) = \; \frac{\pi}{4\beta} \; \int_{-\infty}^{\infty} \; \frac{\sqrt{\beta} H(x(\xi, -\sqrt{\beta})/\gamma)}{\cosh^2 \frac{\pi}{2\sqrt{\beta}}(\xi - \xi_0)} d\xi. \tag{3.6}$$

This expression was obtained by Hamilton [6], but without the $\beta^{1/2}$-scaling. This intermediate $\beta^{1/2}$-scale plays an important role indicating the degree of topography smoothing. The rate of smoothing actually depends on $\beta^{1/2}/\gamma$. Numerical examples of the topography smoothing are presented at the end of this paper (Figures 6.1 and 6.2). We call $\beta^{1/2}$ an intermediate scale because we will work with an integer power expansion in β as in Whitham's [16] derivation of Boussinesq's equations.

In the $(\xi, \tilde{\zeta})$ coordinate system the varying bottom topography is straightened out and a variable coefficient, namely $\tilde{y}_{\tilde{\zeta}}$, will appear in the free surface condition. Before writing the equations in the $(\xi, \tilde{\zeta})$ system note that

$$\int_{-\infty}^{\infty} \frac{\pi}{4\sqrt{\beta}} \operatorname{sech}^2 \left[\frac{\pi}{2\sqrt{\beta}}(x-y) \right] dx \;=\; \frac{1}{2} \tanh \left[\frac{\pi}{2\sqrt{\beta}} x \right]_{-\infty}^{\infty} = 1. \tag{3.7}$$

Hence as $\sqrt{\beta} \downarrow 0$ the kernel in (3.6) goes to a delta function and the bottom is felt at the free surface level without any smoothing. Smoothing takes place only for $\beta^{1/2} > 0$.

At the undisturbed level we define the *variable free surface coefficient*

$$M(\xi) \equiv \tilde{y}_{\tilde{\zeta}}(\xi, 0) = 1 + m(\xi),$$

where

$$m(\xi; \sqrt{\beta}, \gamma) \equiv \frac{\pi}{4\sqrt{\beta}} \int_{-\infty}^{\infty} \frac{n(x(\xi_0, -\sqrt{\beta})/\gamma)}{\cosh^2 \frac{\pi}{2\sqrt{\beta}}(\xi_0 - \xi)} d\xi_0 \;=\; (K * (n \circ x))(\xi). \tag{3.8}$$

The $(\sqrt{\beta}, \gamma)$ parameter dependence will be omitted for brevity. We are ready to rewrite the potential theory equations in the $(\xi, \tilde{\zeta})$ coordinate system. The velocity potential $\phi(\xi, \tilde{\zeta}, t)$ when represented in the curvilinear coordinate system is such that

$$\phi_\xi = \phi_x \; x_\xi(\xi, \tilde{\zeta}) + \phi_{\tilde{y}} \; \tilde{y}_\xi(\xi, \tilde{\zeta})$$

and

$$\phi_{\tilde{\zeta}} = \phi_x \; x_{\tilde{\zeta}}(\xi, \tilde{\zeta}) + \phi_{\tilde{y}} \; \tilde{y}_{\tilde{\zeta}}(\xi, \tilde{\zeta}).$$

In particular, at the undisturbed free surface or for linear problems,

$$\phi_\xi(\xi, 0) = M(\xi) \phi_x$$

and

$$\phi_{\tilde{\zeta}}(\xi, 0) = M(\xi) \phi_{\tilde{y}}.$$

Note that we have used the Cauchy–Riemann equations. Inverting the relation given above, we have

$$\phi_x = \frac{1}{|J|} \left[\tilde{y}_{\tilde{\zeta}} \phi_\xi - \tilde{y}_\xi \phi_{\tilde{\zeta}} \right] \tag{3.9}$$

and

$$\phi_{\tilde{y}} = \frac{1}{|J|} \left[-x_{\tilde{\zeta}} \phi_\xi + x_\xi \phi_{\tilde{\zeta}} \right], \tag{3.10}$$

where

$$|J| = x_\xi \tilde{y}_{\tilde{\zeta}} - \tilde{y}_\xi x_{\tilde{\zeta}} = \tilde{y}_{\tilde{\zeta}}^2 + \tilde{y}_\xi^2.$$

Moreover

$$\phi_x^2 + \phi_{\tilde{y}}^2 = \frac{1}{|J|} \left(\phi_\xi^2 + \phi_{\tilde{\zeta}}^2 \right), \tag{3.11}$$

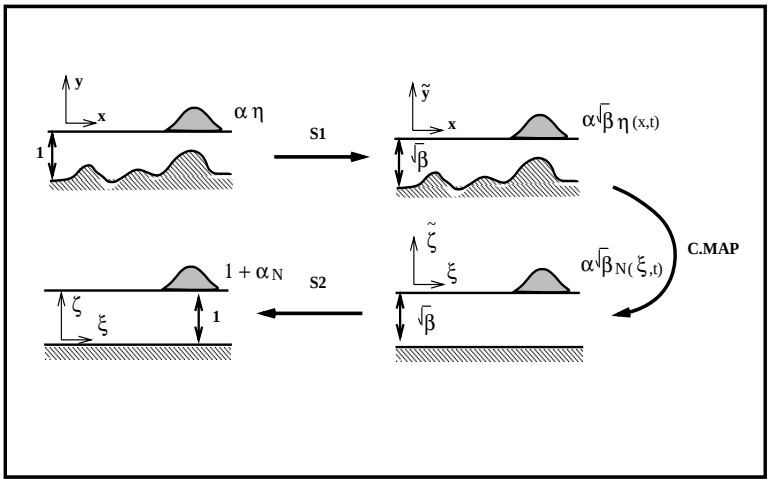


FIG. 3.2. *This is a schematic figure showing the changes of scales performed, indicated by* S1*,* S2*, and the conformal transformation. The geometries on the left are in the shallow water framework (with unit depth). The geometries on the right are used so that the velocity potential solves Laplace's equation in the corresponding variables.*

which relates the speeds in the two coordinate systems.

In the new coordinate system the position of the free surface will be described by some function, here denoted by $N(\xi,t)$, such that

$$\tilde{\zeta} = \alpha\sqrt{\beta}N(\xi,t). \tag{3.12}$$

The new free surface profile, represented by $N(\xi,t)$, does not necessarily resemble $\eta(x,t)$ as indicated in the schematic Figure 3.2. Nevertheless this is a material curve. In the Cartesian coordinates $(x,\tilde{y})$ the dimensionless kinematic condition was given by

$$\sqrt{\beta}\eta_t + \alpha\phi_x\sqrt{\beta}\eta_x - \phi_{\tilde{y}} = 0,$$

which is the same as

$$\frac{D}{Dt}\left(\tilde{y} - \alpha\sqrt{\beta}\eta(x,t)\right) = 0,$$

with the convective derivative

$$\frac{D}{Dt} \equiv \partial_t + \alpha\left(\phi_x\partial_x + \phi_{\tilde{y}}\partial_{\tilde{y}}\right).$$

In the curvilinear coordinates it becomes

$$\frac{\mathcal{D}}{\mathcal{D}t} \equiv \partial_t + \frac{\alpha}{|J|}\left(\phi_\xi\partial_\xi + \phi_{\tilde{\zeta}}\partial_{\tilde{\zeta}}\right),$$

where we have used that $\partial_x = |J|^{-1}(\tilde{y}_{\tilde{\zeta}}\partial_\xi - \tilde{y}_\xi\partial_{\tilde{\zeta}})$ and the corresponding expression for $\partial_{\tilde{y}}$ (cf. (3.9) and (3.10)). The new kinematic condition is obtained by performing

$$\frac{\mathcal{D}}{\mathcal{D}t}\left(\tilde{\zeta} - \alpha\sqrt{\beta}N(\xi,t)\right) = 0$$

and is presented below.

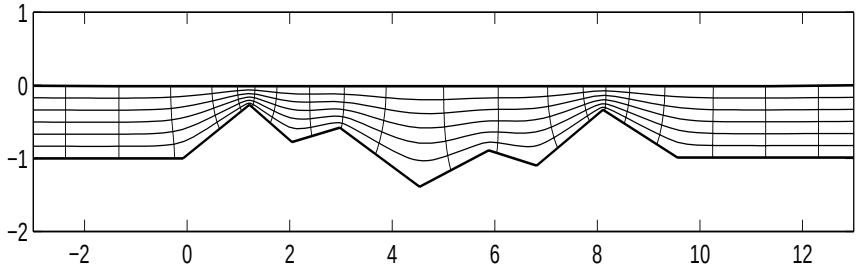


FIG. 4.1. *A schematic figure showing a slowly varying topography in the xy coordinate system together with the ξ and $\tilde{\zeta}$ level-curves. This figure was generated using SC-Toolbox* [4].

4. Nonlinear potential theory equations in terrain-following coordinates. The scaled water wave equations in the fixed orthogonal curvilinear coordinates $(\xi, \tilde{\zeta})$ (cf. Figure 4.1) are

$$\phi_{\xi\xi} + \phi_{\tilde{\zeta}\tilde{\zeta}} = 0, \qquad -\sqrt{\beta} < \tilde{\zeta} < \alpha\sqrt{\beta}N(\xi, t), \tag{4.1}$$

with free surface conditions

$$|J|N_t + \alpha\phi_\xi N_\xi - \frac{1}{\sqrt{\beta}}\phi_{\tilde{\zeta}} = 0 \tag{4.2}$$

and

$$\phi_t + \eta + \frac{\alpha}{2|J|}\left(\phi_\xi^2 + \phi_{\tilde{\zeta}}^2\right) = 0 \tag{4.3}$$

at $\tilde{\zeta} = \alpha\sqrt{\beta}N(\xi, t)$. The bottom condition is

$$\phi_{\tilde{\zeta}} = 0 \quad \text{at} \quad \tilde{\zeta} = -\sqrt{\beta}. \tag{4.4}$$

Regarding the initial conditions note that by starting with a pulse over a region of uniform depth, the initial data are not affected by the conformal mapping (cf. Figure 4.1 at $x \le -1$ and $x \ge 11$). We need only to replace x by ξ in the initial wave conditions. Away from the rough region

$$1 + m(\xi) = \tilde{y}_{\tilde{\zeta}}(\xi, 0) \approx 1,$$

where this adjustment is exponentially fast. Numerical evidence of this adjustment is presented in Figures 6.1 and 6.2. The exponential adjustment rate depends on $\beta^{1/2}$, the important scale we retained in the coefficient (3.8). Hence from the Cauchy–Riemann equations, $x_\xi(\xi, 0) \approx 1$. This was not noticed in Hamilton's work [6] because the analysis was not done in dimensionless variables. Scales were used only to argue which terms could be dropped.

Moreover, it is important to point out that these equations are the same as those obtained by Hamilton [6], using a variational principle. The difference is in our use of dimensionless variables. The corresponding dimensionless variational principle is given below for completeness. Hamilton [6, Appendix A] showed that the $(\xi, \tilde{\zeta})$ change

of variables can be used within the functional given in the books of Whitham [16] and Mei [7]. The variational pressure principle is

$$I = \int \int_{\mathbf{R}} \mathcal{L} dx dt, \tag{4.5}$$

where $\mathbf{R}$ is a region in (x,t) space and the dimensionless Lagrangian is given by

$$\mathcal{L} = -\sqrt{\beta} \int_{-\sqrt{\beta}h(x,y)}^{\alpha\sqrt{\beta}\eta(x,t)} \left[\phi_t + \tilde{y} + \frac{\alpha}{2}\left(\nabla\phi \cdot \nabla\phi\right)\right] d\tilde{y}. \tag{4.6}$$

Under the change of variables the functional becomes

$$I = -\sqrt{\beta} \int \int_{\mathbf{R}} \int_{-\sqrt{\beta}}^{\alpha\sqrt{\beta}N} \left[|J| \left(\phi_t + \frac{1}{\sqrt{\beta}}\tilde{y}(\xi,\tilde{\zeta})\right) + \frac{\alpha}{2}\left(\nabla\phi \cdot \nabla\phi\right)\right] d\tilde{\zeta} d\xi dt, \tag{4.7}$$

where now $\nabla = (\partial_\xi, \partial_{\tilde{\zeta}})$. Minimizing this functional, by taking variations in ϕ and N, Hamilton obtained a system of equations similar to the dimensionless one derived above.

5. Asymptotic theory. The asymptotic theory is performed at the level of equations. We are interested in the shallow water/long wave regime, more specifically in the weakly dispersive, weakly nonlinear regime. As mentioned in section 3 it is convenient to normalize in the vertical direction and work with a unit depth channel (cf. scaling S2, Figure 3.2) when performing a long wave (shallow water) asymptotic analysis. The geometry is kept fixed and we can focus on the leading order terms of the asymptotic expansion. Let the origin of the curvilinear coordinate system be at the bottom and define $\tilde{\zeta} = \sqrt{\beta}(\zeta - 1)$. Substitute in the equations above to get

$$\beta\phi_{\xi\xi} \; + \; \phi_{\zeta\zeta} \; = \; 0 \qquad \text{at} \;\; 0 < \zeta < 1 + \alpha N(\xi,t), \tag{5.1}$$

with free surface conditions

$$|J| N_t + \alpha\phi_\xi N_\xi - \frac{1}{\beta}\phi_\zeta = 0, \tag{5.2}$$

$$\eta + \phi_t + \frac{\alpha}{2|J|}\left(\phi_\xi^2 + \frac{1}{\beta}\phi_\zeta^2\right) = 0 \tag{5.3}$$

at $\zeta = 1 + \alpha N(\xi,t)$ and

$$\phi_\zeta = 0 \qquad \text{at} \;\; \zeta = 0. \tag{5.4}$$

As in Whitham [16] consider a power series expansion near the bottom of the channel in the form

$$\phi(\xi,\zeta,t) = \sum_{n=0}^{\infty} \zeta^n \; f_n(\xi,t). \tag{5.5}$$

This function satisfies the scaled Laplace equation when

$$f_{m+2} = \frac{-\beta}{(m+2)(m+1)} \frac{\partial^2 f_m}{\partial\xi^2}.$$

Moreover, the Neumann condition at the bottom is satisfied when all odd terms (i.e., f_{2m+1}) are zero. Hence by denoting $f_0(\xi,t) = f(\xi,t)$ for simplicity, we have

$$\phi(\xi,\zeta,t) = \sum_{n=0}^{\infty} \frac{(-\beta)^n}{(2n)!}\zeta^{2n}\frac{\partial^{2n} f(\xi,t)}{\partial \xi^{2n}}, \tag{5.6}$$

a power series expansion in β. For this reason the $\beta^{1/2}$-scale was called an intermediate scale earlier. The velocity potential, represented by the power series above, satisfies the scaled Laplace equation and the homogeneous Neumann condition along the bottom. Up to this point we have done exactly the same steps as in Whitham's book [16] for a flat channel in Cartesian coordinates.

We must now satisfy the free surface conditions. At $\zeta = \zeta_{FS} \equiv 1 + \alpha N(\xi,t)$ we have that

$$\phi_\xi(\xi,\zeta_{FS},t) = f_\xi - \frac{\beta}{2}(1+\alpha N)^2 f_{\xi\xi\xi} + O(\beta^2), \tag{5.7}$$

$$\phi_\zeta(\xi,\zeta_{FS},t) = -\beta(1+\alpha N) f_{\xi\xi} + \frac{\beta^2}{3!}(1+\alpha N)^3 f_{\xi\xi\xi\xi} + O(\beta^3), \tag{5.8}$$

and

$$\phi_t(\xi,\zeta_{FS},t) = f_t - \frac{\beta}{2}(1+\alpha N)^2 f_{\xi\xi t} + O(\beta^2). \tag{5.9}$$

Note that both (new) free surface conditions have a time dependent coefficient. This is due to the presence of the Jacobian, of the time independent change of coordinates, evaluated at the time dependent free surface. Within the weakly nonlinear, weakly dispersive approximation considered, it is possible to eliminate this time dependence as will be shown below. Time independent coefficients are better for analysis and computations. As a first step, in this direction, it is worth noticing that at the smooth free surface $\tilde{\zeta}_{FS} = \alpha\sqrt{\beta}N(\xi,t)$ the Jacobian is

$$|J|(\xi,t) = \tilde{y}_\xi^2(\xi,\tilde{\zeta}_{FS}) + \tilde{y}_{\tilde{\zeta}}^2(\xi,\tilde{\zeta}_{FS})$$

and by the Taylor polynomial formula

$$|J|(\xi,t) = \tilde{y}_{\tilde{\zeta}}^2(\xi,0) + \alpha^2 R_J(\xi,\tilde{\zeta}_M) = M(\xi)^2 + O(\alpha^2), \quad 0 < |\tilde{\zeta}_M| < |\tilde{\zeta}_{FS}|. \tag{5.10}$$

The Jacobian can be well approximated by an $O(1)$ time independent coefficient. The time dependent correction term is $O(\alpha^2)$ due to the fact that the curvilinear coordinate system is symmetric about $\tilde{y} = \tilde{\zeta} = 0$. There are no $O(\alpha)$ terms. For the same reason, approximating $\tilde{\zeta}(x,\tilde{y}_{FS})$ in $\tilde{y}$ leads to

$$N(\xi,t) = \frac{1}{M(\xi)}\eta(x(\xi),t) + \alpha^2\beta R_N(\xi,\tilde{y}_M), \quad 0 < |\tilde{y}_M| < |\tilde{y}_{FS}|. \tag{5.11}$$

Substituting expressions (5.7)–(5.11) in the free surface conditions, we get

$$\eta + f_t - \frac{\beta}{2} f_{\xi\xi t} + \frac{\alpha}{2M^2(\xi)} f_\xi^2 = O(\alpha\beta, \beta^2),$$

$$M(\xi)\ \eta_t + \left[\left(1+\frac{\alpha}{M(\xi)}\eta\right) f_\xi\right]_\xi - \frac{\beta}{6} f_{\xi\xi\xi\xi} = O(\alpha^2, \alpha\beta, \beta^2).$$

The variable coefficients are time independent and depend only on $\tilde{y}_{\tilde{\zeta}}(\xi, 0)$ as mentioned before in the conformal mapping section.

As in Whitham [16] it is useful to work with a conveniently averaged velocity component. From (5.7) we have that

$$\phi_\xi(\xi, \zeta_{FS}, t) \equiv u(\xi, \zeta_{FS}, t) = \tilde{u}(\xi, t) - \frac{\beta}{2}\zeta_{FS}^2 \tilde{u}_{\xi\xi}(\xi, t) + O(\beta^2),$$

where $\tilde{u} \equiv f_\xi$ is the "slip-velocity" along the bottom topography. In analogy with the σ-coordinates used in atmospheric flows [5], we call ϕ_ξ the terrain-following velocity component [3]. Its transversal average (along a ξ level-curve) is defined as

$$U(\xi, t) \equiv \frac{1}{\zeta_{FS}} \int_0^{\zeta_{FS}} \phi_\xi(\xi, \zeta, t) d\zeta. \tag{5.12}$$

In physical space this integral represents a weighted average of the terrain-following velocity components. We provide further insight about the averaging in Figure 5.1. Therefore using the potential's power series representation,

$$U(\xi, t) = \sum_{n=0}^{\infty} \frac{(-\beta)^n}{(2n)!} \left[\frac{1}{\zeta_{FS}} \int_0^{\zeta_{FS}} \zeta^{2n} d\zeta\right] \frac{\partial^{2n+1} f(\xi, t)}{\partial \xi^{2n+1}} = \tilde{u} - \frac{\beta}{6}\tilde{u}_{\xi\xi} + O(\alpha\beta, \beta^2).$$

Differentiate the first free surface condition with respect to ξ and substitute for depth-averaged velocity through the expansion

$$\tilde{u} = U + \frac{\beta}{6} U_{\xi\xi} + O(\alpha\beta, \beta^2).$$

We get a *terrain-following Boussinesq system*

$$\begin{cases} M(\xi)\eta_t + \left[(1 + \alpha\eta/M(\xi))\, U\right]_\xi = 0, \\ \\ U_t + \eta_\xi + \frac{\alpha}{2}\left[U^2/M^2(\xi)\right]_\xi - \frac{\beta}{3} U_{\xi\xi t} = 0. \end{cases} \tag{5.13}$$

This is a weakly nonlinear, weakly dispersive system with variable coefficients, all depending on $M(\xi; \sqrt{\beta}, \gamma)$. The $O(\alpha^2, \alpha\beta, \beta^2)$ truncation terms are small and were dropped. This was not the case when working with a Cartesian coordinate system. The truncation terms could be large. Note that working with the terrain-following velocity we generated (formally) a converging asymptotic expansion. This is true even in the presence of rapidly varying topographies ($\gamma \ll 1$), as opposed to the original (nonconverging) expansion in Cartesian coordinates. Our physical interpretation is that the Boussinesq system with weighted terrain-following velocities averages is more representative, as an asymptotic simplification of the full set of equations. Namely, the bulk of the terrain-following velocity components do not deviate too much from their average as opposed to the horizontal components. Figure 5.1 corroborates this interpretation from a graphical perspective. The weighted velocity averages are performed over the ξ level-curves (vertical curves in Figure 5.1; curves over which ζ varies, but ξ is fixed). Note that even with high graphing resolution (Figure 5.1(b)) the effective flow domain penetrates very little into the narrow valleys. The bulk of the average

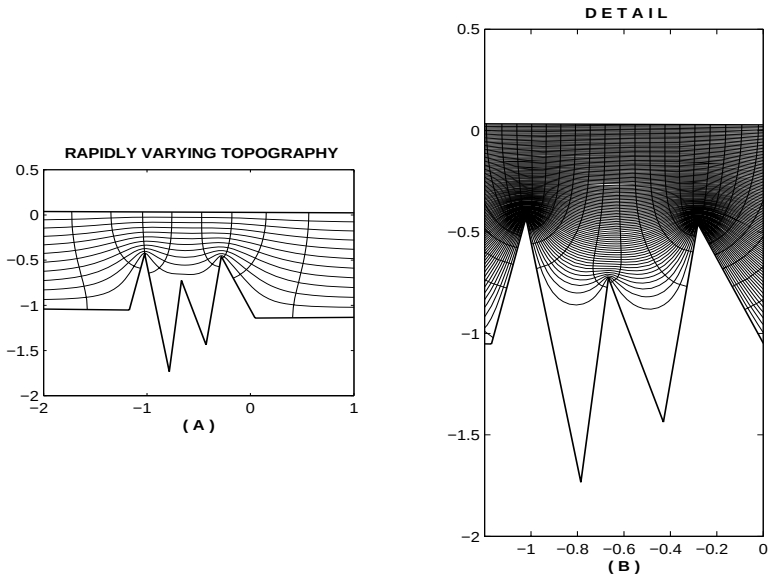


FIG. 5.1. *The effective flow domain. The horizontal resolution is equal to* (a) 10 ζ *level-curves,* (b) 100 ζ *level-curves. This figure was generated using the SC-Toolbox* [4].

is performed over the smoothly varying "outer" terrain-following velocities. In other words the contribution from the highly corrugated valleys is very small: 10% below the lowest ζ level-curve in 5.1(a) and 1% in 5.1(b). Clearly the outer flow prevails in the weighted average (5.12). In contrast if the topography is slowly varying the weights for the inner valley regions will be larger as indicated in Figure 4.1.

Regarding the use of averaged velocities, we do this change of dependent variables because standard long wave models use depth-averaged velocities. Moreover, the use of this dependent variable improves the stability for the high wavenumber regime of the Boussinesq system.

When the channel is flat the *metric term* ($M(\xi) = 1+m(\xi)$) reduces to one and we recover the standard dimensionless Boussinesq system [16]. Note that system (5.13) is different from that obtained by Hamilton (cf. equations (A14) and (A15) in [6]), which has powers of the metric term and an η_{tt} term. A comparison with standard Boussinesq systems is presented in the appendix of the present paper. When linearized they all have the same dispersion relation.

6. Applications and conclusions. We now point out some important features regarding applications with the terrain-following Boussinesq system:

(i) The derivatives of the topography-related coefficients are all well defined, even in the presence of corners or steps along the bottom. The metric term $M(\xi)$ is an analytic function. This is not the case in the classical Boussinesq equations (cf. Appendix A) as pointed out in Hamilton's work [6].

(ii) Under the curvilinear coordinate formulation, the *nonlinear shallow water equations* are viewed as an $O(\alpha, \sqrt{\beta})$ approximation to the full potential theory equations. Dropping the $O(\beta)$ term from system (5.13) above, we are left with a hyperbolic

system, but we are still keeping an $O(\sqrt{\beta})$ term inside the metric term (cf. (3.8)):

$$\begin{cases} M(\xi;\sqrt{\beta},\gamma)\eta_t + \left[\left(1+\alpha\eta/M(\xi;\sqrt{\beta},\gamma)\right)U\right]_\xi = 0, \\ \\ U_t + \eta_\xi + \frac{\alpha}{2}\left[U^2/M^2(\xi;\sqrt{\beta},\gamma)\right]_\xi = 0. \end{cases} \tag{6.1}$$

This has a great advantage in opposition to the $\beta \to 0$ limit used in the standard shallow water derivations. In the limiting $\beta \to 0$ case derivatives at corners along the bottom are not defined. In the present formulation consistency is preserved between the two formulations since in the vanishing β limit the kernel of the metric term goes to a Dirac delta function (cf. (3.7)) and singularities at the bottom are recovered. But retaining the intermediate $O(\sqrt{\beta})$ term leads to a (hyperbolic) shallow water system having a smooth, well-defined coefficient, even if the topography has corners. The degree of smoothing/averaging performed by the metric-kernel is controlled by $\sqrt{\beta}$ and γ (cf. (3.8)). This last point was not observed in Hamilton's work [6] since the metric term was not scaled in terms of depth/wavelength scales. By averaging we mean cases where the bottom is rapidly varying ($\gamma \ll 1$).

A schematic example is presented in Figure 6.1. The conformal mapping was computed numerically [12] as well as the metric term. We considered periodic mountain ranges having identical triangular mountains. The height of each triangle is equal to 0.5 and the base equal to 2.0 units. A natural question to ask is, What is the underlying value of β in this example? As will be discussed below, the averaging effect is related to the ratio between the scales $\sqrt{\beta}$ and γ. This was pointed out in Nachbin [12, p. 366] where a rule was devised in order to incorporate both scales. In order to incorporate these two scales, the Schwarz–Christoffel mapping was performed on mountains varying on the "modified" length scale

$$\tilde{l}_b \equiv l_b/(\sigma\sqrt{\beta}).$$

Note the difference of notations from [12], $l_p \equiv \sigma$ and ($\beta = h_0/\lambda$)—there is the $\sqrt{\beta}$ here. In other words

$$\sqrt{\beta} = \gamma/\tilde{l}_b.$$

Since for the triangular mountains $\tilde{l}_b = 2$, then we have that $\beta = 0.25\gamma^2$. A typical γ for rapidly varying topographies is 0.1 (10 mountain peaks per pulse width).

In the first example we contrast the metric term for a periodic mountain range with 5 mountains with that for a mountain range (of the same period) with 7 mountains. We clearly observe the exponential adjustment mentioned earlier. We also observe the smoothing along the sharp summit of the triangular mountains and the averaging as we put the mountains closer together (cf. Figure 6.2). In this case we have $\tilde{l}_b = 1$ and therefore $\beta = \gamma^2$. As indicated above averaging (in a homogenization sense) takes place as soon as the mountain scale parameter γ is smaller than the support of the sech^2-kernel, controlled by $\beta^{1/2}$ (see (3.7)). Note that the topography's derivative scales like $1/\gamma$ while the metric term's derivative always scales like $1/\sqrt{\beta}$. One of the advantages of this formulation is that the latter is independent of γ. This is a useful fact in the discussion that follows.

(iii) It is worth pointing out that first order derivatives of the metric term are produced. These terms are $O(1/\sqrt{\beta})$ (cf. (3.8)) and are generated by the N_ξ term, in the kinematic condition, and by differentiating Bernoulli's Law with respect to ξ. In particular, these $O(1/\sqrt{\beta})$ terms are present in the $O(\alpha\beta, \beta^2)$ terms that were

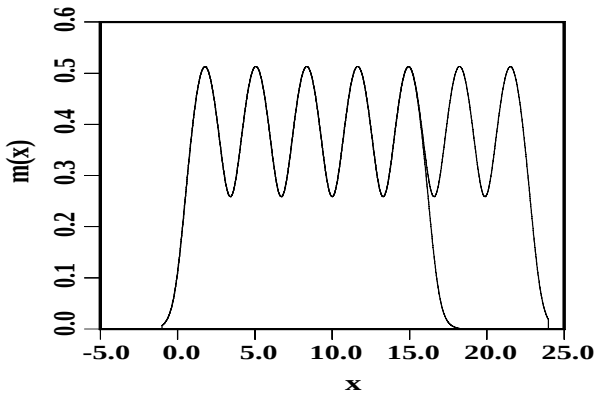


FIG. 6.1. *The metric term* $m(\xi) = M(\xi) - 1$ *for two periodic topographies. One was generated by* 5 *triangular mountains along the bottom while the other by* 7 *triangular mountains. The two metric terms match very well over the first* 5 *mountains, confirming the exponentially fast adjustment mentioned earlier. The* $\beta^{1/2}$*-scale plays a role in this adjustment.*

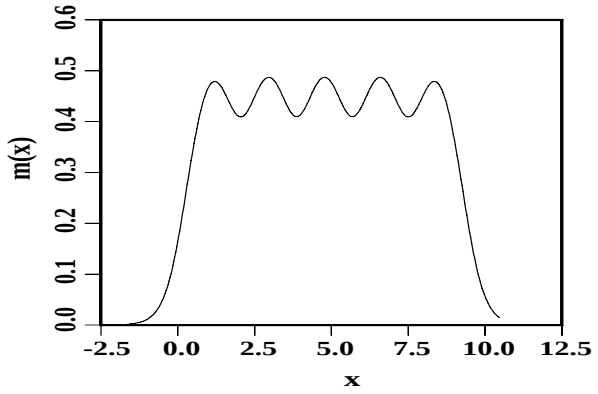


FIG. 6.2. *Averaging observed through the metric term* $m(\xi) = M(\xi) - 1$*. The triangular mountains are closer together, i.e., varying on a faster scale. The base of each triangle is* $1/2$ *of those in Figure* 6.1*.*

dropped from Bernoulli's Law. Nevertheless, one should not be concerned with the loss of validity of the asymptotics. The $O(1/\sqrt{\beta})$ term does not alter the ordering of the terms presented in the asymptotic analysis above. It can be easily verified that when the equation

$$\eta + f_t - \frac{\beta}{2} f_{\xi\xi t} + \frac{\alpha}{2M^2(\xi)} f_\xi^2 = O(\alpha\beta, \beta^2)$$

is differentiated with respect to ξ, the $1/\sqrt{\beta}$ term will only affect the $O(\alpha\beta)$ terms. For example, if we take $O(\alpha\beta)/\sqrt{\beta}$ and compare it to the leading $O(\alpha)$ nonlinearity term, we easily note that reordering takes place only if $\sqrt{\beta} > 1$, which is not the long wave regime we started with. Hence as long as $0 < \beta \ll 1$ the equations derived are valid weakly nonlinear, weakly dispersive approximations to the nonlinear potential theory equations. Namely they are valid as the leading order corrections in α and in β. The question of including the next order terms (in either parameter) was not

relevant to our application of interest [9]. Perhaps there are applications where these terms might play a role. We have not attempted to identify these regimes. Having set that $\beta > 0$ guarantees that the variable coefficients can be differentiated in the standard way. Delta functions do not arise at discontinuities of the topography.

(iv) Regarding applications, these new equations enable the study of the interaction of linear and nonlinear long surface waves with rapidly varying features of the topography. To our knowledge up to the present date, only small amplitude or mild slope topographies have been considered for Boussinesq-type models. The present model expands the regime of applicability and is consistent with previous Cartesian formulations in the case where $\gamma \gg 1$. In this regime the conformal mapping is nearly the identity map at the free surface level.

Many new theoretical and numerical problems can now be addressed with this new curvilinear formulation. A particular theoretical problem that we investigated through this new model is the O'Doherty–Anstey approximation of weakly dispersive waves. This theory characterizes the apparent diffusion of transmitted pulses due to disordered multiple scattering. This theory, initially developed for acoustic waves [1], was generalized to weakly dispersive waves by Muñoz Grajales and Nachbin [9]. Using the terrain-following Boussinesq system, we obtained expressions that capture both the attenuation of the transmitted pulse as well as the forward scattering radiation. Numerical validation experiments include the apparent diffusion of solitary waves in the presence of a highly disordered topography. We are currently also investigating an interesting numerical problem regarding the use of the terrain-following formulation [10]. The classical system (A.2) contains a term difficult to handle numerically in the presence of a topography. This is because the topography-related coefficients multiply the terms with higher order derivatives. In the curvilinear formulation the higher order term has a constant coefficient. We have obtained numerical evidence [10] that the "repositioning" of topography related terms (namely the metric terms) leads to improvements in the performance of Boussinesq solvers. We have also obtained results regarding the time reversed refocusing (i.e., waveform inversion) for weakly dispersive waves [10].

The Boussinesq system (5.13) can be cast into a second order differential equation. Take the system at the stage where we had

$$\eta + f_t - \frac{\beta}{2} f_{\xi\xi t} + \frac{\alpha}{2M^2(\xi)} f_\xi^2 = O(\alpha^2, \alpha\beta, \beta^2),$$

$$M(\xi)\ \eta_t + \left[\left(1 + \frac{\alpha}{M(\xi)}\eta\right) f_\xi\right]_\xi - \frac{\beta}{6} f_{\xi\xi\xi\xi} = O(\alpha^2, \alpha\beta, \beta^2).$$

From these equations we have that

$$\eta_t = -f_{tt} + O(\alpha, \beta), \quad M(\xi)\eta_t = -f_{\xi\xi} + O(\alpha, \beta), \quad \text{and} \quad f_{\xi\xi} = M(\xi) f_{tt} + O(\alpha, \beta).$$

These approximations are used after we calculate the time derivative of the first equation, multiply it by $M(\xi)$, and subtract it from the second equation to get

$$f_{\xi\xi} - M f_{tt} + \frac{\beta}{3} M f_{\xi\xi tt} - \frac{\alpha}{M}\left[f_\xi^2 + \frac{M}{2} f_t^2\right]_t + \alpha\left(\frac{M'}{M^2}\right) f_t f_\xi = O(\alpha^2, \alpha\beta, \beta^2). \tag{6.2}$$

By dropping the $O(\alpha^2, \alpha\beta, \beta^2)$ term we get the second order Boussinesq equation governing the velocity potential along the bottom topography. It is interesting to note

that in the Boussinesq system (5.13) no new terms were generated by the curvilinear coordinates. The only novelty is the presence of the metric term $M(\xi)$. Nevertheless, a new term appears when the system is reduced to a second order differential equation. The new term contains the derivative of the metric term. Again, in the absence of a topography, this differential equation reduces to the standard second order Boussinesq equation [7].

Appendix A. Boussinesq models and their linearization. Consider Peregrine's model (Mei [7, p. 512]):

$$\eta_t + \nabla \cdot [(h+\eta)U] = 0, \tag{A.1}$$

$$U_t + U \cdot \nabla U + g\nabla\eta = \frac{h}{2}\nabla\left[\nabla \cdot (hU_t)\right] - \frac{h^2}{6}\nabla\left[\nabla \cdot U_t\right].$$

Restricting to one-dimensional flows it becomes

$$\eta_t + \partial_x [(h+\eta)U] = 0, \tag{A.2}$$

$$U_t + UU_x + g\eta_x = \frac{h}{2}\left[\partial_x^2(hU_t)\right] - \frac{h^2}{6}U_{xxt}.$$

Linearize this system by letting $\tilde{\eta} = \varepsilon\eta$ and $\tilde{U} = \varepsilon U$. Drop the $\tilde{}$ to get

$$\eta_t + (hU)_x = 0, \tag{A.3}$$

$$U_t + g\eta_x = \frac{h}{2}(hU_t)_{xx} - \frac{h^2}{6}U_{xxt}.$$

Moreover, if the bottom is flat ($h(x) \equiv h_0$),

$$\eta_t + h_0 U_x = 0, \tag{A.4}$$

$$U_t + g\eta_x = \frac{h_0^2}{3}U_{xxt}.$$

The dimensionless Boussinesq equation derived in Whitham (page 467 of [16], using the dispersion relation) is

$$\eta_t + U_x = 0, \tag{A.5}$$

$$U_t + \eta_x = -\frac{\beta}{3}\eta_{xtt}.$$

Moreover, the linearized terrain-following Boussinesq equation is

$$M(\xi)\eta_t + U_\xi = 0, \tag{A.6}$$

$$U_t + \eta_\xi = \frac{\beta}{3}U_{\xi\xi t}.$$

All these Boussinesq systems share *the same dispersion relation*:

$$\omega^2 = \frac{k^2}{1 + \frac{1}{3}\beta k^2}, \tag{A.7}$$

which is stable as $k \to \infty$. In (A.4) the variables are in their original dimensions. Thus, for the sake of comparison, we set $gh_0 = 1$ and $h_0^2 = \beta$. This means that the characteristic wavelength is taken to be equal to one.

REFERENCES

[1] R. Burridge and L. Berlyand, *The accuracy of the O'Doherty-Anstey approximation for wave propagating in highly disordered stratified media*, Wave Motion, 21 (1995), pp. 357-373.

[2] R. Camassa, D.D. Holm, and C.D. Levermore, *Long–time shallow–water equations with a variable bottom*, J. Fluid Mech., 349 (1997), pp. 173–189.

[3] T.L. Clark, *A small scale dynamical model using a terrain–following coordinate transformation*, J. Comput. Phys., 24 (1977), pp. 136–215.

[4] T. Driscoll, *The Schwarz-Christoffel Toolbox for MATLAB*, http://www.math.udel.edu/~driscoll/software/SC.

[5] A.E. Gill, *Atmosphere–Ocean Dynamics*, Academic Press, New York, 1982.

[6] J. Hamilton, *Differential equations for long-period gravity waves on a fluid of rapidly varying depth*, J. Fluid Mech., 83 (1977), pp. 289–310.

[7] C.C. Mei, *The Applied Dynamics of Ocean Surface Waves*, John Wiley, New York, 1983.

[8] P. Milewski, *Long wave interaction over varying topography*, Phys. D, 123 (1998), pp. 36–47.

[9] J.C. Muñoz Grajales and A. Nachbin, *Dispersive wave attenuation due to orographic forcing*, SIAM J. Appl. Math., submitted.

[10] J.C. Muñoz Grajales and A. Nachbin, *Dispersive Pulse Stabilization and Solitary Wave Refocusing*, manuscript, 2002..

[11] A. Nachbin, *Modelling of Water Waves in Shallow Channels*, WIT Press, Southampton, U.K., 1993.

[12] A. Nachbin, *The localization length of randomly scattered water waves*, J. Fluid Mech., 296 (1995), pp. 353–372.

[13] A. Nachbin and G.C. Papanicolaou, *Water waves in shallow channels of rapidly varying depth*, J. Fluid Mech., 241 (1992), pp. 311–332.

[14] R.R. Rosales and G.C. Papanicolaou, *Gravity waves in a channel with a rough bottom*, Stud. Appl. Math., 68 (1983), pp. 89–102.

[15] H.A. Schäffer and P.A. Madsen, *Further enhancements of Boussinesq-type equations*, Coastal Eng., 26 (1995), pp. 1–14.

[16] G.B. Whitham, *Linear and Nonlinear Waves*, John Wiley, New York-London-Sydney, 1974.

SIAM J. APPL. MATH.
Vol. 63, No. 3, pp. 923–936

ON THE CONVEXITY AND RISK-SENSITIVITY OF THE PRICE OF AMERICAN INTEREST RATE DERIVATIVES*

LUIS H. R. ALVAREZ†

Abstract. We consider the form and sensitivity to risk of the price of perpetual American interest rate derivatives for a broad class of one-factor diffusion models of interest rates. We first present, in terms of the infinitesimal coefficients of the underlying interest rate dynamics, a set of usually satisfied conditions under which the value of the contingent claim is convex, at least on the set where exercising the contract is suboptimal. In line with previous parametrized models considering the valuation of perpetual interest rate derivatives, we find that given our general conditions, the convexity of the exercise payoff is preserved under rational valuation. Consequently, we are able to establish a set of typically satisfied conditions under which increased volatility unambiguously increases the price of the claim and postpones rational exercise by expanding the region where exercising the claim is suboptimal.

Key words. term structure, interest rate derivatives, minimal excessive mappings, convexity, optimal stopping

AMS subject classifications. 60J60, 60G40, 62L15, 60H30, 93E20

PII. S0036139901384674

1. Introduction.

The sign of the relationship between volatility and the value of a contingent contract depends on the form of the value as a function of the current state of the underlying asset. Volatility, being a second-order property, affects the price of the contingent claim through the quadratic variation process and, therefore, increases or decreases the price depending on whether the price is convex or concave. As was proven in [22] and [24], the convexity of the exercise payoff is preserved under risk-neutral valuation for most path-independent claims. Consequently, these authors found that the sign of the relationship between volatility and the arbitrage free price of a contingent contract is positive and, therefore, that increased volatility increases the arbitrage free price of path-independent claims. This property was shown to be valid also for the American-type path-independent contingent claims in both [22] (in the absence of dividends) and [1] (in the presence of dividends). In [2], this result was subsequently shown to be valid for a broad class of path-dependent European interest rate derivatives as well. In [27], the convexity of the price and the positivity of the relationship between volatility and the exercise incentives of a rational investor were established for a special case of the Cox–Ingersoll–Ross model. A similar result was found in [15], which considered the valuation of perpetual variable loan contracts subject to an explicit parametrized short rate model. However, before now not much has been done in considering the robustness of the qualitative results on American interest rate derivatives obtained by applying explicit parametrized interest rate models (cf. [15], [27], and [30]). This is somewhat surprising in light of the large extent of the literature on term structure models of interest rates (see, for example, [7], [8, chapters 15–20], [9], [10], [12], [13, chapter 19], [14], [15], [16], [17], [19], [21], [23], [25],

*Received by the editors February 7, 2001; accepted for publication (in revised form) August 8, 2002; published electronically January 23, 2003. This research was supported by the Foundation for the Promotion of the Actuarial Profession (Aktuaaritoiminnan Kehittämissäätiö) and the Yrjö Jahnsson Foundation.

http://www.siam.org/journals/siap/63-3/38467.html

†Department of Economics, Quantitative Methods in Management, Turku School of Economics and Business Administration, FIN-20500 Turku, Finland (luis.alvarez@tukkk.fi).

[26], [27], [30], [31], [32], [34], [35], and [36]) and in light of the fact that this class of valuation problems arises frequently in real option studies considering irreversible investment in the presence of interest rate uncertainty (cf. [18, pp. 48–51] and [27]) and in the valuation of perpetual variable rate loan contracts (cf. [15]). Moreover, as is stated in [30], "...volatility is a key variable in pricing contingent claims such as interest rate options and bonds." Consequently, the sign of the relationship between volatility and the value of interest rate derivatives is an important factor characterizing the exercise incentives of a rational investor as a function of the volatility of the underlying stochastic interest rate dynamics.

Given these arguments, we plan to consider in this study both the form and the comparative static properties of the value of potentially infinitely lived American interest rate derivatives for a broad class of one-factor models of the short rate. For the sake of generality we model the short rate as a linear diffusion with known functional infinitesimal characteristics (i.e., known drift and volatility coefficient). By then applying the classical theory of diffusions, we first demonstrate that given a set of usually satisfied continuity and monotonicity conditions (which are satisfied by, among others, the Black–Karasinski, Cox–Ingersoll–Ross, Dothan, Merton, and Vasiček models), the minimal excessive mappings for the considered one-factor short rate diffusion model are globally convex. Since any nontrivial excessive mapping for the diffusion can be expressed in terms of the minimal ones and the value of an American contingent contract is excessive, we find that the value is convex, at least on the continuation region where waiting is optimal. Consequently, we find that increased volatility increases or leaves unchanged the pre-exercise value of these claims and, therefore, postpones the rational exercise of the contract by expanding or leaving unchanged the continuation region where waiting is optimal. In other words, we are able to demonstrate that increased volatility unambiguously increases the required exercise premium of a rational investor. The main reason for this finding is shown to be the fact that while increased volatility may (as it typically does) increase the value of the underlying payoff, it simultaneously also increases the value of holding the option alive. Since the latter effect dominates the former, we find that the net effect of increased volatility on waiting is unambiguously positive. Moreover, since the curvature of the minimal excessive mappings is independent of the exercise payoff, we also observe that the sign of the relationship between volatility and rational exercise is a process-specific property independent of the form of the underlying payoff. (See [1] for a similar observation in the path-independent case.)

The contents of this study are as follows. In section two we present the one-factor interest rate dynamics and the considered valuation problem (an optimal stopping problem). In section three we then consider the impact of increased volatility on the value of the interest rate derivative. Finally, our principal results are illustrated explicitly in section four.

2. Valuation and optimal stopping. As is well known from the literature on interest rate derivatives, the term structure of interest rates is entirely determined by specifying the dynamic behavior of the short rate of interest under the equivalent martingale measure $\mathbb{Q}$ [7], [8, chapters 16–17], [13, chapter 19], [20], [21]. In line with this argument, consider now the case where the interest rate process $\{r(t); t \geq 0\}$ is described under the risk-neutral measure $\mathbb{Q}$ on the state-space $(a, b) = \mathcal{I} \subseteq \mathbb{R}$ by the time homogeneous stochastic differential equation

$$dr(t) = \mu(r(t))dt + \sigma(r(t))d\hat{W}(t), \quad r(0) = r, \tag{1}$$

where $\hat{W}(t)$ is $\mathbb{Q}$-Brownian motion and the mappings $\mu : \mathcal{I} \mapsto \mathbb{R}$ and $\sigma : \mathcal{I} \mapsto \mathbb{R}_+$ are given sufficiently smooth mappings (at least continuous) for which both the existence and the uniqueness of a solution for (1) are guaranteed (cf. [11, pp. 46–47]). In accordance with most applications, we assume that $\sigma(r) > 0$ for all $r \in \mathcal{I}$ and that the upper boundary $b > 0$ of the state-space $\mathcal{I}$ is unattainable for the diffusion r (typically, $b = \infty$). Thus, even while the short rate may be expected to increase, it is never expected to attain the maximal possible state b in finite time. We will also assume that the lower boundary $a \leq 0$ is either natural (as 0 is for the Dothan and Merton models and $-\infty$ is for the Vasiček model), entrance, exit, or regular. It is also worth pointing out that the boundary behavior of the short rate process $r(t)$ typically depends on the precise parametrization of the model. For example, depending on the relative sizes of the parameters $\kappa, \theta, \sigma \in \mathbb{R}_+$, 0 may be either regular, exit, or entrance for the familiar Cox–Ingersoll–Ross model:

$$dr(t) = \kappa(\theta - r(t))dt + \sigma\sqrt{r(t)}d\hat{W}(t), \quad r(0) = r.$$

Given these assumptions, it is our purpose now to consider the optimal stopping problem

$$V(r) = \sup_{\tau} E_r \left[e^{-\int_0^\tau r(s)ds} \Phi(r(\tau)) \right], \tag{2}$$

where τ is an arbitrary $\mathcal{F}_t$-stopping time and $\Phi : \mathcal{I} \mapsto \mathbb{R}_+$ is a continuous mapping satisfying the intuitively clear boundedness condition

$$E_r \left[e^{-\int_0^t r(s)ds} \Phi(r(t)) \right] < \infty \tag{3}$$

for all $(t, r) \in \mathbb{R}_+ \times \mathcal{I}$. That is, we plan to consider the determination of the value and rational exercise price of a perpetual American contingent claim with exercise payoff $\Phi(r)$. This type of valuation problems arise frequently in studies considering either the pricing of American-type interest rate derivatives or the impact of interest rate uncertainty on real investment opportunities (cf. [27]). It is worth emphasizing that the value $V(r)$ should be viewed as the value of a derivative on a traded asset, since interest rates themselves are not traded. Put somewhat differently, typically $V(r)$ constitutes the value of a compound contingent contract. For example, if $\Phi(r) = (p(r, T) - c)^+$, where $p(r, T)$ is the value of a zero coupon bond maturing T periods from exercise (a T-bond) and $c \in (0, 1)$ is a known strike price of the option (cf. [27]), then $V(r)$ is the value of a perpetual call option on a T-bond. However, since our approach admits more complex contracts as well, we stick to the general notation. Moreover, since the time horizon of the valuation may be finite (as for the Cox–Ingersoll–Ross model), we observe that (2) constitutes a valuation subject to a potentially finite expiration date. In order to analyze the valuation problem (2), we first have to consider the form of the fundamental solutions $\psi : \mathcal{I} \mapsto \mathbb{R}_+$ and $\varphi : \mathcal{I} \mapsto \mathbb{R}_+$ of the ordinary second-order differential equation $(\mathcal{A}u)(r) - ru(r) = 0$, where

$$\mathcal{A} = \frac{1}{2}\sigma^2(r)\frac{d^2}{dr^2} + \mu(r)\frac{d}{dr} \tag{4}$$

denotes the differential operator representing the infinitesimal generator of r in the domain of $\mathcal{A}$. It is known from the classical theory of diffusions that $\psi(r)$ is monotonically increasing and that $\varphi(r)$ is monotonically decreasing on $\mathcal{I}$; that $\lim_{r \to b} \psi'(r)/S'(r) =$

∞ and $\lim_{r\to b}\varphi'(r)/S'(r)=0$, where

$$S'(r)=\exp\left(-\int^{r}\frac{2\mu(y)}{\sigma^2(y)}dy\right)$$

denotes the density of the scale function S of r; and that $\psi'(r)\varphi(r)-\varphi'(r)\psi(r)=BS'(r)$, where B denotes the constant Wronskian of the solutions. (For a thorough characterization of the fundamental solutions, see [11, pp. 18–19]; see also [28, chapter 4 and especially section 4.6].) Moreover, $\psi(r)$ and $\varphi(r)$ are minimal in the sense that any nontrivial excessive function for the interest rate process $r(t)$ killed at the rate r can be written in terms of these mappings (i.e., there is an integral representation for nontrivial excessive mappings in terms of the mappings $\psi(r)$ and $\varphi(r)$; see [11, p. 32]), and for all $r,q\in\mathcal{I}$ we have

$$E_r\left[e^{-\int_0^{\tau(q)}r(t)dt}\right]=\min\left(\frac{\psi(r)}{\psi(q)},\frac{\varphi(r)}{\varphi(q)}\right),$$

where $\tau(q)=\inf\{t\geq 0: r(t)=q\}$ denotes the first hitting time of the short rate process to the state $q\in\mathcal{I}$. Thus, the price of a zero coupon bond expiring at a random date $\tau(q)$ can be expressed in terms of either $\psi(r)$ or $\varphi(r)$, depending on whether the current short rate is below or above the state q.

Define now the mapping $\eta:\mathcal{I}\mapsto\mathbb{R}$ measuring the *net percentage growth rate of the short rate process killed at the rate* r as

$$\eta(r)=\frac{\mu(r)}{r}-r.$$

Our main results characterizing the form of the minimal excessive mappings for the interest rate process $r(t)$ are now summarized in our next theorem.

THEOREM 1. *Assume that the mapping $\eta(r)$ is nonincreasing on $\mathcal{I}$. Then the decreasing fundamental solution $\varphi(r)$ is convex on $\mathcal{I}$. The increasing fundamental solution $\psi(r)$ is convex on $\mathcal{I}$ as well if either a is unattainable for $r(t)$ or a is attainable for $r(t)$ and $\lim_{r\downarrow a}\eta(r)\leq 0$.*

Proof. Applying Dynkin's theorem to the linear mapping $r\mapsto r$ yields (cf. [1] and [3])

$$E_r\left[e^{-\int_0^{\tau(\hat{a},\hat{b})}r(s)ds}r(\tau(\hat{a},\hat{b}))\right]=r+E_r\int_0^{\tau(\hat{a},\hat{b})}e^{-\int_0^{s}r(t)dt}r(s)\eta(r(s))ds, \tag{5}$$

where $\tau(\hat{a},\hat{b})=\{t\geq 0: r(t)\notin(\hat{a},\hat{b})\}$ denotes the first exit time of the interest rate process from the bounded open set $(\hat{a},\hat{b})\subset\mathcal{I}$. Define now the functionals $u_1:\mathcal{I}\mapsto\mathbb{R}$ and $u_2:\mathcal{I}\mapsto\mathbb{R}$ as

$$u_1(r)=E_r\left[e^{-\int_0^{\tau(\hat{a},\hat{b})}r(s)ds}r(\tau(\hat{a},\hat{b}))\right]\quad\text{and}\quad u_2(r)=E_r\int_0^{\tau(\hat{a},\hat{b})}e^{-\int_0^{s}r(t)dt}r(s)\eta(r(s))ds.$$

As is shown in [29, pp. 199–224] (see also chapter 9 in [33]), the functional $u_1(r)$ is on $(\hat{a},\hat{b})$ the solution of the boundary value problem $(\mathcal{A}u_1)(r)=ru_1(r), u_1(\hat{a})=\hat{a}, u_1(\hat{b})=\hat{b}$, and the functional $u_2(r)$ is the solution of the boundary value problem $(\mathcal{A}u_2)(r)-ru_2(r)+r\eta(r)=0, u_2(\hat{a})=u_2(\hat{b})=0$. Consequently, we find that (5) can

be rewritten as

$$\hat{a}\frac{\tilde{\varphi}(r)}{\tilde{\varphi}(\hat{a})}+\hat{b}\frac{\tilde{\psi}(r)}{\tilde{\psi}(\hat{b})}=r+\tilde{B}^{-1}\tilde{\varphi}(r)\int_{\hat{a}}^{r}\tilde{\psi}(y)y\eta(y)m'(y)dy + \tilde{B}^{-1}\tilde{\psi}(r)\int_{r}^{\hat{b}}\tilde{\varphi}(y)y\eta(y)m'(y)dy, \tag{6}$$

where $m'(r) = 2/(\sigma^2(r)S'(r))$ denotes the density of the speed measure m of the diffusion r,

$$\tilde{\varphi}(r)=\varphi(r)-\frac{\varphi(\hat{b})}{\psi(\hat{b})}\psi(r),\quad \tilde{\psi}(r)=\psi(r)-\frac{\psi(\hat{a})}{\varphi(\hat{a})}\varphi(r),$$

and $\tilde{B} = (1-\psi(\hat{a})\varphi(\hat{b})/(\psi(\hat{b})\varphi(\hat{a})))B$ denotes the constant (with respect to the scale) Wronskian of the functions $\tilde{\psi}(r)$ and $\tilde{\varphi}(r)$. Differentiating (6) with respect to r and reordering terms then yield

$$1=\hat{a}\frac{\tilde{\varphi}'(r)}{\tilde{\varphi}(\hat{a})}+\hat{b}\frac{\tilde{\psi}'(r)}{\tilde{\psi}(\hat{b})}-\tilde{B}^{-1}\tilde{\varphi}'(r)\int_{\hat{a}}^{r}\tilde{\psi}(y)y\eta(y)m'(y)dy-\tilde{B}^{-1}\tilde{\psi}'(r)\int_{r}^{\hat{b}}\tilde{\varphi}(y)y\eta(y)m'(y)dy.$$

Dividing this equation first with the term $\tilde{\psi}'(r)$ and then differentiating the resulting equation yield (after a simplification)

$$\tilde{\psi}''(r)=\frac{2rS'(r)}{\sigma^2(r)}\left[\int_{\hat{a}}^{r}\tilde{\psi}(y)y\eta(y)m'(y)dy-\frac{\eta(r)\tilde{\psi}'(r)}{S'(r)}-\frac{\tilde{B}\hat{a}}{\tilde{\varphi}(\hat{a})}\right], \tag{7}$$

since $\tilde{\varphi}''(r)\tilde{\psi}'(r)-\tilde{\psi}''(r)\tilde{\varphi}'(r)=2r\tilde{B}S'(r)/\sigma^2(r)$. Analogously, we can establish that

$$\tilde{\varphi}''(r)=\frac{2rS'(r)}{\sigma^2(r)}\left[\frac{\tilde{B}\hat{b}}{\tilde{\psi}(\hat{b})}-\eta(r)\frac{\tilde{\varphi}'(r)}{S'(r)}-\int_{r}^{\hat{b}}\tilde{\varphi}(y)y\eta(y)m'(y)dy\right]. \tag{8}$$

However, since both $\tilde{\psi}(r)$ and $\tilde{\varphi}(r)$ satisfy the ordinary differential equation $(\mathcal{A}u)(r) = ru(r)$, we find that (cf. [11, p. 18])

$$\frac{\varphi'(\hat{b})}{S'(\hat{b})}-\frac{\varphi'(r)}{S'(r)}=\int_{r}^{\hat{b}}y\varphi(y)m'(y)dy,\quad \frac{\psi'(r)}{S'(r)}-\frac{\psi'(\hat{a})}{S'(\hat{a})}=\int_{\hat{a}}^{r}y\psi(y)m'(y)dy,$$

implying that (7) and (8) can be rewritten as

$$\tilde{\psi}''(r)=\frac{2rS'(r)}{\sigma^2(r)}\left[\int_{\hat{a}}^{r}\tilde{\psi}(y)y(\eta(y)-\eta(r))m'(y)dy-\frac{\eta(r)\tilde{\psi}'(\hat{a})}{S'(\hat{a})}-\frac{\tilde{B}\hat{a}}{\tilde{\varphi}(\hat{a})}\right] \tag{9}$$

and

$$\tilde{\varphi}''(r)=\frac{2rS'(r)}{\sigma^2(r)}\left[\int_{r}^{\hat{b}}\tilde{\varphi}(y)y(\eta(r)-\eta(y))m'(y)dy+\frac{\tilde{B}\hat{b}}{\tilde{\psi}(\hat{b})}-\eta(r)\frac{\tilde{\varphi}'(\hat{b})}{S'(\hat{b})}\right]. \tag{10}$$

The assumed monotonicity of the mapping $\eta(r)$ then implies that

$$\tilde{\psi}''(r)\geq\frac{2rS'(r)}{\sigma^2(r)}\left[-\frac{\eta(r)\tilde{\psi}'(\hat{a})}{S'(\hat{a})}-\frac{\tilde{B}\hat{a}}{\tilde{\varphi}(\hat{a})}\right]\quad\text{and}\quad\tilde{\varphi}''(r)\geq\frac{2rS'(r)}{\sigma^2(r)}\left[\frac{\tilde{B}\hat{b}}{\tilde{\psi}(\hat{b})}+\eta(r)\frac{\tilde{\varphi}'(\hat{b})}{S'(\hat{b})}\right].$$

Observing that $\tilde{\psi}(r) \uparrow \psi(r)$ as $\hat{a} \downarrow a$ and that $\tilde{\varphi}(r) \uparrow \varphi(r)$ as $\hat{b} \uparrow b$ then implies that

$$\psi''(r) \geq -\frac{2rS'(r)}{\sigma^2(r)}\eta(r)\lim_{\hat{a}\downarrow a}\frac{\tilde{\psi}'(\hat{a})}{S'(\hat{a})} \quad \text{and} \quad \varphi''(r) \geq \frac{2rS'(r)}{\sigma^2(r)}\eta(r)\lim_{\hat{b}\uparrow b}\frac{\tilde{\varphi}'(\hat{b})}{S'(\hat{b})} = 0,$$

since $b > 0$ was assumed to be unattainable for r and $a \leq 0$. If a is unattainable for r, then $\lim_{r\downarrow a}\psi'(r)/S'(r) = 0$, proving the convexity of $\psi(r)$ in that case. On the other hand, if a is attainable for r and $\lim_{r\downarrow a}\eta(r) \leq 0$, then $\eta(r) \leq 0$ for all $r \in \mathcal{I}$, proving the convexity of $\psi(r)$ in that case as well. □

Theorem 1 states a set of weak conditions under which the convexity of the minimal excessive mappings $\psi(r)$ and $\varphi(r)$ is always guaranteed. The results of Theorem 1 are very general, since the monotonicity of the mapping $\eta(r)$ is satisfied by almost all models subject to mean reversion and all models subject to decreasing per capita growth rates $\mu(r)/r$. Consequently, our results are valid for, among others, the Black–Karasinski [10], Cox–Ingersoll–Ross [15], [17], [27], Dothan [19], Merton [32], and Vasiček models [36] of interest rates. An interesting implication of Theorem 1 generalizing the conditions under which the convexity of the decreasing fundamental solution $\varphi(r)$ is assured is now summarized in the following. (See [3] for an analogous result in the constant discounting case.)

COROLLARY 1. *Assume that there is a threshold $r_0 \in \mathcal{I}$ such that $\mu(r) \geq 0$ on (a, r_0) and $\eta(r)$ is nonincreasing on (r_0, b). Then $\varphi(r)$ is convex on $\mathcal{I}$.*

Proof. Since $\varphi(r)$ is nonnegative and decreasing, we find that

$$\frac{1}{2}\sigma^2(r)\varphi''(r) = r\varphi(r) - \mu(r)\varphi'(r) \geq 0$$

whenever $r \in (a, r_0)$, that is, whenever $\mu(r) \geq 0$. The convexity of $\varphi(r)$ on (r_0, b) then follows from (10). □

Corollary 1 states a set of weak conditions under which the convexity of the decreasing fundamental solution of the ordinary second-order differential equation $(\mathcal{A}u)(r) = ru(r)$ is always guaranteed. This result is of importance since the decreasing fundamental solution $\varphi(r)$ plays a dominant role in most valuation problems of perpetual interest rate derivatives (cf. [15] and [27]).

It is worth pointing out that the result of Theorem 1 could also be derived for one factor models with *positive rates* (i.e., for which $\mathcal{I} \subseteq \mathbb{R}_+$) by considering the minimal excessive mappings for the diffusion

$$d\tilde{r}(t) = \tilde{\mu}(\tilde{r}(t))dt + \tilde{\sigma}(\tilde{r}(t))d\hat{W}(t), \tag{11}$$

where $\tilde{\mu}(r) = \mu(r)/r$ and $\tilde{\sigma}(r) = \sigma(r)/\sqrt{r}$. Define now the random time change $\beta(t)$ with time change rate r as (cf. [33, pp. 146–151])

$$\beta(t) = \int_0^t r(s)ds$$

and the right-hand inverse process as $\alpha(t) = \inf\{s \in \mathbb{R}_+ : \beta(s) > t\}$. Our assumptions imply that $\alpha(t)$ is continuous, that the random time change $\beta(t)$ is continuous and monotonically increasing, and that $\alpha(\beta(t)) = \beta(\alpha(t)) = t$. Moreover, as is demonstrated in Theorem 8.5.1 of [33] (page 146) the process $\tilde{r}(t)$ coincides in law with the process $r(\alpha(t))$. Consequently, we observe that

$$V(r) = \sup_{\bar{\tau}} E_r\left[e^{-\bar{\tau}}\Phi(\tilde{r}(\bar{\tau}))\right]. \tag{12}$$

In other words, the valuation problem (2) can be solved in terms of the associated stopping problem (12), at least when the lower boundary is unattainable for the interest rate process $r(t)$. This result is, of course, intuitively clear after noticing that the differential operator $\tilde{\mathcal{A}}$ representing the infinitesimal generator of the process $\tilde{r}(t)$ can be written as $\tilde{\mathcal{A}} = \frac{1}{r}\mathcal{A}$. Rewriting the ordinary second-order differential equation $(\mathcal{A}u)(r) = ru(r)$ as $(\tilde{\mathcal{A}}u)(r) = u(r)$ then shows why the minimal excessive mappings for $r(t)$ killed at the rate r and the minimal excessive mappings for $\tilde{r}(t)$ killed at the constant rate 1 coincide and, consequently, why we have the representation (12). It is also worth mentioning at this point that the random time change approach illustrated above applies also when the lower boundary 0 is attainable for the interest rate process $r(t)$. However, in order to apply the time change formula presented above in that case, we have to impose extra boundary conditions for the time-changed process $\tilde{r}(t)$, since the boundary behavior of $\tilde{r}(t)$ typically differs considerably from the boundary behavior of $r(t)$. This is actually a delicate issue which is mostly overlooked in studies considering problems of type (2) (see, for example, [27]). To illustrate this point explicitly, consider problem (2) in the presence of the short rate model (a special case of the Cox–Ingersoll–Ross model corresponding to the continuous time analogue of a branching process; cf. [29, p. 239]):

$$dr(t) = -\mu r(t)dt + \sigma\sqrt{r(t)}d\hat{W}(t), \quad r(0) = r,$$

where $\mu \in \mathbb{R}_+$ and $\sigma > 0$ are exogenously determined constants. Since 0 is an exit and ∞ a natural boundary for the interest rate process $r(t)$, we find that in this case the fundamental solutions of the ordinary second-order differential equation $(\mathcal{A}u)(r) = ru(r)$ read as $\psi(r) = e^{ar} - e^{br}$ and $\varphi(r) = e^{br}$, where $a = \mu/\sigma^2 + \sqrt{\mu^2/\sigma^4 + 2/\sigma^2}$ and $b = \mu/\sigma^2 - \sqrt{\mu^2/\sigma^4 + 2/\sigma^2}$ denote the positive and the negative root of the characteristic equation $\sigma^2 k^2/2 - \mu k - 1 = 0$, respectively. On the other hand, since the operator $\tilde{\mathcal{A}}$ coincides with the differential operator of Brownian motion with drift, we observe, by imposing that $\tilde{r}(t)$ should be killed at 0, that the optimal stopping problem (2) can be rewritten as in (12) with

$$d\tilde{r}(t) = \mu dt + \sigma d\hat{W}(t), \quad \tilde{r}(0) = r,$$

subject to killing at 0. Our main result characterizing the form of the value $V(r)$ is now summarized in the following.

THEOREM 2. *Assume that the conditions of Theorem* 1 *are satisfied. Then the value function $V(r)$ is convex on the continuation region $C = \{r \in \mathcal{I} : V(r) > \Phi(r)\}$. Moreover, if $\Phi(r)$ is convex on $\mathcal{I}$, then $V(r)$ is convex on $\mathcal{I}$ as well.*

Proof. The excessivity of the value function $V(r)$ implies that it is continuous and, therefore, that C is open. Assume that $(x, y) \subset C$ is an arbitrary open subset of C with compact support on $\mathcal{I}$. The harmonicity of the value $V(r)$ on C then implies that if $r \in (x, y)$, then

$$V(r) = E_r\left[e^{-\int_0^{\tau(x,y)} r(s)ds} V(r(\tau(x, y)))\right],$$

where $\tau(x, y) = \inf\{t \geq 0 : r(t) \notin (x, y)\}$ denotes the first exit time from (x, y). As in the proof of Theorem 1, we find that

$$V(r) = \frac{\psi(y)V(x) - V(y)\psi(x)}{\varphi(x)\psi(y) - \varphi(y)\psi(x)}\varphi(r) + \frac{\varphi(x)V(y) - \varphi(y)V(x)}{\varphi(x)\psi(y) - \varphi(y)\psi(x)}\psi(r).$$

However, the excessivity of $V(r)$ implies that $V(x)/V(y) \geq \min(\psi(x)/\psi(y), \varphi(x)/\varphi(y))$ for all $x, y \in \mathcal{I}$ and, therefore, that $\psi(y)V(x) \geq V(y)\psi(x)$ and $\varphi(x)V(y) \geq \varphi(y)V(x)$ (cf. [11, p. 32]). Since the sum of two convex functions is convex, we find that $V(r)$ is convex on C. Moreover, since $V(r) = \Phi(r)$ on the stopping region, we find that if $\Phi(r)$ is convex on $\mathcal{I}$, then $V(r)$ is convex on $\mathcal{I}$ as well. □

Theorem 2 states a set of conditions under which the value is convex, at least in the continuation region C where exercising the option is suboptimal. Interestingly, Theorem 2 also shows that given its conditions, the convexity of the exercise payoff is preserved under valuation. This result is in line with previous results obtained in models considering path-independent American contingent claims (cf. [1] and [22]) and path-dependent European interest rate derivatives (cf. [2]). It is, however, worth emphasizing that *the sufficient conditions for the convexity of the value on the continuation region C do not depend on the payoff $\Phi(r)$ and, therefore, are essentially determined by the infinitesimal coefficients of the diffusion modeling the interest rate dynamics.* This demonstrates that the form of the value of the interest rate derivative on C is *essentially a process-specific and not payoff-specific property* (cf. [1] for similar results in the case of path-independent claims). A set of stronger conditions under which the convexity of the value is also assured is now presented in the following theorem.

THEOREM 3. *Assume that the payoff $\Phi(r)$ is nonincreasing and convex and that*

(a) *$\mu(r)$ is concave,*

(b) *$\mu(r)$ and $\sigma(r)$ are continuously differentiable with Lipschitz-continuous derivatives, and $\sigma'(r)$ satisfies the standard Novikov condition.*

Then the value $V(r)$ is nonincreasing and convex.

Proof. As is shown in [6], given our assumptions (a) and (b), the discount factor $e^{-\int_0^t r(s)ds}$ is decreasing and convex as a mapping of the current short rate r. Therefore, $e^{-\int_0^t r(s)ds} g(r(t))$, being the product of two nonnegative, nonincreasing and convex mappings, is nonnegative, nonincreasing, and convex. Define now the increasing sequence of nonnegative mappings $\{V_n(r)\}_{n\in\mathbb{N}}$ as

$$V_{n+1}(r) = \sup_{t\geq 0} E_r\left[e^{-\int_0^t r(s)ds} V_n(r(t))\right], \quad V_0(r) = \Phi(r).$$

Since $\Phi(r)$ is nonincreasing and convex, and the maximum of a nonincreasing and convex mapping is nonincreasing and convex (by standard duality arguments), we find that $V_n(r)$ is nonincreasing and convex for all n. As is shown in Corollary 10.1.8 of [33], the sequence $V_n(r) \uparrow V(r)$ as $n \to \infty$. Thus, if $q \leq r$ we find that $V(q) \geq V_n(q) \geq V_n(r)$, from which the monotonicity of the value $V(r)$ follows by monotone convergence. Similarly, if $r, q \in \mathcal{I}$ and $\lambda \in [0,1]$ we find that $\lambda V(r) + (1-\lambda)V(q) \geq \lambda V_n(r) + (1-\lambda)V_n(q) \geq V_n(\lambda r + (1-\lambda)q)$, from which the convexity of the value $V(r)$ follows by monotone convergence. □

Theorem 3 states in terms of the drift $\mu(r)$, the volatility coefficient $\sigma(r)$, and the exercise payoff $\Phi(r)$ a set of conditions under which the value of a perpetual American interest rate derivative is nonincreasing and convex as a function of the current short rate. It is worth observing that although the conditions of Theorem 3 are not identical to (and not as general as) the conditions of Theorem 2, the monotonicity condition required in Theorem 2 follows in some cases from the concavity of the drift $\mu(r)$. More precisely, if $\mathbb{R}_+ \subseteq \mathcal{I}$, $\mu(r)$ is concave, and $\lim_{r\downarrow 0} \mu(r) \geq 0$, then $\mu(r)/r$ is decreasing and, therefore, $\eta(r)$ is decreasing as well. An interesting implication of Theorem 3 valid for contracts written on T-bonds (cf. [15] and [27]) is now summarized in the following.

COROLLARY 2. *Denote as $p(r,T)$ the price of a zero coupon bond maturing T periods from exercise, and assume that the mapping $g : \mathbb{R}_+ \mapsto \mathbb{R}_+$ is nondecreasing and convex and that the conditions* (a) *and* (b) *of Theorem* 3 *are satisfied. Then the value*

$$J(r) = \sup_{\tau} E_r \left[e^{-\int_0^\tau r(s)ds} g(p(r(\tau), T)) \right], \tag{13}$$

is nonincreasing and convex.

Proof. The result is a direct implication of Theorem 3. □

Corollary 2 states a set of typically satisfied conditions under which the value of a compound contract written on a T-bond is nonincreasing and convex. Therefore, an important implication of the findings of Corollary 2 is that, given the conditions of Theorem 3, *the value of a call option written on a T-bond is nonincreasing and convex.*

3. The impact of increased volatility. In this section, we plan to consider the comparative static properties of the value function $V(r)$. To accomplish this task, we assume that the interest rate process $\{\hat{r}(t); t \geq 0\}$ evolving on $\mathcal{I}$ is described under the risk-neutral measure $\mathbb{Q}$ by the time homogeneous stochastic differential equation

$$d\hat{r}(t) = \mu(\hat{r}(t))dt + \hat{\sigma}(\hat{r}(t))d\hat{W}(t), \quad \hat{r}(0) = r, \tag{14}$$

where $\hat{\sigma} : \mathcal{I} \mapsto \mathbb{R}_+$ is a given sufficiently smooth (at least continuous) mapping satisfying the condition $\hat{\sigma}(r) \geq \sigma(r)$ for all $r \in \mathcal{I}$. In accordance with the notation in the previous section, we denote the differential operator representing the infinitesimal generator of $\hat{r}$ as

$$\hat{\mathcal{A}} = \frac{1}{2}\hat{\sigma}^2(r)\frac{d^2}{dr^2} + \mu(r)\frac{d}{dr}. \tag{15}$$

We also denote the increasing and decreasing fundamental solutions of the ordinary second-order differential equation $(\hat{\mathcal{A}}u)(r) = ru(r)$ as $\hat{\psi}(r)$ and $\hat{\varphi}(r)$, respectively. Our first important auxiliary result is now summarized in the following lemma.

LEMMA 1. *Assume that the conditions of Theorem* 1 *are satisfied. Then*

$$\min(\hat{\psi}(r)/\hat{\psi}(q), \hat{\varphi}(r)/\hat{\varphi}(q)) \geq \min(\psi(r)/\psi(q), \varphi(r)/\varphi(q))$$

for all $r, q \in \mathcal{I}$. Thus, the class of excessive mappings for the diffusion $\hat{r}(t)$ killed at the rate r belongs into the class of excessive mappings for the diffusion $r(t)$ killed at the rate r.

Proof. Given the conditions of our lemma, we know that $\psi(r)$ and $\varphi(r)$ are convex on $\mathcal{I}$. Thus, we observe that for all $r \in \mathcal{I}$ we have that $(\hat{\mathcal{A}}\psi)(r) - r\psi(r) = ((\hat{\mathcal{A}} - \mathcal{A})\psi)(r) = \frac{1}{2}(\hat{\sigma}(r) - \sigma(r))\psi''(r) \geq 0$ and $(\hat{\mathcal{A}}\varphi)(r) - r\varphi(r) = \frac{1}{2}(\hat{\sigma}(r) - \sigma(r))\varphi''(r) \geq 0$; that is, $\psi(r)$ and $\varphi(r)$ are subharmonic for $\hat{r}(t)$ on $\mathcal{I}$. Consequently, we observe that if $\hat{\tau}(q) = \inf\{t \geq 0 : \hat{r}(t) = q\}$ and $r \leq q$, then

$$\psi(q)\frac{\hat{\psi}(r)}{\hat{\psi}(q)} = E_r \left[e^{-\int_0^{\hat{\tau}(q)} \hat{r}(s)ds} \psi(\hat{r}(\hat{\tau}(q))) \right] \geq \psi(r).$$

Similarly,

$$\varphi(r)\frac{\hat{\varphi}(q)}{\hat{\varphi}(r)} = E_q \left[e^{-\int_0^{\hat{\tau}(r)} \hat{r}(s)ds} \varphi(\hat{r}(\hat{\tau}(r))) \right] \geq \varphi(q),$$

where $\hat{\tau}(r) = \inf\{t \geq 0 : \hat{r}(t) = r\}$. Combining these two inequalities then prove that $\min(\hat{\psi}(r)/\hat{\psi}(q), \hat{\varphi}(r)/\hat{\varphi}(q)) \geq \min(\psi(r)/\psi(q), \varphi(r)/\varphi(q))$ for all $r, q \in \mathcal{I}$.

An arbitrary nontrivial mapping $f : \mathcal{I} \mapsto \mathbb{R}_+$ is excessive for the diffusion $\hat{r}(t)$ killed at the rate r if and only if $f(r)$ is continuous and nonnegative and satisfies for any $(x, y) \subset \mathcal{I}$, $a < x < y < b$ the inequality

$$u(r) = E_r\left[e^{-\int_0^{\hat{\tau}(x,y)} \hat{r}(s)ds} f(\hat{r}(\hat{\tau}(x, y)))\right] \leq f(r),$$

where $\hat{\tau}(x, y) = \inf\{t \geq 0 : \hat{r}(t) \notin (x, y)\}$ (cf. [11, p. 32]). Since $(\hat{\mathcal{A}}u)(r) = ru(r)$ for all $r \in (x, y)$ and $u(r)$ is convex on (x, y) by Theorem 2, we find that $(\mathcal{A}u)(r) - ru(r) = ((\mathcal{A} - \hat{\mathcal{A}})u)(r) = \frac{1}{2}(\sigma^2(r) - \hat{\sigma}^2(r))u''(r) \leq 0$ for all $r \in (x, y)$. Applying Dynkin's theorem to the mapping $r \mapsto u(r)$ and invoking the continuity of the mapping at x and y then imply that

$$f(r) \geq u(r) \geq E_r\left[e^{-\int_0^{\tau(x,y)} r(s)ds} u(r(\tau(x, y)))\right] = E_r\left[e^{-\int_0^{\tau(x,y)} r(s)ds} f(r(\tau(x, y)))\right],$$

which demonstrates that $f(r)$ is excessive for the diffusion $r(t)$ killed at the rate r as well. □

Define now the value of the contingent contract with exercise payoff $\Phi(r)$ and defined with respect to the more volatile interest rate process $\hat{r}(t)$ as

$$\hat{V}(r) = \sup_{\hat{\tau}} E_r\left[e^{-\int_0^{\hat{\tau}} \hat{r}(s)ds} \Phi(\hat{r}(\hat{\tau}))\right], \tag{16}$$

where $\hat{\tau}$ is an arbitrary stopping time. The key implication of our Lemma 1 is now summarized in our main theorem.

THEOREM 4. *Assume that the conditions of Theorem* 1 *are satisfied. Then* $\hat{V}(r) \geq V(r)$ *for all* $r \in \mathcal{I}$ *and* $C = \{r \in \mathcal{I} : V(r) > \Phi(r)\} \subseteq \{r \in \mathcal{I} : \hat{V}(r) > \Phi(r)\} = \hat{C}$.

Proof. As was demonstrated in Lemma 1, the class of excessive mappings for the diffusion $\hat{r}(t)$ killed at the rate r belongs to the class of excessive mappings for the diffusion $r(t)$ killed at the rate r. Thus, $\hat{V}(r)$, being the least excessive majorant of $\Phi(r)$ for the diffusion $\hat{r}(t)$ killed at the rate r, is an excessive majorant of the payoff $\Phi(r)$ for the diffusion $r(t)$ killed at the rate r as well. Since $V(r)$ is the least of such majorants, we find that $\hat{V}(r) \geq V(r)$ for all $r \in \mathcal{I}$. Finally, if $r \in C$, then $\hat{V}(r) \geq V(r) > \Phi(r)$, proving that $r \in \hat{C}$ as well and, therefore, that $C \subseteq \hat{C}$. □

Theorem 4 demonstrates that given the conditions of our Theorem 1, increased volatility increases the value of the contingent claim and postpones the rational exercise of the contingent contract by expanding the continuation region where exercising is suboptimal. Put somewhat differently, Theorem 4 demonstrates that given the conditions of our Theorem 1, the required exercise premium is an increasing function of the volatility of the process (i.e., an increasing mapping of risk). This result is of interest since it demonstrates how a misspecified volatility coefficient affects the value of the considered derivative instruments. If we overestimate the true market volatility, then our theoretical prices will exceed the true ones, and vice versa. An interesting implication of these findings is now summarized in our next corollary.

COROLLARY 3. *Assume that the conditions of Corollary* 2 *are satisfied. Then increased volatility increases the value of derivatives written on zero coupon bonds. Put formally, increased volatility increases the value* $J(r)$ *defined in* (13).

Proof. As was shown in Corollary 2, $J(r)$ is nonincreasing and convex as a function of the current short rate r. Thus, the conditions of Theorem 4 are satisfied, and the alleged result follows. □

Corollary 3 shows that given the general conditions of Corollary 2, increased volatility increases the value of American contingent contracts written on T-bonds. Since increased volatility increases the value of the T-bonds as well, it is not obvious whether the impact of increased volatility on the continuation region is positive or not. Fortunately, there is a broad class of cases for which the positivity of this relationship can be unambiguously established. Our main result on this topic is now proven in the following theorem.

THEOREM 5. *Assume that $\hat{\sigma}(r) = \kappa\sigma(r)$, where $\kappa \in (1, \infty)$ is a known constant. Assume also that the conditions of Corollary 2 are satisfied. Then an increase in the parameter κ increases the value $J(r)$ and postpones rational exercise by expanding the continuation region where exercising the contract is suboptimal.*

Proof. Denote now as $r_\kappa(t)$ the solution of the stochastic differential equation (14) when $\hat{\sigma}(r) = \kappa\sigma(r)$. Denote also as $J_\kappa(r)$ the value $J(r)$ and as $p_\kappa(r,T)$ the price of a T-bond as a function of the known parameter κ. The assumed smoothness of the infinitesimal coefficients $\mu(r)$ and $\sigma(r)$ imply that $r_\kappa(t)$ is continuous as a function of κ. This, in turn, implies that both the value $J_\kappa(r)$ and the price of a T-bond are continuous as functions of the parameter κ. We have already established that $J_\kappa(r) \geq J(r)$ and we also know from [6] that $p_\kappa(r,T) \geq p(r,T)$. Define now the exercise regions $\Gamma = \{r \in \mathcal{I} : J(r) = g(p(r,T))\}$ and $\Gamma_\kappa = \{r \in \mathcal{I} : J_\kappa(r) = g(p_\kappa(r,T))\}$ and let $r \in \Gamma_\kappa$. Then

$$J_\kappa(r) = g(p_\kappa(r,T)) \geq J(r) \geq g(p(r,T)) \Rightarrow g(p_\kappa(r,T)) - g(p(r,T)) \geq J(r) - g(p(r,T)) \geq 0.$$

The continuity of $g(p_\kappa(r,T))$ as a mapping of the parameter κ then implies that there is a sequence $\kappa_n \downarrow 1$ for which the difference $|g(p_{\kappa_n}(r,T)) - g(p(r,T))| < \epsilon/n$ when $n > \bar{N}$. Since $\epsilon > 0$ is arbitrary, we find that $J(r) = g(p(r,T))$ as well and, therefore, that $\Gamma_\kappa \subseteq \Gamma$. Since $C = \{r \in \mathcal{I} : J(r) > g(p(r,T))\} = \mathcal{I}\backslash\Gamma$ and $C_\kappa = \{r \in \mathcal{I} : J_\kappa(r) > g(p_\kappa(r,T))\} = \mathcal{I}\backslash\Gamma_\kappa$, we find that $C \subseteq C_\kappa$, thus proving the alleged claim. □

4. Illustration. It is our purpose in this section to illustrate the results of the two previous sections by considering a class of problems arising frequently in the literature on interest rate derivatives. We assume throughout this section that $\Phi \in C^2(\mathcal{I})$, although this assumption may be relaxed since continuous mappings can be approximated by a sequence of twice continuously differentiable mappings converging uniformly on compacts towards Φ (a mollification of Φ; cf. [4] and [33, pp. 299–302]). Define now the continuous mapping $f : \mathcal{I} \mapsto \mathbb{R}$ as $f(r) = (\mathcal{A}\Phi)(r) - r\Phi(r)$. Our main result is now summarized in the following.

THEOREM 6. *Assume that there is a threshold $\tilde{r} \in (a,b)$ for which $f(r) \gtreqless 0$, when $r \gtreqless \tilde{r}$, that $\lim_{r\uparrow b} \Phi(r) < \infty$, and that*

$$E_r \int_0^\infty e^{-\int_0^s r(t)dt} |f(r(s))| ds < \infty.$$

(A) *If a is either natural or exit for the interest rate process $r(t)$ killed at the rate r, then $\tau(r^*) = \inf\{t \geq 0 : r(t) \leq r^*\}$ is the optimal exercise date and the value reads as*

$$V(r) = \varphi(r) \sup_{q\in(a,r]} \frac{\Phi(q)}{\varphi(q)} = \begin{cases} \Phi(r^*)\frac{\varphi(r)}{\varphi(r^*)}, & r \in (r^*, b), \\ \Phi(r), & r \in (a, r^*], \end{cases} \tag{17}$$

where $r^* = \operatorname{argmax}\{\Phi(r)/\varphi(r)\} \in (a, \tilde{r})$, denoting the optimal exercise price, is the unique root of the smooth pasting condition $\Phi'(r^*)\varphi(r^*) = \Phi(r^*)\varphi'(r^*)$.

(B) *If a is either entrance or killing boundary for the interest rate process $r(t)$ killed at the rate r and the equation $\Phi'(r)\varphi(r) = \Phi(r)\varphi'(r)$ has an interior root on $(a, \tilde{r})$, then the value reads as in* (17) *and $\tau(r^*) = \inf\{t \geq 0 : r(t) \leq r^*\}$ is the optimal exercise date.*

Proof. Denote as $V_p(r)$ the proposed value function. It is then clear that the excessivity of the value $V(r)$ implies that $V(r) \geq V_p(r)$ (cf. [11, p. 32]). To prove the opposite, we first observe that if an optimal threshold r^* exists, then the proposed value function $V_p(r)$ is nonnegative, dominates the exercise payoff $\Phi(r)$, is twice continuously differentiable outside the threshold r^*, and satisfies the variational inequalities $\min\{rV_p(r) - (\mathcal{A}V_p)(r), V_p(r) - \Phi(r)\} = 0$. Thus, we find that if an optimal threshold r^* exists, then the proposed value function $V_p(r)$ is an excessive majorant of the exercise payoff $\Phi(r)$. However, since $V(r)$ is the least of these majorants, we find that $V_p(r) \geq V(r)$ whenever r^* exists. Consequently, it is sufficient to demonstrate that an optimal threshold r^* exists. Applying Dynkin's theorem to the mapping $r \mapsto \Phi(r)$ and following the proof of our Theorem 1 yield that

$$\Phi(\hat{a})\frac{\tilde{\varphi}(r)}{\tilde{\varphi}(\hat{a})} + \Phi(\hat{b})\frac{\tilde{\psi}(r)}{\tilde{\psi}(\hat{b})} = \Phi(r) + \tilde{B}^{-1}\tilde{\varphi}(r)\int_{\hat{a}}^{r} \tilde{\psi}(y)f(y)m'(y)dy + \tilde{B}^{-1}\tilde{\psi}(r)\int_{r}^{\hat{b}} \tilde{\varphi}(y)f(y)m'(y)dy,$$

where $a < \hat{a} < \hat{b} < b$. Dividing this equation with the mapping $\tilde{\varphi}(r)$, differentiating the resulting equation, and reordering terms then yield

$$\frac{\Phi'(r)}{S'(r)}\tilde{\varphi}(r) - \frac{\tilde{\varphi}'(r)}{S'(r)}\Phi(r) = \frac{\Phi(y)}{\tilde{\psi}(\hat{b})}\tilde{B} - \int_{r}^{\hat{b}} \tilde{\varphi}(y)f(y)m'(y)dy.$$

Letting $\hat{b} \uparrow b$ and invoking the assumptions of our theorem then yield

$$\frac{\Phi'(r)}{S'(r)}\varphi(r) - \frac{\varphi'(r)}{S'(r)}\Phi(r) = -\int_{r}^{b} \varphi(y)f(y)m'(y)dy. \tag{18}$$

As is shown in the proof of part B of Theorem 4 in [5], (18) has, under the assumptions of our theorem, a unique root whenever 0 is either natural or exit. Moreover, $\Phi'(r)\varphi(r) \gtreqless \Phi(r)\varphi'(r)$ when $r \lesseqgtr r^*$, demonstrating that $r^* = \operatorname{argmax}\{\Phi(r)/\varphi(r)\}$, thus completing the proof of part (A) of our theorem. The proof of part (B) is then analogous. □

Theorem 6 states a set of typically satisfied conditions under which the optimal stopping problem (2) can always be explicitly solved in terms of the decreasing fundamental solution of the ordinary second-order differential equation $(\mathcal{A}u)(r) - ru(r) = 0$. It is worth observing that the optimal exercise boundary always exists when the lower boundary a is either natural or exit for the interest rate process $r(t)$ killed at the rate r. However, if a is either regular or entrance, then the equation $\Phi'(r)\varphi(r) = \Phi(r)\varphi'(r)$ may or may not have an interior root in $\mathcal{I}$. If such root does not exist, then the contract is never exercised before expiration. An important corollary of our theorem is now summarized in the following corollary.

COROLLARY 4. *Assume that the conditions of Theorems* 1 *and* 6 *are satisfied. Then increased volatility increases the value $V(r)$ and postpones exercise by expanding the continuation region $C = \{r \in \mathcal{I} : V(r) > \Phi(r)\}$ where exercising the contract is suboptimal.*

Proof. The result is a direct consequence of Theorems 4 and 6. □

In line with our theoretical findings, we observe from Theorem 6 and Corollary 4 that the value of the contingent contract is decreasing on $\mathcal{I}$ and convex on C. Consequently, we find that increased volatility increases its value and expands the continuation region where exercising the contract is suboptimal. Moreover, since in the absence of stochasticity (i.e., when $\sigma(r) \equiv 0$) the optimal exercise threshold, denoted now as $\bar{r}$, satisfies the ordinary first-order condition $\mu(\bar{r})\Phi'(\bar{r}) = \bar{r}\Phi(\bar{r})$ and increased volatility increases r^*, we find that increased volatility increases the required exercise premium $r^* - \bar{r}$ as well.

Acknowledgments. The author is grateful to Paavo Salminen and to two anonymous referees for their constructive comments and suggested improvements on earlier versions of this study.

REFERENCES

[1] L. H. R. Alvarez, *On the Convexity and Comparative Static Properties of a Class of r-harmonic Mappings*, Research Report A40, Institute of Applied Mathematics, University of Turku, Turku, Finland, 2000.

[2] L. H. R. Alvarez, *On the form and risk-sensitivity of zero coupon bonds for a class of interest rate models*, Insurance Math. Econ., 28 (2001), pp. 83–90.

[3] L. H. R. Alvarez, *Does increased stochasticity speed up extinction?*, J. Math. Biol., 43 (2001), pp. 534–544.

[4] L. H. R. Alvarez, *Solving optimal stopping problems of linear diffusions by applying convolution approximations*, Math. Methods Oper. Res., 53 (2001), pp. 89–99.

[5] L. H. R. Alvarez, *Reward functionals, salvage values and optimal stopping*, Math. Methods Oper. Res., 54 (2001), pp. 315–337.

[6] L. H. R. Alvarez and E. Koskela, *Wicksellian Theory of Forest Rotation under Interest Rate Variability*, CESifo Working Paper Series 606, University of Munich, Munich, Germany, 2001.

[7] T. Björk, *Interest rate theory*, in Financial Mathematics, CIME Lectures 1996, W. J. Runggaldier, ed., Lecture Notes in Math. 1656, Springer-Verlag, Berlin, 1997, pp. 53–122.

[8] T. Björk, *Arbitrage Theory in Continuous Time*, Oxford University Press, Oxford, UK, 1998.

[9] F. Black, E. Derman, and W. Toy, *A one-factor model of interest rates and its application to treasury bond options*, Financial Anal. J., 46 (1990), pp. 33–39.

[10] F. Black and P. Karasinski, *Bond and option pricing when short rates are lognormal*, Financial Anal. J., 47 (1991), pp. 52–59.

[11] A. Borodin and P. Salminen, *Handbook on Brownian Motion—Facts and Formulae*, Birkhäuser, Basel, 1996.

[12] M. J. Brennan and E. S. Schwartz, *A continuous time approach to pricing bonds*, J. Banking and Finance, 3 (1979), pp. 133–155.

[13] J. H. Cochrane, *Asset Pricing*, Princeton University Press, Princeton, NJ, 2001.

[14] G. Courtadon, *The pricing of options on default-free bonds*, J. Financial and Quantitative Anal., 17 (1982), pp. 75–100.

[15] J. C. Cox, J. E. Ingersoll, and S. A. Ross, *An analysis of variable rate loan contracts*, J. Finance, 35 (1980), pp. 389–403.

[16] J. C. Cox, J. E. Ingersoll, and S. A. Ross, *A re-examination of traditional expectation hypotheses about the term structure of interest rates*, J. Finance, 36 (1981), pp. 769–799.

[17] J. C. Cox, J. E. Ingersoll, and S. A. Ross, *A theory of the term structure of interest rates*, Econometrica, 53 (1985), pp. 385–407.

[18] A. K. Dixit and R. S. Pindyck, *Investment under Uncertainty*, Princeton University Press, Princeton, NJ, 1994.

[19] U. Dothan, *On the term structure on interest rates*, J. Financial Economics, 6 (1978), pp. 59–69.

[20] D. Duffie, *Dynamic Asset Pricing Theory*, 2nd ed., Princeton University Press, Princeton, NJ, 1996.

[21] D. Duffie and R. Kan, *A yield factor model of interest rates*, Math. Finance, 6 (1996), pp. 379–406.
[22] N. El Karoui, M. Jeanblanc-Picqué, and S. E. Shreve, *Robustness of the Black-Scholes formula*, Math. Finance, 8 (1998), pp. 93–126.
[23] D. Heath, R. Jarrow, and A. Morton, *Bond pricing and the term structure of interest rates*, Econometrica, 60 (1992), pp. 77–106.
[24] D. G. Hobson, *Volatility mis-specification, option pricing and super-replication via coupling*, Ann. Appl. Probab., 8 (1998), pp. 193–205.
[25] J. Hull and A. White, *Pricing interest-rate derivative securities*, Rev. Financial Studies, 3 (1990), pp. 573–592.
[26] J. Hull and A. White, *One-factor interest rate models and the valuation of interest rate derivative securities*, J. Financial and Quantitative Anal., 28 (1993), pp. 235–254.
[27] J. E. Ingersoll, Jr., and S. A. Ross, *Waiting to invest: Investment and uncertainty*, J. Business, 65 (1992), pp. 1–29.
[28] K. Ito and H. P. McKean, Jr., *Diffusion Processes and Their Sample Paths*, Springer, Berlin, 1974.
[29] S. Karlin and H. Taylor, *A Second Course in Stochastic Processes*, Academic Press, Orlando, 1981.
[30] F. A. Longstaff and E. S. Schwartz, *Interest rate volatility and the term structure*, J. Finance, 47 (1993), pp. 1259–1282.
[31] R. C. Merton, *A rational theory of option pricing*, Bell J. Economics and Management Science, 41 (1973), pp. 141–183.
[32] R. C. Merton, *An asymptotic theory of growth under uncertainty*, Rev. Economic Studies, 42 (1975), pp. 375–393.
[33] B. Øksendal, *Stochastic Differential Equations: An Introduction with Applications*, 5th ed., Springer, Berlin, 1998.
[34] K. Sandmann and D. Sondermann, *A term structure model and the pricing of interest rate derivatives*, Rev. of Future Markets, 12 (1993), pp. 391–423.
[35] W. M. Schmidt, *On a general class of one-factor models for the term-structure of interest rates*, Finance Stoch., 1 (1997), pp. 3–24.
[36] O. Vasiček, *An equilibrium characterization of the term structure*, J. Financial Economics, 5 (1977), pp. 177–188.

SIAM J. APPL. MATH.
Vol. 63, No. 3, pp. 937–961

THE COUPLING OF MOTION AND CONDUCTIVE HEATING OF A GAS BY LOCALIZED ENERGY SOURCES*

ANTONIO L. SÁNCHEZ†, JOSÉ L. JIMÉNEZ-ÁLVAREZ†, AND AMABLE LIÑÁN‡

Abstract. This paper investigates the time evolution of the near-isobaric flow field produced in a gas after the sudden application of a constant heat flux from a localized energy source. The problems of plane, line, and point heat sources are all investigated, with a power law for the temperature dependence of the thermal conductivity, after reduction to a quasi-linear heat equation for the temperature. In the planar and spherical cases, the constant heat flux defines scales for the length and time, which are used to nondimensionalize the problem. Numerical integration is used to provide the evolution of the temperature and velocity, and limiting solutions corresponding to small and large rescaled times are obtained. In the axisymmetric case, due to the absence of characteristic length and time scales, the solution is seen to admit a self-similar description in terms of the nondimensional heat flux. Profiles of temperature and radial velocity are provided for different values of this parameter, and the asymptotic limits of both small and large heating rates are addressed separately. The analysis reveals, in particular, the existence of front solutions when the resulting temperatures become much larger than the initial temperature, as occurs for sufficiently large times for the planar source, for sufficiently small times for the point source, and for sufficiently large heating rates for the line source.

Key words. self-similar solutions, asymptotic methods, nonlinear heat conduction, front solutions

AMS subject classifications. 34E15, 80A32

PII. S0036139902403895

1. Introduction.

The expansion accompanying the heating of a gas after the application of an energy source sets the fluid in motion away from the source. The purpose of this paper is to give a description of the associated nonlinear heating process when the induced velocities are much smaller than the velocity of sound, so that one can neglect pressure variations in the first approximation. Furthermore, the analysis treats the energy source as being of negligible size and neglects the effect of gravity, two simplifications that are simultaneously valid when the size of the heated region is much larger than that of the energy source and still sufficiently small so that the buoyancy-induced velocity remains smaller than the thermal-expansion velocity.

We shall consider the one-dimensional transient solutions appearing with plane, line, and point energy sources when a constant heat flux is applied. Numerical and asymptotic techniques will be employed to describe the evolution with time of the temperature and velocity fields. The solution will be seen to depend on the combined effect of outward convection, due to the gas expansion, and of nonlinear heat conduction, associated with the temperature dependence of the thermal conductivity. The results of the analysis should be useful for understanding the ignition process of a reactive gas mixture by a localized energy source, as can be realized in practice by a

*Received by the editors March 11, 2002; accepted for publication (in revised form) September 13, 2002; published electronically February 25, 2003. This research was supported by the Fifth Framework program of the European Commission under the Energy, Environment, and Sustainable Development contract EVG1-CT-2001-00042 EXPRO and by the Spanish MCYT under project 2001-4603-E.

http://www.siam.org/journals/siap/63-3/40389.html

†Departamento de Ingeniería Térmica y de Fluidos, Universidad Carlos III de Madrid, 28911 Leganés, Spain (asanchez@ing.uc3m.es).

‡Departamento de Motopropulsión y Termofluidodinámica, E. T. S. I. Aeronáuticos, Universidad Politécnica de Madrid, 28040 Madrid, Spain (linan@tupi.dmt.upm.es).

laser beam or by passing an electric current through a thin wire.

A simple order-of-magnitude analysis serves to anticipate the characteristic scales of the problem. Consider an energy source located in an infinite gas medium at rest with initial temperature and density T_o and ρ_o, respectively. If a constant energy flux is applied after time $\mathrm{t} = 0$, the characteristic time t_c required to heat a region of characteristic size r_c, so that the temperature is increased by an amount of order T_o, is determined by the energy balance

$$\mathrm{q}_j \mathrm{t}_c \sim \rho_o c_p \mathrm{T}_o \mathrm{r}_c^{j+1}, \tag{1.1}$$

where c_p is the specific heat at constant pressure, assumed to be constant, and the index j takes the values $j = (0, 1, 2)$ for planar, cylindrical, and spherical geometries. Correspondingly, q_0, q_1, and q_2 represent, respectively, the heating rate per unit surface for the planar source, the heating rate per unit length for the line source, and the heating rate of the point source. The above equation must be supplemented by the condition

$$\mathrm{q}_j \sim \mathrm{r}_c^{j-1} k_o \mathrm{T}_o, \tag{1.2}$$

which states that the energy flux is conducted across the heated region, with k_o representing the value of the thermal conductivity at the initial temperature T_o.

For the planar and point sources, the above two balances give the characteristic scales of length and time

$$\mathrm{r}_c \sim \left(\frac{\mathrm{q}_j}{k_o \mathrm{T}_o}\right)^{1/(j-1)} \quad \text{and} \quad \mathrm{t}_c \sim \alpha_o^{-1} \left(\frac{\mathrm{q}_j}{k_o \mathrm{T}_o}\right)^{2/(j-1)}, \tag{1.3}$$

where $\alpha_o = k_o/(\rho_o c_p)$ is the unperturbed thermal diffusivity. On the other hand, the characteristic velocity due to thermal expansion associated with relative changes in density of order unity, which can be anticipated from the continuity equation to be of order $\mathrm{v}_c = \mathrm{r}_c/\mathrm{t}_c$, becomes in this case

$$\mathrm{v}_c \sim \alpha_o \left(\frac{\mathrm{q}_j}{k_o \mathrm{T}_o}\right)^{-1/(j-1)}. \tag{1.4}$$

As shown below, use of these scales enables the problems $j = 0$ and $j = 2$ to be written in a convenient parameter-free form. On the other hand, no characteristic scales can be constructed for the line source, for which the radial extent of the heated region increases with time according to

$$\mathrm{r}_c \sim \left[\frac{\mathrm{q}_1}{(k_o \mathrm{T}_o)}\right]^{1/2} (\alpha_o \mathrm{t})^{1/2}, \tag{1.5}$$

while the characteristic velocity is given by

$$\mathrm{v}_c \sim \left[\frac{\mathrm{q}_1}{(k_o \mathrm{T}_o)}\right]^{1/2} (\alpha_o/\mathrm{t})^{1/2}. \tag{1.6}$$

Because of the absence of characteristic scales, the problem will be seen to admit a similarity solution in terms of the self-similar coordinate $\mathrm{r}/(\alpha_o \mathrm{t})^{1/2}$, with $q = \mathrm{q}_1/(2\pi k_o \mathrm{T}_o)$ entering as a governing parameter.

The ranges of validity for the different assumptions employed in the paper can be delineated by using the above scaling laws (1.3)–(1.6). For instance, the assumption that the source is localized is valid only when r_c is much larger than the size of the energy source, while gravity-induced velocities, of order $(g\mathrm{r}_c)^{1/2}$, can be neglected only when $\mathrm{v}_c \gg (g\mathrm{r}_c)^{1/2}$ for $\mathrm{r}_c \ll (\alpha_o^2/g)^{1/3}$. The assumption of isobaric heating holds only when the induced velocities v_c are much smaller than the velocity of sound, given in order of magnitude by $(c_p\mathrm{T}_o)^{1/2}$, thereby producing pressure variations, of order $\rho_o v_c^2$, that are much smaller than the ambient value. It should also be noted that the above considerations provide, for a given gas mixture, the range of heating rates for which the analyses of the planar and point sources remain accurate. On the other hand, since the characteristic scales given in (1.5)–(1.6) change with time, the above considerations give the time range for which the analysis of the line source holds.

Neither finite-size sources nor buoyancy and compressibility effects are addressed in the present work. When gravity enters, the symmetric solution determined here is expected to evolve to give a steady plume for large times, giving a flow pattern that has been extensively studied in the past (see, e.g., [8, 9] for entries into the literature of thermal plumes from line and point sources). When compressibility effects are significant, a strong shock wave can be expected to form, a phenomenon also observed following the instantaneous localized deposition of a finite amount of energy [18, 19, 20]. This shock wave weakens as it moves away from the source, eventually leading to an acoustic wave as the pressure settles everywhere to the ambient value for sufficiently large times.

The structure of the paper is as follows. After formulating the problem, we will address the similarity solution emerging in the case of a line source. The self-similar temperature and velocity profiles will be given for different values of the heat release rate, and the asymptotic limit of large heat release rates will be described in detail. Next, we will present the solution corresponding to planar and point sources, which involve integration of a parameter-free nonlinear parabolic equation for the temperature. The analysis is extended to include the asymptotic limits of small and large rescaled times. Finally, some concluding remarks will be given.

It should be noted that the problem of near-isobaric heat propagation in a gas from a plane source was addressed previously by Clarke, Kassoy, and Riley in their study of heating of a gas slot confined between infinite parallel walls [5]. In particular, the nonlinear heat equation that governs the problem was derived. They showed that, when the heating rate is applied for a sufficiently long time, the characteristic temperature of the heated region becomes much larger than T_o, so that the thermal wave becomes a front solution with an edge that clearly defines the hot region surrounding the source. We shall see that a front solution also appears with the line source in the limit of large heat release rates and with the point source for sufficiently small times. As seen below, the structure of the solution includes a locally planar thin layer of warm gas, identical for all three configurations, that separates the region of hot gas from the outer cold gas, at temperature $\mathrm{T} = \mathrm{T}_o$. It is worth mentioning that similar front solutions have been previously identified in asymptotic analyses of heat conduction problems when the thermal conductivity depends strongly on the temperature [22], e.g., in electronic conduction in plasmas [16, 23], or in the presence of large temperature variations in gases, as occurs in supercritical droplet evaporation [15]. Front solutions are also encountered in problems of mass diffusion [7] and in flows in porous media [2].

2. Formulation. In the near-isobaric limit considered here, the momentum equation becomes secondary for the computation of the one-dimensional problems addressed, in the sense that the resulting velocity and temperature fields can be determined by integrating the continuity equation

$$\frac{\partial \rho}{\partial \mathrm{t}} + \frac{1}{\mathrm{r}^j}\frac{\partial}{\partial \mathrm{r}}(\mathrm{r}^j \rho \mathrm{v}) = 0 \tag{2.1}$$

and the energy equation

$$\frac{\partial}{\partial \mathrm{t}}(\rho c_v \mathrm{T}) + \frac{1}{\mathrm{r}^j}\frac{\partial}{\partial \mathrm{r}}\left(\mathrm{r}^j \rho \mathrm{v} c_p \mathrm{T} - \mathrm{r}^j k \frac{\partial \mathrm{T}}{\partial \mathrm{r}}\right) = 0, \tag{2.2}$$

supplemented with the equation of state for the ideal gas

$$\rho \mathrm{T} = \rho_o \mathrm{T}_o, \tag{2.3}$$

which is written with pressure differences neglected. These pressure differences, which are much smaller than the ambient pressure in this near-isobaric limit, can be computed a posteriori by integrating the momentum balance equation. In the formulation, ρ, T, and v denote, respectively, the density, temperature, and velocity of the gas, while c_v represents the specific heat at constant volume. For generality, the thermal conductivity k is allowed to vary in our analysis from its initial value k_o, with a temperature dependence given by

$$\frac{k}{k_o} = \left(\frac{\mathrm{T}}{\mathrm{T}_o}\right)^{\sigma}, \tag{2.4}$$

where the exponent σ is typically in the range $0 \le \sigma \le 1$ in gases and takes the value $\sigma = 5/2$ for electronic conduction in plasmas. The initial and boundary conditions for (2.1) and (2.2) corresponding to an infinitesimally small heat source located at $\mathrm{r} = 0$ are

$$\mathrm{t} = 0, \quad \mathrm{r} > 0 : \ \mathrm{T} = \mathrm{T}_o, \quad \rho = \rho_o \tag{2.5}$$

and

$$\mathrm{t} > 0 \begin{cases} \mathrm{r} = 0 : & \mathrm{v} = 0, \ -2^j \pi^{\delta_j} \mathrm{r}^j k \partial \mathrm{T}/\partial \mathrm{r} = \mathrm{q}_j, \\ \mathrm{r} = \infty : & \mathrm{T} = \mathrm{T}_o, \end{cases} \tag{2.6}$$

where $\delta_j = 0$ if $j = 0$ and $\delta_j = 1$ otherwise.

The approximation (2.3) eliminates the time derivative in (2.2), because in this limit of near-isobaric heating, the internal energy does not accumulate locally in the flow field. Integrating the resulting equation, using the boundary condition at $\mathrm{r} = 0$, yields

$$2^j \pi^{\delta_j} \mathrm{r}^j \left(\mathrm{v} \rho_o c_p \mathrm{T}_o - k \frac{\partial \mathrm{T}}{\partial \mathrm{r}}\right) = \mathrm{q}_j. \tag{2.7}$$

As can be seen, the heat released at $\mathrm{r} = 0$ is transported partly by convection and partly by heat conduction. Introducing the dimensionless temperature $T = \mathrm{T}/\mathrm{T}_o = \rho_o/\rho$ and substituting (2.7) into (2.1) finally gives

$$\frac{1}{T^2}\frac{\partial T}{\partial \mathrm{t}} - \frac{1}{\mathrm{r}^j}\frac{\partial}{\partial \mathrm{r}}\left[\frac{\alpha_o}{T}\left(\frac{\mathrm{q}_j}{2^j \pi^{\delta_j} k_o \mathrm{T}_o} + \mathrm{r}^j T^{\sigma}\frac{\partial T}{\partial \mathrm{r}}\right)\right] = 0 \tag{2.8}$$

to be integrated with initial and boundary conditions

$$\left\{ \begin{array}{l} \mathrm{t} = 0 : \ T = 1, \\ t > 0 : \left\{ \begin{array}{l} \mathrm{r} = 0 : \ \mathrm{r}^j T^\sigma \partial T / \partial \mathrm{r} = -\mathrm{q}_j / (2^j \pi^{\delta_j} k_o \mathrm{T}_o), \\ \mathrm{r} = \infty : \ T = 1. \end{array} \right. \end{array} \right. \tag{2.9}$$

This nonlinear heat problem describes the buoyancy-free isobaric evolution of the gas temperature, subject to a localized energy source of rate q_j, which can vary with time.

3. The line source of heat. As previously mentioned, for the line source the solution to (2.8) when the heating rate q_1 is constant is of the self-similar form $T = T(\eta)$, involving the similarity variable $\eta = \mathrm{r}/(\alpha_o \mathrm{t})^{1/2}$, so that $T(\eta)$ is given by the solution of

$$\left[\eta^{-1}(q + \eta T^\sigma T_\eta) - \frac{\eta}{2}\right] T_\eta = T\eta^{-1}(\eta T^\sigma T_\eta)_\eta \left\{ \begin{array}{ll} \eta = 0 : & q + \eta T^\sigma T_\eta = 0, \\ \eta = \infty : & T = 1. \end{array} \right. \tag{3.1}$$

To simplify the notation throughout the text, subscripts will be utilized to denote differentiation with respect to a given variable, so that, for instance, $T_\eta = dT/d\eta$ in the above equation. Apart from the thermal-conductivity exponent σ, only the dimensionless heating rate

$$q = \frac{\mathrm{q}_1}{2\pi k_o T_o} \tag{3.2}$$

enters as a parameter in (3.1). As can be seen, besides the thermal-expansion velocity

$$u = \frac{\mathrm{v}}{(\alpha_o/\mathrm{t})^{1/2}} = \frac{1}{\eta}(q + \eta T^\sigma T_\eta), \tag{3.3}$$

the convective term incorporates an apparent negative velocity $-\eta/2$ due to the growing length scale used in the definition of η.

Sample distributions of $T(\eta)$ are shown in Figure 1 for different values of q, with a value $\sigma = 0.5$ adopted in the calculations for the temperature dependence of the conductivity. Integration by a shooting method was initiated near $\eta = 0$, where the temperature profile is of the form

$$T^{\sigma+1} \simeq -(\sigma + 1) q \ln \eta + B, \tag{3.4}$$

with $B(q, \sigma)$ representing an unknown constant that was varied in the shooting procedure to satisfy the boundary condition $T = 1$ at $\eta = \infty$. The resulting value of B is shown as an inset in Figure 1 for $\sigma = (0, 0.5, 1.0)$. Note that the local high-temperature description (3.4) can be of interest for the analysis of some related problems, such as the ignition of a reactive gas mixture by hot wires or by laser beams [10].

The temperature distribution can be used to determine from (3.3) the gas velocity induced by thermal expansion. This velocity is zero at the heat source and also at $\eta = \infty$ and reaches a maximum at an intermediate location, a result clearly seen in the velocity profiles exhibited in Figure 2. The effect of the heat source on the far field is that of a volumetric source of fluid, inducing radial velocities that decay according to $u \simeq q/\eta$ for $\eta \gg 1$.

The solution corresponding to small heating rates can be determined by introducing an expansion for $T - 1$ in increasing powers of q. Since in this case both ρ and k change by only a small amount from their unperturbed values ρ_o and k_o, the resulting solution is, in the first approximation, that corresponding to a solid with constant

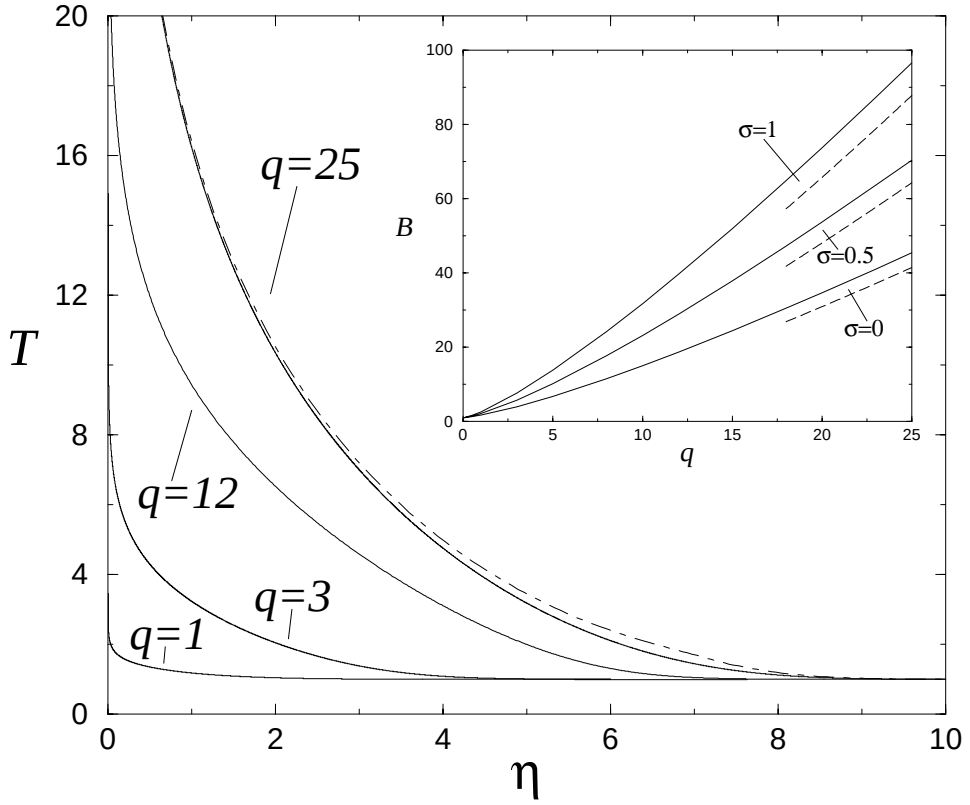


FIG. 1. *The temperature profiles obtained by integration of* (3.1) *with* $\sigma = 0.5$ *(solid lines) and from the large-q composite expansion (dot-dashed lines); the inset shows the variation with q of the constant B, along with the large-q prediction* $B = q[b + (\sigma + 1)\ln(q)/2]$.

thermal conductivity $(T-1)/q = \frac{1}{2}E_1(\eta^2/4)$ (see [4]), where E_1 is the exponential integral function [1]. The effect of the thermal expansion emerges in the solution, giving a small modification of order q^2 to the temperature increment $T-1$ and inducing small radial velocities, of order q, that can be determined from (3.3) to give $u = (q/\eta)[1-\exp(-\eta^2/4)]$. This description can be expected to fail as $T-1$ increases to values of order unity for $\eta \to 0$, in an exponentially small region around the axis corresponding to $\eta \sim \exp(-1/q)$. This region can be studied by employing $\ln(\eta)/q$ as an appropriately stretched coordinate, an analysis that gives $T^{\sigma+1} = 1-(\sigma+1)q\ln\eta$ as the leading-order representation for the temperature. This is in agreement with the results shown in the inset of Figure 1, where the constant B approaches unity as $q \to 0$.

The analysis of the limit of large heat release rates, $q \gg 1$, is more complicated and requires consideration of separate spatial regions. As seen in Figure 1, both the extent of the heated domain and the characteristic value of the temperature grow with increasing values of q. A simple order-of-magnitude analysis of (3.1) reveals that the rescaled variables $\xi = q^{-1/2}\eta$ and $\theta = q^{-1/(\sigma+1)}T$ are appropriate replacements for η and T in this limit of large q. Use of these alternative variables enables (3.1) to

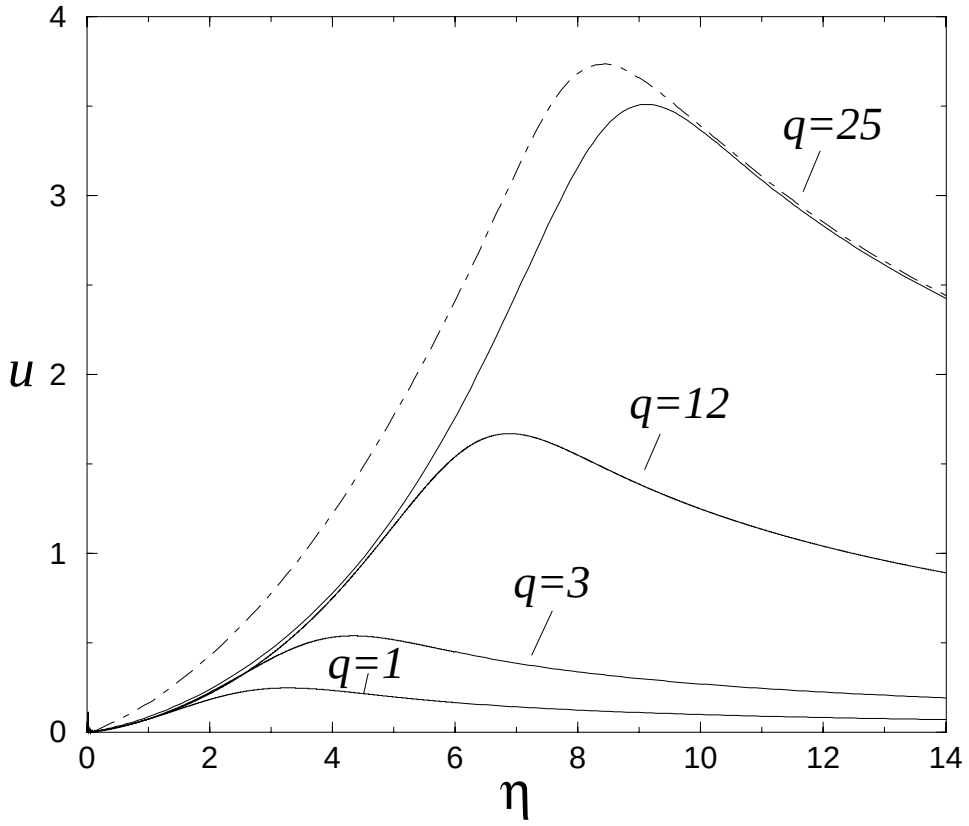


FIG. 2. *The velocity distribution* (3.3) *for* $\sigma = 0.5$ *evaluated with the temperature profiles determined numerically (solid lines) and with the large-q composite expansion (dot-dashed lines).*

be written in the form

$$\left[\xi^{-1}(1+\xi\theta^{\sigma}\theta_{\xi})-\frac{\xi}{2}\right]\theta_{\xi}=\theta\xi^{-1}(\xi\theta^{\sigma}\theta_{\xi})_{\xi}\begin{cases}\xi=0:\xi\theta^{\sigma}\theta_{\xi}=-1,\\ \xi=\infty:\theta=q^{-1/(\sigma+1)}.\end{cases}\tag{3.5}$$

3.1. The high-temperature region. Introducing an expansion of the form

$$\theta=\theta_0+q^{-\mu_1}\theta_1+\cdots\tag{3.6}$$

into (3.5) produces a series of problems that can be solved sequentially, with the order μ_1 for the first-order correction to the leading-order result being determined in the course of the analysis.

The problem emerging at leading order for the function θ_0,

$$\left[\xi^{-1}(1+\xi\theta_0^{\sigma}\theta_{0\xi})-\frac{\xi}{2}\right]\theta_{0\xi}=\theta_0\xi^{-1}(\xi\theta_0^{\sigma}\theta_{0\xi})_{\xi}\begin{cases}\xi=0:\xi\theta_0^{\sigma}\theta_{0\xi}=-1,\\ \xi=\sqrt{2}:\theta_0=0,\end{cases}\tag{3.7}$$

has a front solution that neatly defines the hot region. The location $\xi=\sqrt{2}$ of the front is determined a priori from inspection of (3.7) by noting that heat conduction vanishes as the temperature approaches its zero boundary value, so that convection remains as the only transport mechanism there. Therefore, the leading edge of the

temperature distribution must lie at $\xi = \sqrt{2}$, where the positive thermal-expansion velocity $U = q^{-1/2}u = \xi^{-1}(1 + \xi\theta^{\sigma}\theta_{\xi}) \simeq \xi^{-1}$ equals the negative apparent velocity $-\xi/2$ associated with the growing length scale. The front nature of the solution is clearly a result of the vanishing boundary temperature seen with the scales of this high-temperature region, as occurs in other problems of high-temperature hydrodynamics [13, 15, 22, 23].

The resulting function θ_0 is shown in Figure 3 for four different values of the thermal-conductivity exponent $\sigma = (0, 0.5, 1.0, 2.5)$. The numerical integration was started at $\xi \ll 1$, where $\theta_0^{\sigma+1} \simeq -(\sigma+1)\ln\xi + b$, with $b = (0.0477, 0.1557, 0.2914, 0.7637)$ for $\sigma = (0, 0.5, 1.0, 2.5)$. Note that, in terms of this shooting parameter, the constant B appearing in (3.4) can be expressed in the form $B = q[b + (\sigma + 1)\ln(q)/2]$, an asymptotic prediction tested in the inset of Figure 1. The temperature profiles are seen to approach the cold boundary $\xi = \sqrt{2}$ according to

$$\theta_0 = \left(\frac{1+\sigma}{1-\sigma}\right)^{1/(1+\sigma)} (\sqrt{2} - \xi)^{2/(1+\sigma)} \tag{3.8}$$

if $\sigma < 1$, according to

$$\theta_0 = \sqrt{2}(\sqrt{2} - \xi)\left[\ln\left(\frac{1}{\sqrt{2} - \xi}\right)\right]^{1/2} \tag{3.9}$$

if $\sigma = 1$, and according to

$$\theta_0 = E(\sqrt{2} - \xi)^{1/\sigma} \tag{3.10}$$

if $\sigma > 1$, where E is a constant to be determined as part of the numerical integration (e.g., $E = 2.7449$ for $\sigma = 5/2$).

As previously mentioned, the order μ_1 of the first-order correction to the leading-order results must be determined as part of the solution. Although the boundary condition at $\xi = \infty$ given in (3.5) suggests $\mu_1 = 1/(\sigma + 1)$, corresponding to a correction in temperature T of order unity, it is shown below that the necessary correction is in fact larger when $\sigma < 1$. The cases $\sigma = 1$ and $\sigma > 1$, which give $\mu_1 = 1/(\sigma + 1)$, are treated separately in the appendixes.

The function θ_1 satisfies the equation

$$\left(1 - \frac{\xi^2}{2}\right)\theta_{1\xi} = 2\theta_0(\xi\theta_0^{\sigma-1}\theta_{0\xi})_{\xi}\theta_1 + \theta_0^2(\xi\theta_0^{\sigma-1}\theta_{1\xi})_{\xi} + \theta_0^2(\xi(\sigma-1)\theta_0^{\sigma-2}\theta_{0\xi}\theta_1)_{\xi} \tag{3.11}$$

obtained from linearizing (3.5) about θ_0. The corresponding boundary conditions are

$$\xi = 0: \quad \xi\theta_0^{1+\sigma}\theta_{1\xi} - \sigma\theta_1 = 0 \tag{3.12}$$

and

$$\xi = \sqrt{2}: \quad \theta_1 = 0. \tag{3.13}$$

In addition to the trivial solution $\theta_1 = 0$, for each value of μ_1 the above problem admits a single nontrivial solution that can be determined aside from an arbitrary multiplicative factor. To discriminate the value of μ_1, one needs to investigate the corner layer that appears at distances of order $q^{-1/2}$ about $\xi = \sqrt{2}$, where the temperature becomes of order unity.

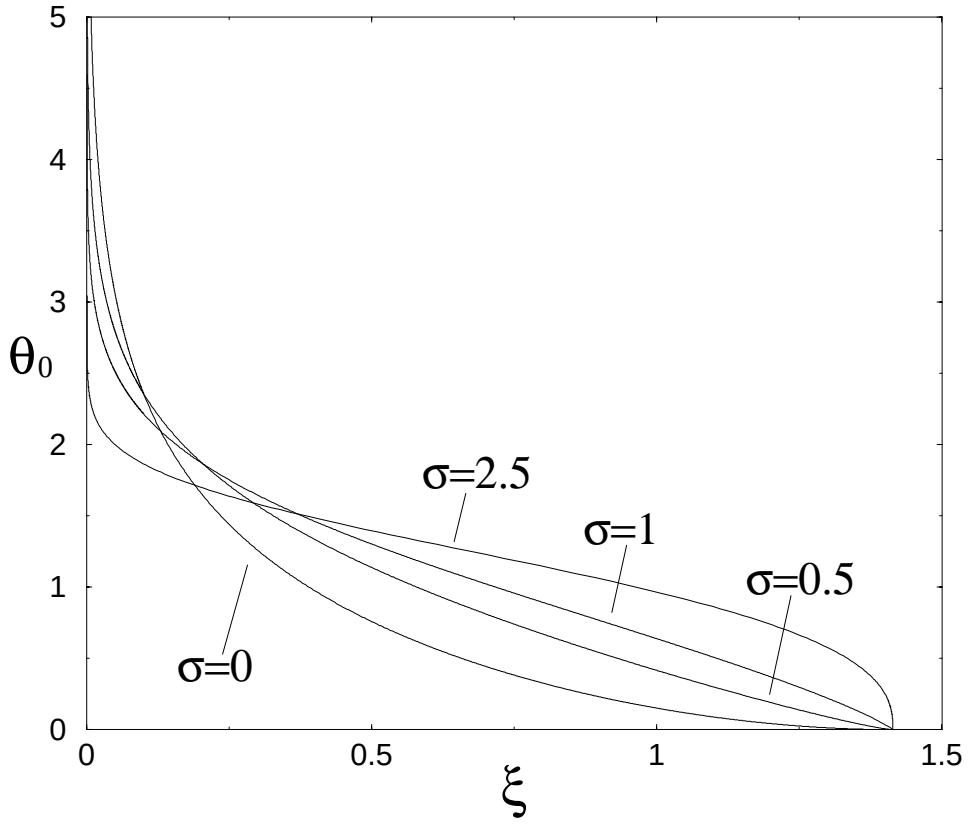


FIG. 3. *The temperature profiles obtained by integration of* (3.7) *for* $\sigma = (0, 0.5, 1.0, 2.5)$.

3.2. The corner layer. Around $\xi = \sqrt{2}$ the temperature must evolve from the cold boundary distribution given in (3.8) to the final asymptotic value $T = 1$. The description of the resulting corner layer must make use of the translated coordinate $\chi = \sqrt{2q} - \eta$. At leading order the problem becomes

$$T^2(T^{\sigma-1}T_\chi)_\chi + \chi T_\chi = 0 \begin{cases} \chi \to -\infty : T = 1, \\ \chi \to \infty : T \to \left(\frac{1+\sigma}{1-\sigma}\right)^{1/(1+\sigma)} \chi^{2/(1+\sigma)}. \end{cases} \tag{3.14}$$

The solution to this problem, which is given in Figure 4, determines in particular the asymptotic behavior for $\chi \to \infty$, where the temperature is seen to approach only slowly its boundary value according to

$$T - \left(\frac{1+\sigma}{1-\sigma}\right)^{1/(1+\sigma)} \chi^{2/(1+\sigma)} = A\chi^{(2-\sigma-\sqrt{2-\sigma^2})/(1+\sigma)}, \tag{3.15}$$

with A being a constant determined as part of the integration. Sample values are $A = (3.816, 4.5080, 5.664, 9.353)$ for $\sigma = (0, 0.25, 0.5, 0.75)$.

3.3. Uniformly valid description. Matching the solution given in (3.15) with the outer expansion $\theta = \theta_0 + t^{-\mu_1}\theta_1 + \cdots$ gives

$$\mu_1 = \frac{\sigma + \sqrt{2 - \sigma^2}}{2(1+\sigma)} \tag{3.16}$$

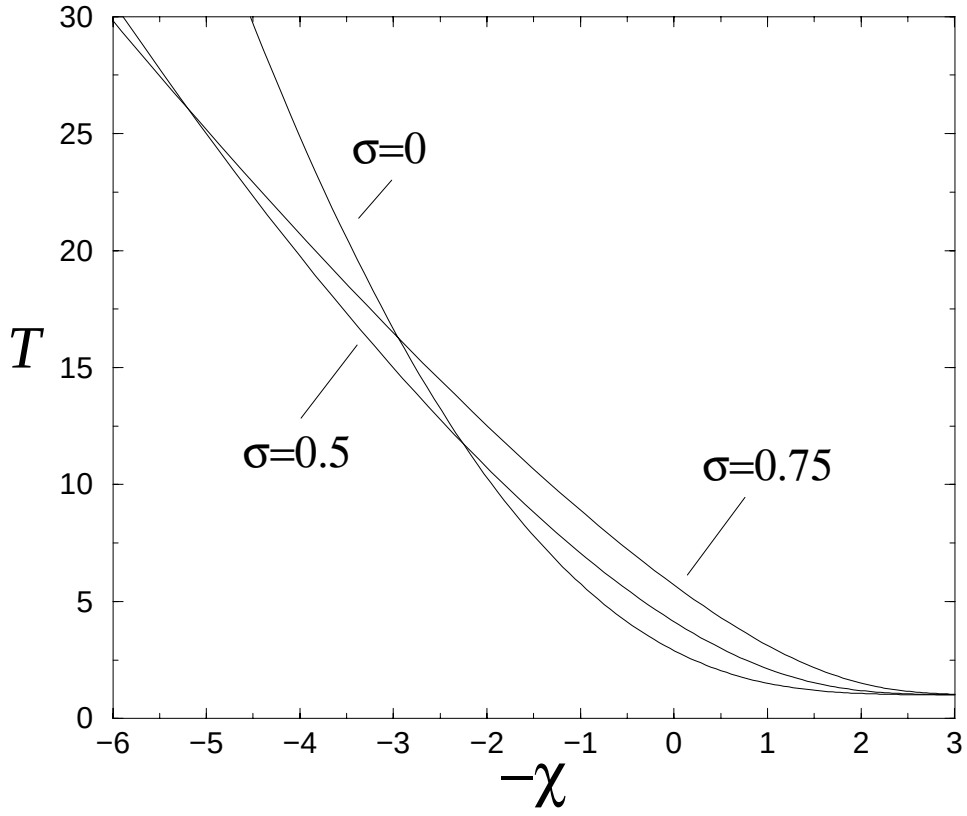


FIG. 4. *The temperature profile across the transition layer for different values of* σ.

for the order of the correction. Furthermore, the matching procedure provides

$$\xi \to \sqrt{2}: \quad \theta_1 = A(\sqrt{2}-\xi)^{(2-\sigma-\sqrt{2-\sigma^2})/(1+\sigma)} \tag{3.17}$$

as a replacement for (3.13), thereby removing all previously noted arbitrariness; i.e., (3.11) subject to (3.12) and (3.17) has a unique solution that must be determined by numerical integration.

A uniformly valid description for the temperature field can be obtained by combining the outer expansion for θ with the corner-layer profile according to the composite expansion

$$\begin{aligned} T(\eta) &= q^{1/(\sigma+1)}\left[\theta_0(\xi)+q^{-\mu_1}\theta_1(\xi)\right] \\ &+ T(\chi) - H(\sqrt{2}-\xi)\left[\left(\frac{1+\sigma}{1-\sigma}\right)^{1/(1+\sigma)}\chi^{2/(1+\sigma)} - A\chi^{(2-\sigma-\sqrt{2-\sigma^2})/(1+\sigma)}\right], \end{aligned} \tag{3.18}$$

where $H(\sqrt{2}-\xi)$ is the Heaviside function with origin $\xi=\sqrt{2}$ and $T(\chi)$ is the temperature profile across the corner layer, with the rescaled variables $\xi = q^{-1/2}\eta$ and $\chi = \sqrt{2q}-\eta$ being those utilized above for the outer region and for the corner layer, respectively. The resulting temperature profile and the accompanying velocity

profile, determined by straightforward substitution of (3.18) into (3.3), are plotted in Figures 1 and 2, showing reasonable agreement for the relatively large value of $q = 25$ considered.

4. Constant heat flux from a plane wall. Use of the characteristic scales identified above in (1.3) and (1.4) provides $t = \alpha_o[\mathrm{q}_0/(k_oT_o)]^2\mathrm{t}$, $r = [\mathrm{q}_0/(k_oT_o)]\mathrm{r}$, and $v = \mathrm{v}/[\mathrm{q}_0\alpha_o/(k_oT_o)]$ as dimensionless variables to describe the planar source. As first shown by Clarke, Kassoy, and Riley [5], the problem then reduces to that of integrating

$$\frac{\partial T}{\partial t} - T^2\frac{\partial}{\partial r}\left[\frac{1}{T}\left(1 + T^\sigma\frac{\partial T}{\partial r}\right)\right] = 0 \tag{4.1}$$

with initial condition

$$t = 0, \qquad 0 \le r < \infty : \quad T = 1 \tag{4.2}$$

and boundary conditions

$$t > 0 \begin{cases} r = 0 : & T^\sigma \partial T/\partial r = -1, \\ r = \infty : & T = 1, \end{cases} \tag{4.3}$$

while the velocity can be computed from (2.7) to give

$$v = 1 + T^\sigma\frac{\partial T}{\partial r}. \tag{4.4}$$

4.1. Temperature and velocity distributions. As can be seen, σ is the only parameter left in the problem. An exact solution is known only for $\sigma = 1$ (see [5]), a case for which the density-weighted coordinate $\mathrm{d}z = T^{-1}\mathrm{d}r$, often introduced for the analysis of variable-density boundary layers [17, 21], reduces (4.1)–(4.3) to the constant-density problem (see [4]), thereby yielding

$$T = 1 + 2t^{1/2}\,\mathrm{i}^1\mathrm{erfc}\left(\frac{z}{2t^{1/2}}\right) \quad \text{and} \quad r = z + t\left[1 - 4\,\mathrm{i}^2\mathrm{erfc}\left(\frac{z}{2t^{1/2}}\right)\right] \tag{4.5}$$

as an implicit representation for the temperature profile $T(r)$, where $\mathrm{i}^1\mathrm{erfc}$ and $\mathrm{i}^2\mathrm{erfc}$ denote repeated integrals of the complementary error function erfc (see [1]). Correspondingly, the velocity profile (4.4) reduces to $v = 1 - \mathrm{erfc}\left[\int_0^r T^{-1}\mathrm{d}r/(2t^{1/2})\right]$.

For $\sigma \neq 1$, the problem needs numerical integration. To handle the unbounded value of $\partial T/\partial t$ at $t = 0$, the initial condition (4.2) must be replaced in the numerical integrations with the leading-order representation of the temperature profile for $t \ll 1$, when the temperature increase from the initial value $T = 1$ is small, of order $t^{1/2}$, and is seen to be confined to a thin layer of characteristic thickness $t^{1/2}$ located in the vicinity of the wall. To describe this initial period it is convenient to introduce the self-similar variables $r/t^{1/2}$ and $(T-1)/t^{1/2}$ into (4.1)–(4.3), yielding in the first approximation the constant-density result (see [4]),

$$\frac{(T-1)}{t^{1/2}} = 2\,\mathrm{i}^1\mathrm{erfc}\left[\frac{r}{(2t^{1/2})}\right], \tag{4.6}$$

while the initial velocity distribution becomes

$$v = 1 - \mathrm{erfc}\left(\frac{r}{2t^{1/2}}\right). \tag{4.7}$$

Characteristic temperature profiles obtained for $\sigma = 0.5$ by numerical integration of (4.1) with boundary conditions (4.3) and with the initial profile (4.6) evaluated at $t \ll 1$ are shown in Figure 5, along with the accompanying velocity profiles determined by evaluating (4.4). The profiles at $t = 0.4$ are compared with the asymptotic predictions

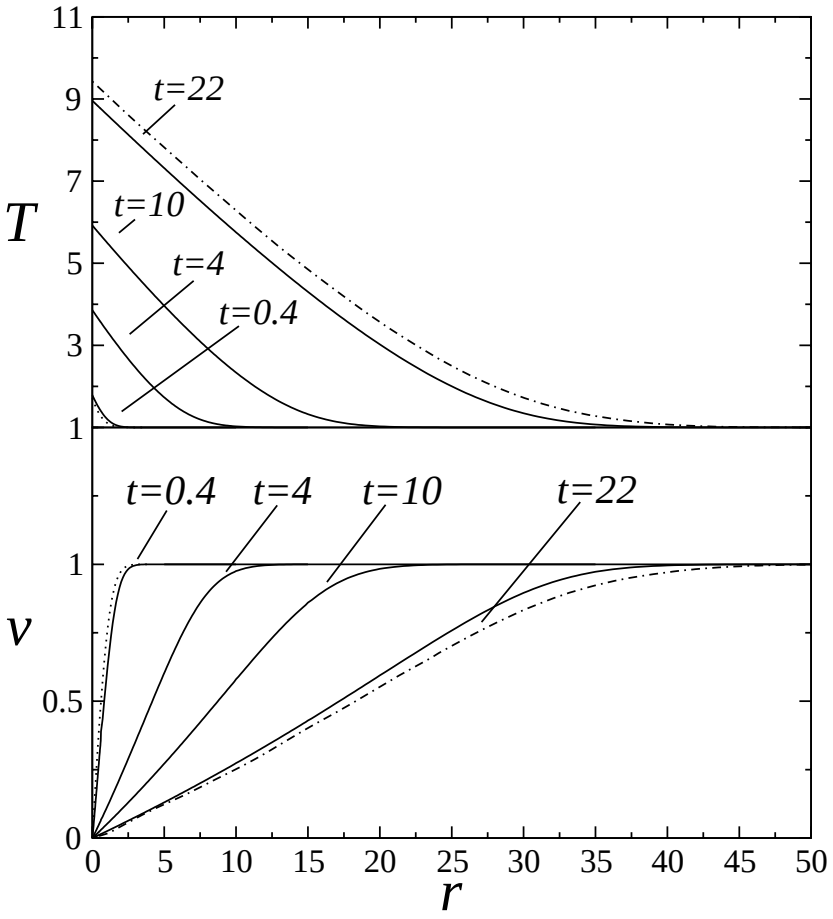


FIG. 5. *The temperature and velocity profiles obtained by integration of* (4.1)–(4.3) *for* $\sigma = 0.5$ *(solid lines); the profiles at* $t = 0.4$ *are compared with the asymptotic predictions given in* (4.6) *and* (4.7) *for* $t \ll 1$, *and the profiles at* $t = 22$ *are compared with the asymptotic predictions for* $t \gg 1$.

for small times given in (4.6) and (4.7), while the profile $t = 22$ is compared with the asymptotic prediction for $t \gg 1$, to be developed below.

In this planar case, the temperature remains bounded everywhere, growing with time. The evolution of the maximum temperature T_w attained at the wall is shown in Figure 6 for $\sigma = 0$ and $\sigma = 0.5$. The numerical solution is compared with the asymptotic description for small t,

$$T_w = 1 + \frac{2}{\pi^{1/2}} t^{1/2}, \tag{4.8}$$

obtained by evaluating (4.6) at $r = 0$, and also with the results given below for $t \gg 1$. Note that (4.8) gives exactly the wall temperature at all times when $\sigma = 1$, as can be seen by evaluating (4.5) at $z = 0$.

4.2. Solution for $t \gg 1$. The solution in this limit parallels that obtained above for the planar case in the limit $q \gg 1$. As previously mentioned, the temperature and the extent of the heated region continues to increase as time progresses. An order-of-magnitude analysis of (4.1) and (4.3) suggests the use of the modified variables

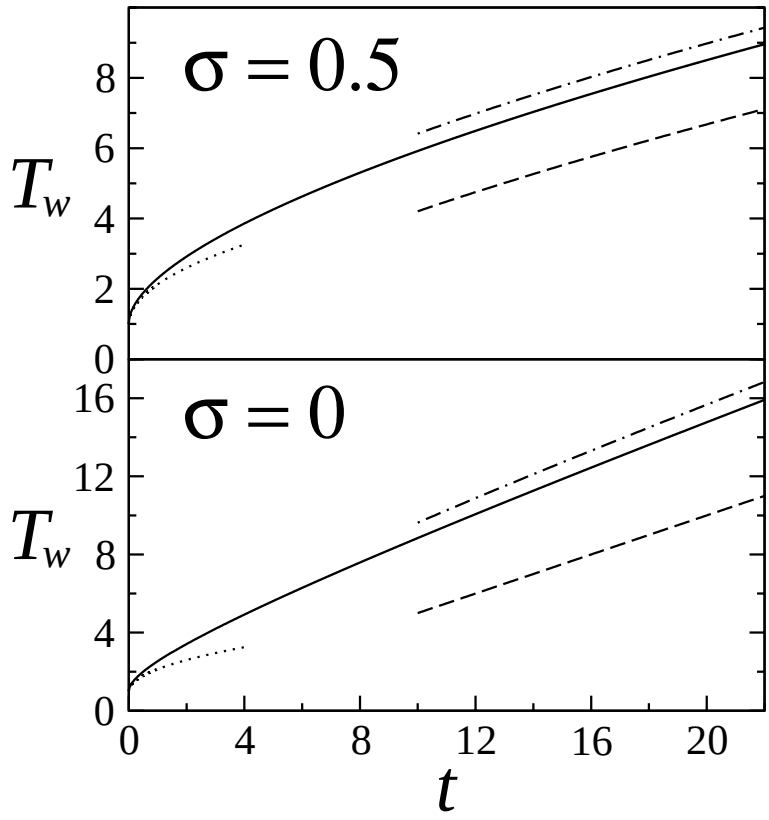


FIG. 6. *The evolution with time of the wall temperature for different values of σ obtained from numerical integration of* (4.1) *(solid lines), from the short-time prediction* (4.8) *(dotted lines), from the leading-order long-time prediction $T_w = \theta_0(0)t^{1/(\sigma+1)}$ (dashed lines), and from the two-term expansion $T_w = t^{1/(\sigma+1)}[\theta_0(0) + \theta_1(0)t^{-\mu_0}]$ (dot-dashed lines).*

$\theta = T/t^{1/(\sigma+1)}$ and $x = r/t$, of order unity, for the analysis of the limit $t \gg 1$, so that (4.1) and (4.3) take the form

$$t\theta_t + (1 + \theta^\sigma\theta_x - x)\theta_x + \frac{\theta}{\sigma+1} = \theta(\theta^\sigma\theta_x)_x \tag{4.9}$$

and

$$\begin{cases} x = 0: & \theta^\sigma\theta_x = -1, \\ x = \infty: & \theta = 1/t^{1/(\sigma+1)}, \end{cases} \tag{4.10}$$

while the velocity is given by $v = 1 + \theta^\sigma\theta_x$. As can be seen, because of the rescaled variables employed in this limit, an additional convective term $-x\theta_x$ appears in (4.9), together with a damping term $\theta/(\sigma+1)$ associated with the growing temperature scale.

Introducing the expansion $\theta(x,t) = \theta_0(x) + t^{-\mu_0}\theta_1(x) + \cdots$ permits us to solve the problem in a sequential manner, with the unknown value of μ_0 being determined as part of the asymptotic development as shown below. The function θ_0 is obtained from

$$(1 + \theta_0^\sigma\theta_{0x} - x)\theta_{0x} + \frac{\theta_0}{\sigma+1} = \theta_0(\theta_0^\sigma\theta_{0x})_x \begin{cases} x = 0: & \theta_0^\sigma\theta_{0x} = -1, \\ x = 1: & \theta_0 = 0. \end{cases} \tag{4.11}$$

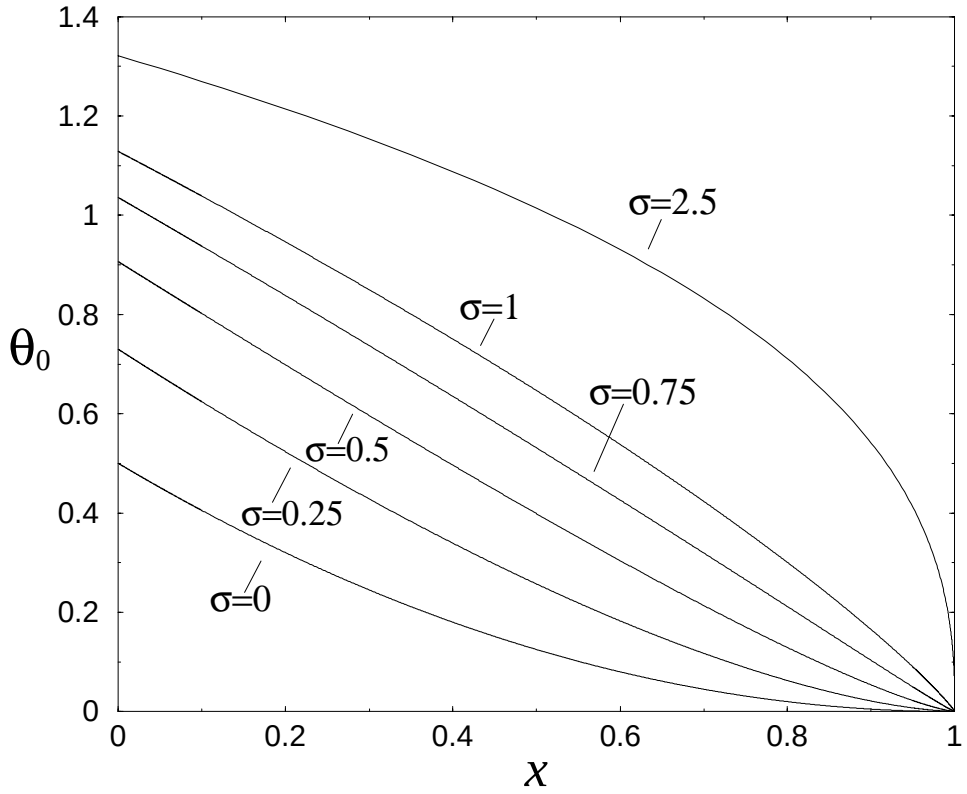


FIG. 7. *The temperature profile* θ_0 *obtained from* (4.9) *for different values of* σ.

The problem (4.11) has a front solution, similar to that seen in the axisymmetric case for $q \gg 1$. Since both the heat-conduction term, $\theta_0(\theta_0^\sigma \theta_{0x})_x$, and the damping term, $\theta_0/(\sigma+1)$, vanish as the temperature approaches its zero boundary value, the front of the temperature distribution must lie at $x = 1$, where the positive velocity due to thermal expansion, $v \simeq 1$, equals the apparent negative velocity $-x$.

When $\sigma = 0$, the problem given in (4.11) has the exact solution (see [5])

$$\begin{cases} \theta_0 = (1-x)^2/2, & v = x, \quad \text{for} \quad 0 < x < 1, \\ \theta_0 = 0, & v = 1, \quad \text{for} \quad 1 < x. \end{cases} \tag{4.12}$$

Another exact solution appears when $\sigma = 1$, as can be seen by rewriting (4.5) in terms of the intermediate coordinate $\bar{z} = z/(2t^{1/2})$ to give the implicit representation [5]

$$\theta_0 = 2\,\mathrm{i}^1\mathrm{erfc}(\bar{z}) \quad \text{and} \quad x = 1 - 4\,\mathrm{i}^2\mathrm{erfc}(\bar{z}). \tag{4.13}$$

Numerical integration is necessary to compute profiles of θ_0 when $\sigma \neq (0, 1)$. The temperature profiles (4.12) and (4.13) are shown in Figure 7 along with the numerical results corresponding to $\sigma = (0.25, 0.5, 0.75, 2.5)$. A shooting technique started at $x = 0$ was used for the integration of (4.11), with $\theta(0)$ utilized as the shooting parameter to be varied in the iteration procedure. This initial value, which equals $\theta_0(0) = 0.5$ for $\sigma = 0$, $\theta_0(0) = 2/\pi^{1/2}$ for $\sigma = 1$, and $\theta_0(0) = (0.7298, 0.9063, 1.0356, 1.3212)$ for $\sigma = (0.25, 0.5, 0.75, 2.5)$, determines the leading-order prediction for the wall temperature

$T_w = \theta_0(0)t^{1/(\sigma+1)}$. The comparisons with the results of the numerical integrations for large times, shown in Figure 6, clearly indicate that the asymptotic description must be carried on to the following order for increased accuracy.

This first-order correction $\theta_1(x)$ must satisfy the conservation equation (4.9) linearized about θ_0

$$\left(\frac{1}{\sigma+1} - \mu_0\right)\theta_1 + (1-x)\theta_{1x} = \theta_0^\sigma \left\{ \left[(\sigma+1)\theta_{0xx} - \sigma(1-\sigma)\frac{\theta_{0x}^2}{\theta_0} \right] \theta_1 - 2(1-\sigma)\theta_{0x}\theta_{1x} + \theta_0\theta_{1xx} \right\}, \tag{4.14}$$

subject to the boundary conditions

$$x = 0: \quad \theta_0^{1+\sigma}\theta_{1x} - \sigma\theta_1 = 0 \tag{4.15}$$

and

$$x = 1: \quad \theta_1 = 0. \tag{4.16}$$

As occurred before with the perturbation problem (3.11)–(3.13), for each value of μ_0 the problem (4.14)–(4.16) admits a single nontrivial solution that can be determined aside from an arbitrary multiplicative factor. The value of μ_0 is determined from matching the first two terms of the high-temperature distribution $\theta = \theta_0 + t^{-1/(\sigma+1)}\theta_1$ with the leading-order temperature representation across the corner layer, located around $x = 1$. It is remarkable that the first-order correction can be determined without taking into account the initial non–self-similar growth period, thereby indicating that memory effects emerge in the asymptotic development for large times only at higher orders. The associated corrections should be computed from matching the asymptotic results with the numerical computations for $t \sim 1$, a development not pursued further here.

As seen before for the line source, the structure of the solution for $\sigma = 1$ and $\sigma > 1$ is different from that encountered with $\sigma < 1$. The analyses of the former solutions, which are given in the appendixes, reveal that $\mu_0 = 1/(\sigma+1)$, corresponding to a correction in temperature T of order unity. The corrections are larger when $\sigma < 1$, when the leading-order temperature profile approaches the boundary $x = 1$ according to the local description

$$\theta_0 = \left(\frac{(1+\sigma)}{2(1-\sigma)}\right)^{1/(\sigma+1)} (1-x)^{2/(\sigma+1)}, \tag{4.17}$$

as can be obtained from (4.11). The corner layer, where the temperature T is of order unity, corresponds to distances $(1-x)$ of order t^{-1}. Introducing the coordinate $\chi = (t/2)^{1/2}(1 - r/t)$ reduces the leading-order problem to that given in (3.14), whose solution matches asymptotically with the boundary distribution (4.17). Furthermore, inspection of (3.15) indicates that the order of the first-order correction must be

$$\mu_0 = \frac{\sigma + \sqrt{2-\sigma^2}}{2(1+\sigma)} \tag{4.18}$$

to complete the matching, and that

$$x = 1: \quad \theta_1 = A[(1-x)/\sqrt{2}]^{(2-\sigma-\sqrt{2-\sigma^2}/(1+\sigma)} \tag{4.19}$$

must replace (4.16) to provide uniqueness for the solution to (4.14). In general, numerical integration is required to compute θ_1, the only exception being the case $\sigma = 0$, for which the exact solution

$$\theta_1 = \left(\frac{A}{2^{1-1/\sqrt{2}}}\right)\left[(1-x)^{2-\sqrt{2}} - \frac{2-\sqrt{2}}{1+\sqrt{2}}(1-x)^{1+\sqrt{2}}\right] \tag{4.20}$$

is available.

The results of the asymptotic analysis can be combined to give a uniformly valid description for the temperature. The corresponding composite expansion is that given in (3.18), with q, ξ, and μ_1 being replaced with t, x, and μ_0, and with the origin for the Heaviside function being $x = 1$. The resulting temperature profile and its accompanying velocity profile are plotted in Figure 5, showing good agreement for the value $t = 22$ considered. The relatively small errors observed, of order unity, correspond to a correction at the following order in the asymptotic analysis, which is not computed here. As seen in Figure 6, an error of order unity is also present in the second-order asymptotic prediction for the wall temperature $T_w = t^{1/(\sigma+1)}(\theta_0(0) + \theta_1(0)t^{-\mu_0})$, where $\theta_1(0) = (2.359, 2.223, 1.927, 1.486)$ for $\sigma = (0, 0.25, 0.5, 0.75)$.

5. The point source of heat. The characteristic scales for this problem, defined in order of magnitude in (1.3) and (1.4), were used to define the dimensionless variables $t = \alpha_o[\mathrm{q}_2/(4\pi k_o T_o)]^{-2}\mathrm{t}$, $r = \mathrm{r}/[\mathrm{q}_2/(4\pi k_o T_o)]$, and $v = \mathrm{v}[\mathrm{q}_2/(4\pi k_o T_o)]/\alpha_o$. The temperature T is determined by integrating

$$\frac{\partial T}{\partial t} - \frac{T^2}{r^2}\frac{\partial}{\partial r}\left[\frac{1}{T}\left(1 + T^\sigma r^2 \frac{\partial T}{\partial r}\right)\right] = 0, \tag{5.1}$$

with initial condition

$$t = 0, \;\; 0 \le r < \infty : \;\; T = 1 \tag{5.2}$$

and boundary conditions

$$t > 0 \begin{cases} r = 0 : & r^2 T^\sigma \partial T/\partial r = -1, \\ r = \infty : & T = 1, \end{cases} \tag{5.3}$$

while the velocity can be computed from

$$v = \frac{1}{r^2}\left(1 + r^2 T^\sigma \frac{\partial T}{\partial r}\right). \tag{5.4}$$

As in the planar case, σ remains as the only parameter left in the problem. Because of the boundary condition at $r = 0$, the temperature profile presents an infinite value at the point source for $t > 0$. Hence, the numerical integration of the problem (5.1)–(5.3) must account for the singular character of the solution near the origin, where

$$T^{\sigma+1} = \frac{(\sigma+1)}{r} + C(t). \tag{5.5}$$

In particular, the initial profile $T = 1$ must be replaced with the leading-order representation emerging for $t \ll 1$.

5.1. Initial temperature growth. As previously anticipated, the structure of the solution in this limit $t \ll 1$ is that found for $q \gg 1$ in the axisymmetric case and for $t \gg 1$ in the planar case, that is, a neatly defined central region of high temperature separated from the outer cold gas at temperature $T = T_o$ by a thin corner layer of warm fluid. The appropriate scales for length and temperature to describe the hot region, $t^{1/3}$ and $t^{-1/[3(\sigma+1)]}$, can be anticipated from the balance of the three terms in (5.1). Correspondingly, the associated rescaled variables $\theta = t^{1/[3(\sigma+1)]}T$ and $y = r/t^{1/3}$ reduce (5.1) and (5.3) to

$$t\theta_t + \left[y^{-2}(1 + y^2\theta^\sigma\theta_y) - \frac{y}{3}\right]\theta_y - \frac{\theta}{3(\sigma+1)} = \theta y^{-2}(y^2\theta^\sigma\theta_y)_y \tag{5.6}$$

and

$$\begin{cases} y = 0: & y^2\theta^\sigma\theta_y = -1, \\ y = \infty: & \theta = t^{1/[3(\sigma+1)]}. \end{cases} \tag{5.7}$$

Because of the growing length scale that has been introduced, besides the rescaled thermal-expansion velocity of order unity, $t^{2/3}v = y^{-2}(1+y^2\theta^\sigma\theta_y)$, there exists in (5.6) a negative apparent velocity $-y/3$. Similarly, the decreasing scale used for the temperature leads to the negative damping term $-\theta/[3(\sigma+1)]$.

Introducing the expansion $\theta(y,t) = \theta_0(y) + t^{\mu_2}\theta_1(y) + \cdots$ yields at leading order

$$\left[y^{-2}(1 + y^2\theta_0^\sigma\theta_{0y}) - \frac{y}{3}\right]\theta_{0y} - \frac{\theta_0}{3(\sigma+1)} = \theta_0 y^{-2}(y^2\theta_0^\sigma\theta_{0y})_y \tag{5.8}$$

to be integrated with boundary conditions

$$\begin{cases} y = 0: & y^2\theta_0^\sigma\theta_{0y} = -1, \\ y = 3^{1/3}: & \theta_0 = 0. \end{cases} \tag{5.9}$$

As before, the balance between the thermal-expansion velocity and the apparent velocity determines the location of the front $y = 3^{1/3}$. A shooting method was used to integrate (5.8). Integration was initiated near $y = 0$, where the temperature profile is of the form $\theta_0^{\sigma+1} = (\sigma+1)/y + c$. The unknown shooting parameter c was varied in the numerical integration to satisfy the boundary condition at $y = 3^{1/3}$, yielding the profiles shown in Figure 8. The negative constant $c = -(0.8390, 1.2173, 1.5806, 2.6397)$ for $\sigma = (0, 0.5.1.0, 2.5)$ provides $C = ct^{-1/3}$ for the initial evolution of the constant C in (5.5).

If $\sigma < 1$, the function θ_0 is seen to approach the boundary according to

$$\theta_0 = \left(\frac{7(1+\sigma)}{6(1-\sigma)}\right)^{1/(1+\sigma)} (3^{1/3} - y)^{2/(1+\sigma)}, \tag{5.10}$$

to be matched with the temperature profile across the transition layer, which is determined at leading order by (3.14), with the similarity coordinate being defined as $\chi = t^{-1/6}(6/7)^{1/2}(3^{1/3} - r/t^{1/3})$. Matching the first two terms in the expansion for θ with (3.15) yields in this case

$$\mu_2 = \frac{\sigma + \sqrt{2 - \sigma^2}}{6(1+\sigma)} \tag{5.11}$$

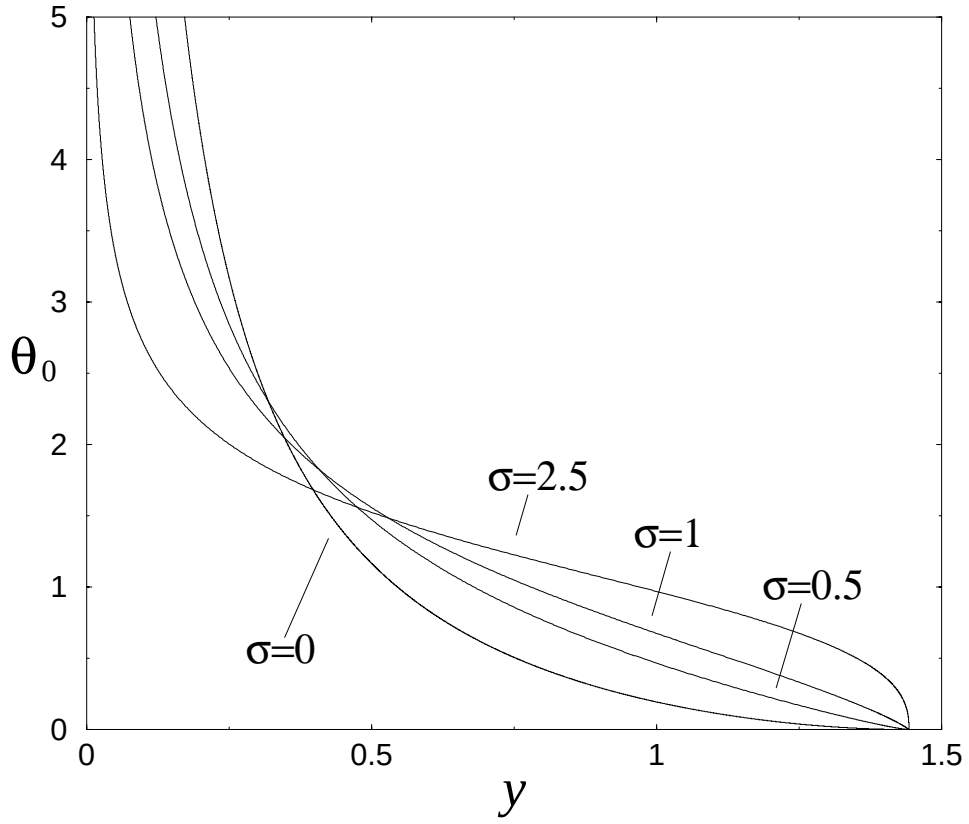


FIG. 8. *The temperature profiles obtained by integration of* (5.8) *for* $\sigma = (0, 0.5, 1.0, 2.5)$.

along with the asymptotic value

$$\theta_1 \to A\left[(7/6)^{1/2}(3^{1/3}-y)\right]^{(2-\sigma-\sqrt{2-\sigma^2})/(1+\sigma)} \tag{5.12}$$

to be used as a boundary condition at $y = 3^{1/3}$ when computing the first-order correction θ_1. The singular case $\sigma = 1$, when

$$\theta_0 = (3^{1/3}-y)\left[\frac{7}{3}\ln\left(\frac{1}{3^{1/3}-y}\right)\right]^{1/2} \tag{5.13}$$

for $0 < 3^{1/3} - y \ll 1$, is described separately in Appendix A, while the the case $\sigma > 1$, when $\theta_0 \propto (3^{1/3}-y)^{1/\sigma}$ near the boundary, is described in Appendix B.

5.2. Temperature and velocity distributions. As previously mentioned, the results of the asymptotic analysis for $t \ll 1$ were employed to enable integrations of (5.1). The two-term expansion $\theta = \theta_0 + t^{\mu_2}\theta_1$ was combined with the solution in the corner layer to provide the corresponding composite expansion, which is that given in (3.18) with the exponents $-1/[3(\sigma+1)]$ and μ_2 and the variables y and t replacing $1/(\sigma+1)$, $-\mu_1$, ξ, and q, respectively, and with the origin for the Heaviside function being $y = 3^{1/3}$. This composite expansion evaluated at $t = 0.01$ was used as an initial condition in the integrations shown in Figure 9, where the temperature

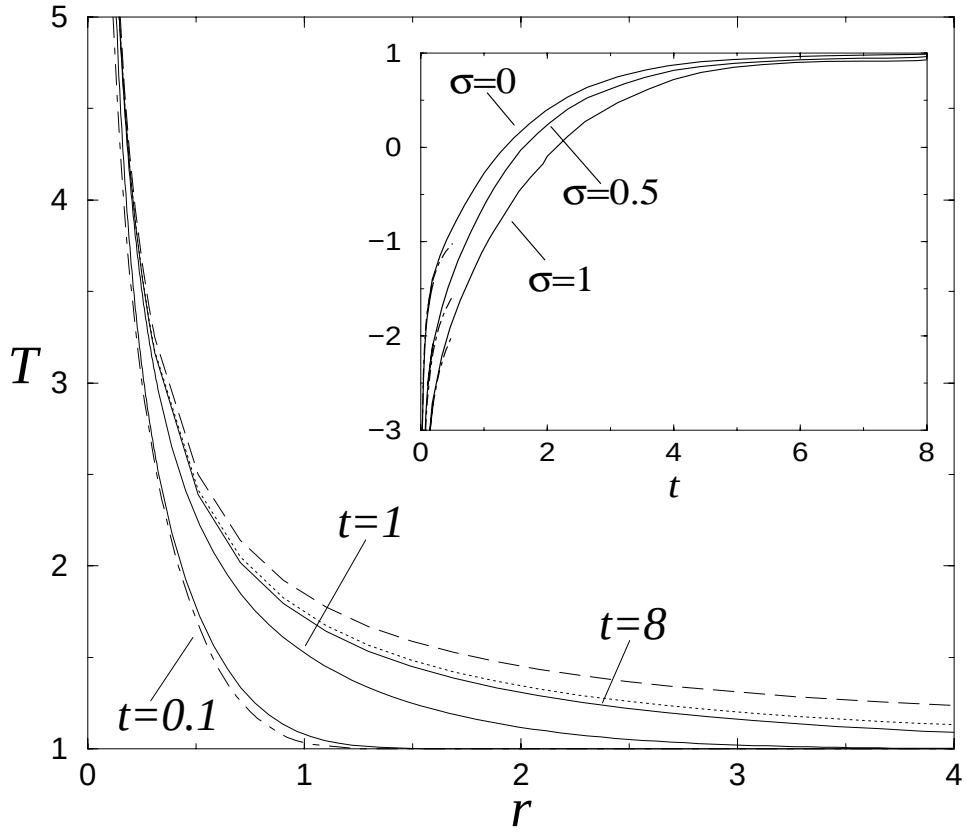


FIG. 9. *The temperature profile corresponding to the final steady solution* (5.14) *with* $\sigma = 0.5$ *(dashed line), along with those obtained by integration of* (5.1) *(solid lines), from the short-time composite expansion (dot-dashed line), and from the long-time quasi-steady expression* (5.16) *(dotted line); the inset shows the variation of the constant* C *together with the short-time prediction* $C = ct^{-1/3}$ *(dot-dashed lines).*

profiles corresponding to $t = (0.1, 1.0, 8.0)$ are shown. As can be observed, the comparison with the short-time composite expansion for $t = 0.1$ still gives reasonably good agreement. For completeness, the plot exhibits in an inset the variation with time of the constant C corresponding to the near-origin temperature distribution (5.5), along with the short-time prediction $C = ct^{-1/3}$.

The temperature profiles can be used in (5.4) to provide the associated velocity profiles, which are shown in Figure 10. The solution is seen to evolve rapidly from the initial large velocities of order $t^{-2/3}$ to the final quasi-stagnant solution corresponding to $t \gg 1$, which is described below.

5.3. Quasi-steady long-time solution. For asymptotically large values of t, the solution evolves to approach the profile

$$T_s = \left(1 + \frac{\sigma + 1}{r}\right)^{1/(\sigma+1)}. \tag{5.14}$$

This steady solution (5.14) and its associated velocity field $v = 0$ are correct to all algebraic orders at distances r of order unity; i.e., the investigation in the limit $t \gg 1$

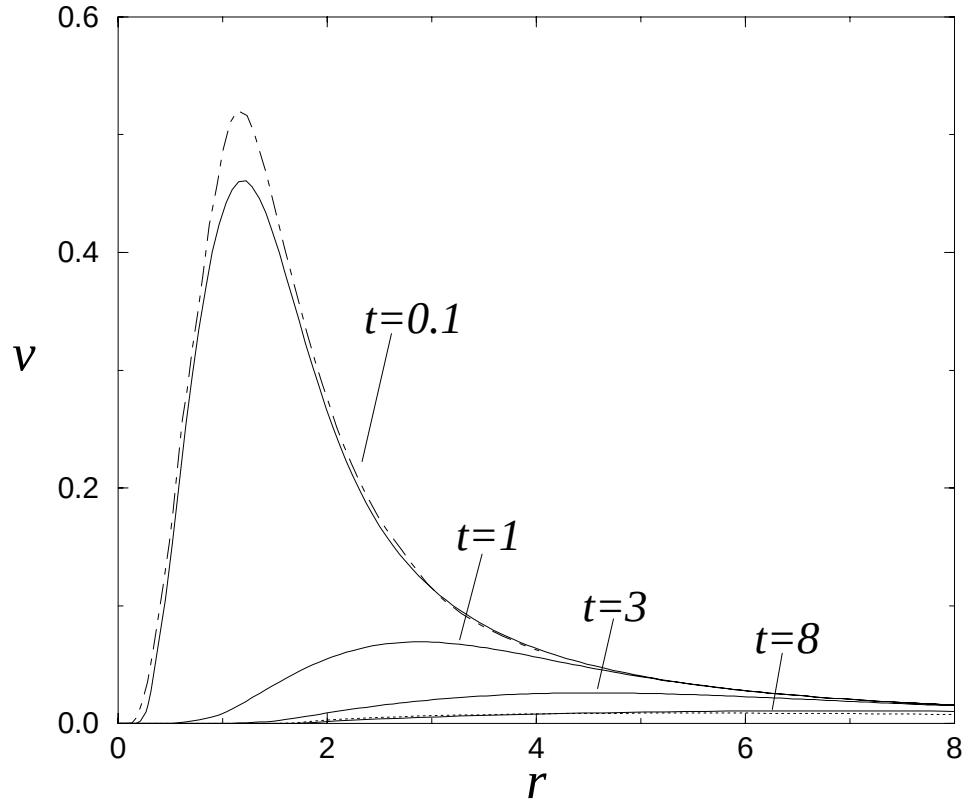


FIG. 10. *The velocity profiles obtained by evaluating* (5.4) *with the numerical temperature profile (solid lines) and with the short-time composite expansion (dot-dashed lines). The dotted line represents the long-time quasi-steady solution* (5.17).

of perturbations of the form $T = T_s(r) + t^{-\alpha}T_\alpha(r)$ yields $T_\alpha = 0$ irrespective of the value of α.

Unsteady effects are seen to enter farther from the heat source, in a far-field region corresponding to distances of order $t^{1/2}$ where only small temperature increments $T - 1$ of order $t^{-1/2}$ exist. To study this far-field region, it is convenient to employ the similarity coordinate $\eta = r/t^{1/2}$, along with the rescaled temperature increment $T - 1 = t^{-1/2}\Theta$, where an expansion of the form $\Theta(\eta, t) = \Theta_0(\eta) + t^{-1/2}\Theta_1(\eta) + \cdots$ is assumed. Introducing these new variables into (5.1) yields at leading order

$$(\eta^2/2)(\Theta_0 + \eta\Theta_{0\eta}) + (\eta^2\Theta_{0\eta})_\eta \begin{cases} \eta \to 0: & \Theta_0 = 1/\eta, \\ \eta \to \infty: & \Theta_0 = 0, \end{cases} \tag{5.15}$$

where the boundary condition as $\eta \to 0$ comes from matching with the steady solution (5.14). Straightforward integration gives $\Theta_0 = \eta^{-1}\mathrm{erfc}(\eta/2)$. Now combining the inner steady-state profile with the far-field transient solution provides the composite expansion $T = T_s + t^{-1/2}(\Theta_0 - 1/\eta)$, which gives the solution for the large-time temperature evolution with errors of order t^{-1}. At the same level of approximation, one may write

$$T = \left[1 + \frac{\sigma + 1}{r}\mathrm{erfc}\left(\frac{r}{2t^{1/2}}\right)\right]^{1/(\sigma+1)} \tag{5.16}$$

for the temperature profile, a compact expression that can be used in (5.4) to obtain the velocity profile

$$v = \frac{1}{r^2}\left[\operatorname{erf}\left(\frac{r}{2t^{1/2}}\right) - \frac{r}{(\pi t)^{1/2}}\exp\left(\frac{-r^2}{4t}\right)\right]. \tag{5.17}$$

Note that, when $\sigma = 0$, (5.16) corresponds to the exact self-similar solution achieved with constant density and constant conductivity [4]. These large-time predictions are compared in Figures 9 and 10 with results of numerical integrations for $t = 8$, yielding reasonably good agreement.

6. Conclusions. The transient, one-dimensional, near-isobaric, buoyancy-free flow field induced by a localized energy source of constant rate has been analyzed for planar, cylindrical, and spherical geometries. The convection induced by thermal expansion is seen to aid the transport of heat away from the source, in a nonlinear process of evolution that has been computed with account taken of the variable thermal conductivity typical of gases. Our study shows a self-similar solution for the line source, with the dimensionless heat release q entering as a parameter, while both the planar source and the point source require consideration of a nonlinear parabolic equation for the time evolution of the temperature.

The analysis reveals that front solutions emerge when the resulting temperatures become much larger than the initial temperature, with the front location being determined a priori from a convective balance. It is shown that the inner structure of the planar thin front, which is identical for all three geometrical configurations, determines the first-order correction in the hot region. Note that front solutions can also be expected to emerge as limiting solutions when a variable heating rate $\mathrm{q_j(t)}$ is applied, a problem to be addressed in future work. In that case, unsteady effects are likely to emerge in the hot region at leading order, while the inner structure of the thin front is expected to evolve in a quasi-steady manner. Also of interest is the investigation of the effect of compressibility on the heat propagation process from point and line sources, as done for planar sources by Clarke, Kassoy, and Riley [6]. Future research should also consider the solution emerging after the heat source is switched off. A related study is that of Meerson [12], who considered the conductive cooling of a gas heated by a localized deposition of heat.

The quantitative information provided here can be of interest, for instance, in analyses of ignition processes of a reactive gas mixture by localized energy sources [10]. The corresponding energy conservation equation should incorporate a heat-release term, and should be supplemented by conservation equations for the chemical species. The ignition process typically involves an initial quasi-frozen period with negligible chemical heat release, in which the description given here holds, followed by a period of significant exothermicity. For instance, for ignition of hydrogen-oxygen mixtures [10, 11], the initial branched-chain explosion [3, 14] produced after the heat source is turned on could be computed with the temperature and velocity fields given above. It can be anticipated that, since ignition often requires temperatures that are much larger than the normal ambient value, the front solutions described above will be particularly useful for these ignition studies.

Appendix A. The front solution for $\sigma = 1$. The structure of the front when $\sigma = 1$ is different from that described in the text for σ in the range $0 \le \sigma < 1$. We give first the solution corresponding to the line source of heat, and describe later the small modifications required for the planar and spherical geometries.

To construct the solution one needs to match the leading-order solution across the corner layer with the the two-term expansion $\theta = \theta_0 + q^{-1/2}\theta_1$, where we already anticipate that the order of the correction is $\mu_1 = 1/2$. The first-order correction θ_1 is determined by integrating

$$\left(1 - \frac{\xi^2}{2}\right)\theta_{1\xi} = 2\theta_0(\xi\theta_{0\xi})_\xi\theta_1 + \theta_0^2(\xi\theta_{1\xi})_\xi, \tag{A.1}$$

with boundary conditions

$$\begin{cases} \xi = 0: & (\theta_0\theta_1)_\xi = 0, \\ \xi = \sqrt{2}: & \theta_1 = 2D\ln[1/(\sqrt{2}-\xi)], \end{cases} \tag{A.2}$$

where D is an unknown constant to be determined as part of the matching procedure. Near the front, the two-term expansion $\theta = \theta_0 + q^{-1/2}\theta_1$ can be written as

$$\theta = \sqrt{2}(\sqrt{2}-\xi)\left[\ln\left(\frac{1}{\sqrt{2}-\xi}\right)\right]^{1/2} + q^{-1/2}2D\ln\left(\frac{1}{\sqrt{2}-\xi}\right), \tag{A.3}$$

where use has been made of (3.9).

Observation of (A.3) reveals that the corner layer, where the temperature becomes of order unity, is a factor $(\ln q)^{-1/2}$ thinner than that found with $\sigma < 1$, and is displaced towards the cold outer gas. More precisely, the front extends over distances of order $q^{-1/2}(\ln q)^{-1/2}$ around $\xi = \sqrt{2} + Dq^{-1/2}(\ln q)^{1/2}$, where D is the unknown constant appearing in (A.2). The appropriate inner coordinate must incorporate both a translation and a dilatation according to $\zeta = q^{1/2}(\ln q)^{1/2}[\sqrt{2} + Dq^{-1/2}(\ln q)^{1/2} - \xi]$. The problem reduces to that of integrating

$$T^2T_{\zeta\zeta} - DT_\zeta = 0 \begin{cases} \zeta \to -\infty: & T - 1 \to 0, \\ \zeta \to \infty: & T_\zeta - 1 \to 0, \end{cases} \tag{A.4}$$

where the boundary condition as $\zeta \to \infty$ comes from matching with (A.3). Integrating once with use made of the boundary condition $T(-\infty) = 1$ yields $T_\zeta = D(1 - 1/T)$, whereas imposing the linear profile on the hot boundary finally determines $D = 1$. This value can be used in (A.2) to complete the boundary conditions necessary to uniquely determine the first-order perturbation θ_1. Note that the second quadrature for the corner-layer equation, $T + \ln(T-1) = \zeta + \zeta_o$, contains an arbitrary translation ζ_o, which could be computed from higher-order terms in the asymptotic expansion.

The solutions encountered for $j = 0$ and $j = 2$ also respond to the same structure. Thus, for the planar heat source, the first-order perturbation θ_1 in the expansion $\theta = \theta_0 + t^{-1/2}\theta_1$ is determined from

$$(1-x)\theta_{1x} = 2\theta_0\theta_{0xx}\theta_1 + \theta_0^2\theta_{1xx} \begin{cases} x = 0: & (\theta_0\theta_1)_x = 0, \\ x = 1: & \theta_1 = 2D\ln[1/(1-x)], \end{cases} \tag{A.5}$$

while for the point source the expansion in the hot region becomes $\theta = \theta_0 + t^{1/6}\theta_1$, where θ_1 satisfies

$$\left(1 - \frac{y^3}{3}\right)\theta_{1y} = 2\theta_0(y^2\theta_{0y})_y\theta_1 + \theta_0^2(y^2\theta_{1y})_y, \tag{A.6}$$

with boundary conditions

$$\begin{cases} y = 0: & (\theta_0\theta_1)_y = 0, \\ y = 3^{1/3}: & \theta_1 = 6D\ln[1/(3^{1/3}-y)]. \end{cases} \tag{A.7}$$

On the other hand, the inner coordinates $\zeta = (t/2)^{1/2}(\ln t)^{1/2}[1+D(t/2)^{-1/2}(\ln t)^{1/2} - x]$ for $j = 0$, and $\zeta = [(7/2)\ln(1/t)]^{1/2}t^{-1/6}\left\{3^{1/3} + [(7/2)\ln(1/t)]^{1/2}t^{1/6}(D/3) - y\right\}$ for $j = 2$, reduce the corner-layer problem to (A.4), indicating that $D = 1$ should be used in (A.5) and (A.7).

Appendix B. The front solution for $\sigma > 1$. The structure of the thermal wave near the edge for $\sigma > 1$ is similar to that described above for $\sigma = 1$. As explained in the text, the asymptotic behavior of the leading-order profile θ_0 near the edge is

$$\theta_0 = E[(j+1)^{1/(j+1)} - \xi]^{1/\sigma}, \tag{B.1}$$

where E is a constant to be determined from the numerical integration, and where ξ should be replaced with x and y for $j = 0$ and $j = 2$, respectively, following the notation used in the text.

As before, we shall first give the solution corresponding to $j = 1$, for which we assume the expansion $\theta = \theta_0 + q^{-1/(\sigma+1)}\theta_1$. The first-order correction θ_1 can be calculated by integrating

$$\left(1 - \frac{\xi^2}{2}\right)\theta_{1\xi} = 2\theta_0(\xi\theta_0^{\sigma-1}\theta_{0\xi})_\xi\theta_1 + \theta_0^2(\xi\theta_0^{\sigma-1}\theta_{1\xi})_\xi + (\sigma-1)\theta_0^2(\xi\theta_0^{\sigma-2}\theta_{0\xi}\theta_1)_\xi, \tag{B.2}$$

with boundary conditions

$$\begin{cases} \xi = 0 : \theta_0\theta_{1\xi} + \sigma\theta_1\theta_{0\xi} = 0, \\ \xi = \sqrt{2} : \theta_1 = D(\sqrt{2} - \xi)^{(1-\sigma)/\sigma}, \end{cases} \tag{B.3}$$

where D is a constant to be determined below. Near the edge, the two-term expansion for θ gives

$$\theta = E(\sqrt{2} - \xi)^{1/\sigma} + D(\sqrt{2} - \xi)^{(1-\sigma)/\sigma}. \tag{B.4}$$

As seen before for $\sigma = 1$, the corner layer that appears is thinner than that corresponding to $\sigma < 1$ and is displaced towards the outer cold gas. Its inner structure can be described by introducing the variable $\zeta = (\sigma/E)q^{\sigma/(\sigma+1)}[\sqrt{2} + (\sigma/E)Dq^{-1/(\sigma+1)} - \xi]$ to yield the problem

$$T^2(T^{\sigma-1}T_\zeta)_\zeta - DT_\zeta = 0 \begin{cases} \zeta \to -\infty : T \to 1, \\ \zeta \to \infty : T_\zeta \to 1/\sigma\zeta^{(1-\sigma)/\sigma}. \end{cases} \tag{B.5}$$

The boundary condition as $\zeta \to -\infty$ can be used in a first quadrature to give us $D[(T-1)/T^\sigma] = T_\zeta$, which can be evaluated as $\zeta \to \infty$ to provide $D = 1/\sigma$ for the value of the unknown constant D.

The same structure appears near the edge of the thermal wave when $j = 0$ and $j = 2$. For the planar case, the first-order correction in the expansion $\theta = \theta_0 + t^{-1/(\sigma+1)}\theta_1$ is determined from

$$(1-x)\theta_{1x} = 2\theta_0(\theta_0^{\sigma-1}\theta_{0x})_x\theta_1 + \theta_0^2(\theta_0^{\sigma-1}\theta_{1x})_x + (\sigma-1)\theta_0^2(\theta_0^{\sigma-2}\theta_{0x}\theta_1)_x, \tag{B.6}$$

with boundary conditions

$$\begin{cases} x = 0 : \theta_0\theta_{1x} + \sigma\theta_1\theta_{0x} = 0, \\ x = 1 : \theta_1 = D(1-x)^{(1-\sigma)/\sigma}. \end{cases} \tag{B.7}$$

Similarly, the expansion for $j = 2$ is $\theta = \theta_0 + t^{1/[3(\sigma+1)]}\theta_1$, where θ_1 is computed by integrating

(B.8)
$$\left(1 - \frac{y^3}{3}\right)\theta_{1y} = 2\theta_0(y^2\theta_0^{\sigma-1}\theta_{0y})_y\theta_1 + \theta_0^2(y^2\theta_0^{\sigma-1}\theta_{1y})_y + (\sigma-1)\theta_0^2(y^2\theta_0^{\sigma-2}\theta_{0y}\theta_1)_y,$$

with boundary conditions

(B.9)
$$\begin{cases} y = 0 : \theta_0\theta_{1y} + \sigma\theta_1\theta_{0y} = 0, \\ y = 3^{1/3} : \theta_1 = 3D(3^{1/3} - y)^{(1-\sigma)/\sigma}. \end{cases}$$

Use of the inner coordinates $\zeta = (\sigma/E)(t/2)^{\sigma/(\sigma+1)}[1 + (\sigma/E)D(t/2)^{-1/(\sigma+1)} - x]$ for $j = 0$, and $\zeta = (\sigma/E)t^{-\sigma/[3(\sigma+1)]}[3^{1/3} + (\sigma/E)(D/3)t^{1/[3(\sigma+1)]} - y]$ for $j = 2$, reduces the description of the corner layer to the problem given in (B.5), so that the value $D = 1/\sigma$ is obtained for the constant D appearing in (B.7) and (B.9).

REFERENCES

[1] M. Abramowitz and A. Stegun, *Handbook of Mathematical Functions*, Dover, New York, 1965.

[2] D. G. Aronson, *Regularity of flows in porous media: A survey*, in Nonlinear Diffusion Equations and Their Equilibrium States, W. M. Ni, L. A. Peletier, and J. Serrin, eds., Springer-Verlag, New York, 1988, pp. 35–49.

[3] L. L. Bonilla, A. L. Sánchez, and M. Carretero, *The description of homogeneous branched-chain explosions with slow radical recombination by self-adjusting time scales*, SIAM J. Appl. Math., 61 (2000), pp. 528–550.

[4] H. S. Carslaw and J. C. Jaeger, *Conduction of Heat in Solids*, 2nd ed., Oxford University Press, London, 1959.

[5] J. J. Clarke, D. R. Kassoy, and N. Riley, *Shocks generated in a confined gas due to rapid heat addition at the boundary.* I. *Weak shock waves*, Proc. Roy. Soc. London A, 393 (1984), pp. 309–329.

[6] J. J. Clarke, D. R. Kassoy, and N. Riley, *Shocks generated in a confined gas due to rapid heat addition at the boundary.* II. *Strong shock waves*, Proc. Roy. Soc. London A, 393 (1984), pp. 331–351.

[7] J. Crank, *The Mathematics of Diffusion*, 2nd ed., Clarendon Press, Oxford, UK, 1975.

[8] V. N. Kurdyumov and A. Liñán, *Free convection from a point source of heat, and heat transfer from spheres at small Grashof numbers*, Int. J. Heat Mass Transfer, 42 (1999), pp. 3849–3860.

[9] A. Liñán and V. N. Kurdyumov, *Laminar free convection induced by a line heat source, and heat transfer from wires at small Grashof numbers*, J. Fluid Mech., 362 (1998), pp. 199–227.

[10] U. Maas and J. Warnatz, *Ignition processes in hydrogen-air mixtures*, Combustion and Flame, 74 (1988), pp. 53–69.

[11] U. Maas, B. Raffel, J. Wolfrum, and J. Warnatz, *Observation and simulation of laser induced ignition processes in O_2-O_3 and H_2-O_2 mixtures*, Proc. Comb. Institute, 21 (1986), pp. 1869–1876.

[12] B. Meerson, *On the dynamics of strong temperature disturbances in the upper atmosphere of the Earth*, Phys. Fluids A, 1 (1989), pp. 887–891.

[13] A. A. Samarski, V. A. Galaktionov, S. P. Kurdyumov, and A. P. Mikhailov, *Blow-Up in Quasilinear Parabolic Equations*, Walter de Gruyter, New York, 1995.

[14] A. L. Sánchez, A. Liñán, and F. A. Williams, *Chain-branching explosions in mixing layers*, SIAM J. Appl. Math., 59 (1999), pp. 1335–1355.

[15] C. Sánchez-Tarifa, A. Crespo, and E. Fraga, *A theoretical model for the combustion of droplets in supercritical conditions and gas pockets*, Astronautica Acta, 17 (1972), pp. 685–692.

[16] J. Sanz, A. Liñán, M. Rodríguez, and J. R. Sanmartín, *Quasi-steady expansion of plasma ablated from laser-irradiated pellets*, Phys. Fluids, 24 (1981), pp. 2098–2106.

[17] H. Schlichting, *Boundary-Layer Theory*, Springer-Verlag, Berlin, 2000.

[18] L. I. Sedov, *Propagation of strong shock waves*, Prikl. Mat. Mekh., 10 (1946), pp. 241–250.
[19] G. I. Taylor, *The formation of a blast wave by a very intense explosion. Part* I. *Theoretical discussion*, Proc. Roy. Soc. London A, 201 (1950), pp. 159–174.
[20] G. I. Taylor, *The formation of a blast wave by a very intense explosion. Part* II. *The atomic explosion of* 1945, Proc. Roy. Soc. London A, 201 (1950), pp. 175–186.
[21] F. A. Williams, *Combustion Theory*, Benjamin Cummings, Menlo Park, CA, 1985.
[22] Ya. B. Zeldovich and A. S. Kompaneetz, *Towards a theory of heat conduction with thermal conductivity depending on the temperature*, Collection of papers dedicated to the 70th birthday of Academician A. F. Ioffe, Izd. Akad. Nauk SSSR, Moscow, 1950, pp. 61–71.
[23] Ya. B. Zeldovich and Yu. P. Raizer, *Physics of Shock Waves and High Temperature Hydrodynamics Phenomena*, Fizmatgiz, Moscow, 1963; English translation, Academic Press, New York, 1967.

SIAM J. APPL. MATH.
Vol. 63, No. 3, pp. 962–986

GENERALIZED TAYLOR–ARIS DISPERSION IN SPATIALLY PERIODIC MICROFLUIDIC NETWORKS. CHEMICAL REACTIONS*

K. D. DORFMAN† AND H. BRENNER‡

Abstract. Macrotransport theory governing solute transport in spatially periodic networks is extended so as to account for first-order, irreversible chemical reactions occurring within the network. The otherwise locally continuous interstices of the spatially periodic medium are modeled as a discrete graphical network by the expedient of dividing the repetitive unit cell into a finite number of subvolume elements i ($i = 1, 2, \ldots, n$) representing the nodes of the graph. The solute is assumed to be depleted at the uniform rate $k(i)$ when present in node i, i.e., each node i is modeled as a continuous stirred-tank flow reactor. The edges of the graph embody the solute transport processes occurring between nodes, either via "piggy-back" entrainment in a flowing fluid or external force-driven animation, or both, as well as by molecular diffusion. A Taylor–Aris-like "method-of-moments" scheme is applied to homogenize the resulting master equation governing solute transport within the network, thereby explicitly furnishing (i) a pair of adjoint matrix eigenvalue problems for computing the node-based macrotransport fields $P_0^\infty(i)$ and $A(i)$ (ultimately required to calculate the mean solute velocity $\bar{\mathbf{U}}^*$), as well as the network-scale, effective first-order irreversible reaction rate constant $\bar{K}^*$; (ii) a matrix equation for computing the third node-based macrotransport field $\mathbf{B}(i)$ (ultimately used to determine the Taylor–Aris solute dispersivity $\bar{\mathbf{D}}^*$); and (iii) edge-based summations of the three preceding nodal fields, used to calculate the network-scale solute velocity vector $\bar{\mathbf{U}}^*$ and dispersivity dyadic $\bar{\mathbf{D}}^*$. The computational simplicity of this graphical network scheme, in contrast with the original interstitially continuous Taylor–Aris macrotransport paradigm, is demonstrated in the context of an elementary geometric model of a porous medium.

Key words. Taylor dispersion, stochastic processes, asymptotics

AMS subject classifications. 60G15, 82C31, 82C70

PII. S0036139902401872

1. Introduction.

The integration of microscale reaction protocols with downstream microfluidic chromatographic separation techniques has spearheaded the development of miniaturized total analysis systems (μ-TAS) [18, 20, 23] directed towards low-volume (point-of-use) chemical processes and biological assays in microchip environments. Constructing such devices with precision microfabrication techniques enables the creation of highly reproducible periodic microscale structures of any mode of arrangement, whose unit cell configurations can be designed for optimal performance. Concurrently, the relatively new field of "microreaction engineering" [19] has employed these fabrication techniques to produce increasingly complex microscale reactor architectures. Globally interpreting the performance of these devices necessitates knowledge of the three device-scale parameters serving to quantify the effective transport processes, namely, the mean solute depletion rate $\bar{K}^*$, velocity vector $\bar{\mathbf{U}}^*$, and dispersion dyadic $\bar{\mathbf{D}}^*$. Computing these global parameters from knowledge of the detailed microscale (unit cell) parameters characterizing the device necessitates

*Received by the editors February 1, 2002; accepted for publication (in revised form) August 16, 2002; published electronically February 25, 2003. The research of the first author was supported by a Graduate Research Fellowship from the National Science Foundation. The research of the second author was supported by a grant from Eli Lilly & Company to encourage microfluidic research.
http://www.siam.org/journals/siap/63-3/40187.html

†Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139. Current address: Institut Curie, Physico-Chimie/UMR 168, 26 Rue d'Ulm, 75248 Paris Cedex 5, France (Kevin.Dorfman@curie.fr).

‡Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139 (hbrenner@mit.edu).

creating theoretical tools sufficient to do justice to the technological advances implicit therein, while at the same time being sufficiently simple to render the computations tractable. This is the goal of the scheme outlined herein.

A previous contribution [12] demonstrated that graph theoretical models of microfluidic devices could be satisfactorily employed to compute $\bar{\mathbf{U}}^*$ and $\bar{\mathbf{D}}^*$ for nonreactive systems with a degree of rigor consistent with conventional (i.e., interstitially continuous) spatially periodic Taylor–Aris models, the latter models requiring vastly more detailed microscale data and computational resources. Network models of this type prove especially useful in microfluidic contexts due to their ability to straightforwardly incorporate complicated topologies into the requisite analysis. Such graphical network modeling entails subdividing the otherwise continuous solute transport path within the spatially periodic interstitial domain of the device into discrete volume elements. These constitute the nodes of the graph, which are treated simply as "points." Transport between such volume elements occurs within the channels connecting them. These constitute the edges of the graph, regarded simply as "lines" connecting the points, with the transport rates occurring therein quantified by experimentally measurable, albeit averaged, discrete local-scale transport parameters. As such, the requisite internode convective and diffusive transport rates must be specified for each edge of the graph. In addition, a "mixing rule" must be specified to govern the choice of intersectional egress channel for those solute particles instantaneously situated within the channel intersections and about to exit.

In the absence of the present reactive feature, application of a rigorous Taylor–Aris-like "method-of-moments" scheme to the lumped-parameter, local-scale transport processes occurring within the spatially periodic network produced a generic paradigm [12], enabling the calculation of $\bar{\mathbf{U}}^*$ and $\bar{\mathbf{D}}^*$ from knowledge of the prescribed local-scale data. In the present contribution, we extend this nonreactive network scheme to include the depletion of physicochemically reactive solutes within the network, either via chemical reaction or by irreversible adsorption onto the walls of the medium. In the spirit of previous generalized Taylor–Aris analyses [9], we develop a generic computational scheme for extracting the key macroscale parameters, namely, $\bar{K}^*$, $\bar{\mathbf{U}}^*$, and $\bar{\mathbf{D}}^*$, from the prescribed microscale data. Aside from their Lagrangian definitions, the latter trio of parameters also possess Eulerian interpretations as the transport coefficients appearing in the "macrotransport" equation [9],

$$\frac{\partial \bar{P}}{\partial t} + \bar{\mathbf{U}}^* \cdot \bar{\nabla} \bar{P} - \bar{\mathbf{D}}^* : \bar{\nabla} \bar{\nabla} \bar{P} + \bar{K}^* \bar{P} = A\left(i_0\right) \delta\left(\mathbf{R_I} - \mathbf{R_{I_0}}\right) \delta\left(t\right), \tag{1.1}$$

quantifying the asymptotic, long-time transport processes of the reactive solute probability density. (What is meant by long-time will be established in section 4.) In the latter, $\bar{P}$ is a coarse-grained probability density, $\bar{\nabla}$ is a coarse-grained gradient operator, δ is the Dirac delta function, $\mathbf{R_I}$ is the unit cell location (c.f. (2.1)), and i_0 and $\mathbf{R_{I_0}}$, respectively, identify the local and unit cell scale initial locations of the solute pulse. Importantly, the magnitude of the (fictitious) initial condition appearing in the above macrotransport equation is not necessarily equal to that of the true initial condition [7], with the difference quantified by the fictitious initial condition field $A(i)$ (this field will be determined in our asymptotic analysis). Consequently, although the macrotransport coefficients $\bar{K}^*$, $\bar{\mathbf{U}}^*$, and $\bar{\mathbf{D}}^*$ will themselves prove to be independent of the initial condition, at least for long-times validating (1.1), the effective equation still depends upon the initial condition. This memory effect contrasts directly with the Taylor–Aris description of conservative transport processes, wherein the effective equation "forgets" the initial condition. Indeed, the incorporation of the

fictitious initial condition is an essential feature of effective transport equations for nonconservative transport processes [7].

Numerous schemes, displaying varying degrees of rigor and sophistication, have already been proposed in the literature for formulating an effective equation and calculating the coefficients appearing therein. With the existence of sufficient statistical data created from stochastic simulations, approximate values of these global transport rate parameters may be extracted from unidirectional capillary transport models [2, 3, 40], pore-effectiveness factors [17], or other algorithms for simulating particle transport [32]. Alternatively, analytical techniques, such as effective-medium theories [10, 11, 21, 22, 26, 29], multiple-scales analyses [27], volume-averaging [33], center-manifold theory [5, 6], effective stream-tube ensembles [16], and general lumping analyses [25], have been invoked to homogenize the unsteady convection-diffusion-reaction transport equation governing the solute transport through the interstices of the periodic array. These latter techniques are well adapted to characterize disordered ("random") porous media or nonlinear chemical reaction rates (or both), along with the concomitant degree of mathematical and computational complexity accompanying such schemes. Indeed, variations of these schemes have been employed to analyze transport in randomly connected reactive networks [2, 3, 4, 17, 24, 34, 40], in particular near to the percolation limit [3, 34, 39, 40].

In what follows, only first-order, irreversible reactions occurring in the interstices of a regular, spatially periodic "porous medium" are considered. Analogous to our prior analysis of nonreactive media [12], we adapt the spatially periodic moment scheme proposed originally by Dungan, Shapiro, and Brenner [14] to the network model. Such generalized moment analyses constitute physically reasonable methods for homogenizing linear transport equations, while sidestepping the mathematical and computational intractabilities inherent in the aforementioned, more detailed homogenization theories. As a consequence, our analysis, when brought to fruition, furnishes a straightforward matrix equation/edge summation scheme for computing $\bar{K}^*$, $\bar{\mathbf{U}}^*$, and $\bar{\mathbf{D}}^*$. The computational simplicity of the resulting scheme renders parametric studies of the macrotransport processes computationally feasible, even for large networks with complex architectures.

Apart from the explicit μ-TAS and microreaction engineering applications cited above, the generic paradigm to be developed is of broader interest in applications lying outside of these fields. Indeed, various homogenization procedures, albeit devoid of our rigorous Taylor–Aris network formalism, have previously been invoked to study catalysis [3, 17, 31, 33, 34, 39, 40], reduced kinetic models [25], transport in chemical reactors and porous media [2, 6, 14, 15, 16, 17, 27, 28, 30, 34], and irreversible adsorption phenomena [4, 24, 32, 34, 38]. One particularly interesting use of the notion of homogenization involves extracting macroscopically observable reaction rates from molecular-scale models of coupled reaction-diffusion phenomena [10, 11, 21, 22, 26, 29].

This paper is organized as follows. Section 2 outlines the graph construction and concomitant master equation governing the solute (probability density) transport process. Section 3 details an adaptation of the generalized moment scheme [14] to the graphical master equation. Asymptotic long-time moments of the probability density are evaluated in section 4, thereby furnishing a generic paradigm for computing the macrotransport parameters from the prescribed microscale data. This paradigm is applied in section 5 to a relatively straightforward reactive network model, demonstrating thereby the ready applicability of this scheme towards extracting complex, *nonlinear macroscopic* behavior from otherwise nominally *linear microscale* systems.

2. Microscale description. The general protocol for converting an interstitially continuous spatially periodic model into a graphical network model was discussed at length in our prior contribution [12]. Consequently, the exposition which follows is appropriately abbreviated, making adjustments to the prior discretization technique, where necessary, to properly account for the nontrivial feature of (locally) spatially nonuniform chemical reactions. Reference [12] should be consulted for further details.

2.1. Continuous model. Attention is focused exclusively upon convective-diffusive-reactive transport processes occurring in "strongly connected," spatially periodic networks. The spatially periodic medium is characterized by the existence of a repetitive unit cell, extending indefinitely in all directions. The use of infinitely extended networks eliminates the need to explicitly account for "end effects." As real networks are finite in extent, the present analysis is expected to be strictly asymptotically valid only for circumstances where the number, N, of unit cells comprising the real system is large, i.e., $N \gg 1$. While the present unbounded analysis is mathematically consistent, care must be taken in its direct application to bounded systems. Explicitly, it has been shown [5], in the context of center-manifold theory, that Taylor–Aris dispersion models of this type may not be applicable to finite systems when local-space diffusion is not the shortest time scale. Moreover, sufficiently short residence times in bounded systems may fail to satisfy the requisite long-time criteria (see section 4). In the latter case, the present asymptotic analysis will no longer be valid.

The geometry of the (three-dimensional) unit cell is quantified by a trio of base lattice vectors $\{\mathbf{l}_1, \mathbf{l}_2, \mathbf{l}_3\}$, which are subject to the restriction that the magnitude of their scalar triple product, $|\mathbf{l}_1 \times \mathbf{l}_2 \cdot \mathbf{l}_3|$, is equal to the superficial volume, τ_0, of the unit cell [9]. The location of a given unit cell within the infinite array is identified by a triad of integers $\mathbf{I} \equiv (I_1, I_2, I_3)$ ($I_j = 0, \pm 1, \pm 2, \ldots, \pm\infty$; $j = 1, 2, 3$) whereby the centroid of cell $\mathbf{I}$ is vectorially displaced from an arbitrary origin situated at $\mathbf{R}_0 = (0, 0, 0)$ by an amount represented by the discrete position vector

$$\mathbf{R}_\mathbf{I} = \mathbf{l}_1 I_1 + \mathbf{l}_2 I_2 + \mathbf{l}_3 I_3. \tag{2.1}$$

The interstitial domain of the unit cell is decomposed into a finite number of subvolume elements i ($i = 1, 2, \ldots, n$), here represented as the nodes of the graph.[1] Consequently, the instantaneous location of a solute particle on the graph is denoted by the discrete matrix/integer pair $(\mathbf{I}, i)$. Each node i is characterized by its volume, $v(i)$, and its reaction rate constant, $k(i)$ ($k \geq 0$), the latter quantifying the irreversible, first-order rate of solute depletion (if any) occurring therein. Consequently, the periodic network may be envisioned as composed of a strongly connected network of homogeneous, continuous stirred-tank flow reactors (CSTFRs).

Transport between contiguous subvolume elements, say, nodes i' and i, is represented by the edges of the graph. The edge geometry requires specifying the edge length, $l(j)$, corresponding to the distance between the respective centroids of the subvolume elements i' and i (connected by edge j), and the edge's effective channel cross-sectional area, $A(j)$. In addition to molecular diffusion, solute motion within the

[1] In considering purely convective-diffusive transport [12], each subvolume element was defined (for the sake of definiteness) to consist of the volume of a channel intersection plus half the volume of those channels incident to that intersection. This restriction is relaxed in what follows, since we will require later that the reaction rate be uniform within a given subvolume element.

edges of the network is assumed to arise from passive "piggy-back" entrainment in a flowing solvent and/or by the action of an externally applied force acting on the solute molecules. As in our earlier graphical network model [12], these transport mechanisms here are quantified by the mean solute speed, $U(j)$, and dispersivity, $D(j)$, prevailing in edge j. Both are regarded as being scalars within that edge, their tensorial attributes being associated with the spatial direction (orientation) of the channel centerline equipollent to that edge. The edge dispersivity $D(j)$ includes contributions from both molecular diffusion and Taylor–Aris dispersion, the latter arising from any local flow inhomogeneities existing within the channel. For a channel of sufficiently large aspect ratio, the microscale parameters $U(j)$ and $D(j)$ are calculable, at least in principle, from classical macrotransport theory [9]. In lieu of sufficient hydrodynamic data for effecting their calculation, these parameters may also be measured experimentally. Further details regarding their evaluation in such circumstances are available elsewhere [12].

Since multiple edges j are typically associated with a single node i,[2] the preference for the solute to choose a particular edge j upon exiting node i is assumed to be governed quantitatively by a mixing parameter $K(j)$. The numerous models proposed for estimating this parameter are reviewed in our prior, nonreactive contribution [12]. The hypotheses underlying the three most widely prevalent models are (i) perfect mixing ($K(j) = 1$), where the channel intersections are simply envisioned as large mixing volumes; (ii) flow-rate proportionality ($K(j) \propto U(j)A(j)$), where, for convection-dominated flows, it is assumed that the intersection residence time is insufficient for the particle to cross many streamlines within the intersection before exiting; and (iii) thermodynamic partitioning ($K(j) =$ a function of the solute physicochemical properties), where, for diffusion-dominated flows, the longer intersection residence time suffices to establish thermodynamic equilibrium. With use of the preceding geometrical data and transport parameters, the edge convection rate, $c(j)$, and edge diffusion (dispersion) rate, $d(j)$, are defined as the respective volumetric flow rates,

$$c(j) \overset{\text{def.}}{=} K(j)U(j)A(j), \qquad d(j) \overset{\text{def.}}{=} K(j)\frac{D(j)A(j)}{l(j)}. \tag{2.2}$$

2.2. Graphical model. In order to clarify the preceding discussion, as well as to facilitate construction of the requisite graphs, consider by way of example the reactive medium depicted in Figure 1. The network repeats indefinitely in the x-direction, whereby its unidirectional lattice geometry is characterized by the single base lattice vector $\mathbf{l}_x = \hat{\mathbf{x}} l_x$, with $\hat{\mathbf{x}}$ a unit vector in the x-direction and $l_x = |\mathbf{l}_x|$ the period of the unit cell. The single unit cell, indicated by the box, is shaded to correspond to its graphical decomposition into the trio of volume elements, $i = \{a, b, c\}$.

Figure 2(a) depicts the basic graph [12], Γ_b, constructed from the network in Figure 1. The basic graph consists of those nodes contained within a representative cell, say, $\mathbf{I}$, as well as those nodes located in an adjacent cell, $\mathbf{I}'$, possessing an edge directed into cell $\mathbf{I}$.[3] The edge direction is chosen such that the convective flow

[2] A node possessing a single incident edge corresponds to a "dead-end" bond in the network model. The strong connectivity of the network requires that at least one node in the unit cell possess multiple incident edges.

[3] The notation $\mathbf{I}'$ is invoked to generically denote a cell adjacent to $\mathbf{I}$. For networks with multiple adjacent cells, possessing a number of edges entering cell $\mathbf{I}$, the respective cells would be referred to notationally as $\mathbf{I}'$, $\mathbf{I}''$, etc.

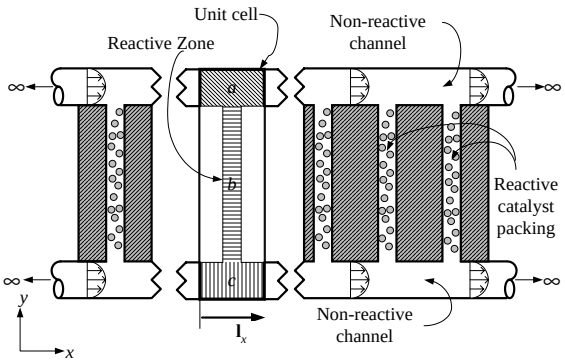


FIG. 1. *Spatially periodic, unidirectional reactive network consisting of two continuous, infinitely extended, nonreactive cylindrical ducts, periodically connected by thin, cylindrical tubes containing a reactive catalyst packing. The periodicity of the network is reflected by the presence of the unit cell, indicated by the highlighted box, with base lattice vector* $\mathbf{l_x}$. *The white portion of the unit cell indicates the inaccessible volume occupied by the blocks separating adjacent reactive domains. The unit cell is subdivided into the three discrete volumetric domains, a, b, and c, so as to facilitate subsequent graphical analysis of the network.*

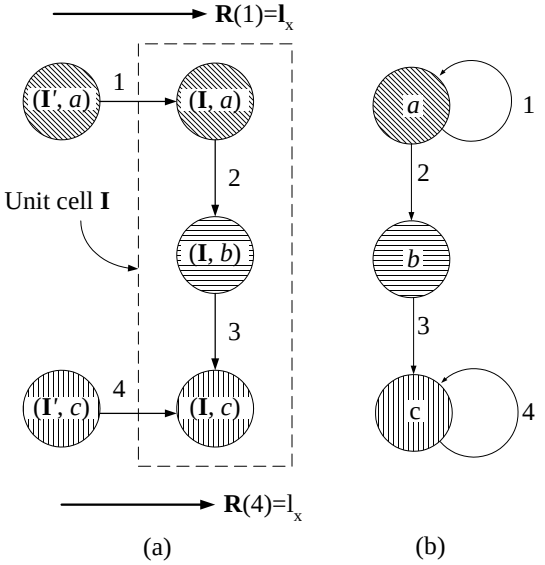


FIG. 2. (a) *Basic graph constructed from the continuous description of Figure* 1. *Vertices* $i = \{a, b, c\}$ *on the basic graph correspond to the volume elements depicted in Figure* 1. *The edges* $j = \{1, 2, 3, 4\}$ *connecting adjacent vertices represent intrachannel transport pathways situated between the individual volume elements* i, *within each edge, in which the solute is transported at the convective rate* $c(j)$ *and diffusive rate* $d(j)$. *The macroscopic jump vector* $\mathbf{R}(j = \{1, 4\}) = \mathbf{l_x}$ *corresponds to a "Darcy-scale" displacement vector drawn between the adjacent cells* $\mathbf{I}'$ *and* $\mathbf{I}$. (b) *Local graph constructed by contracting homologous vertices in the basic graph of Figure* 2(a).

rate, $c(j)$, is positive within that edge, i.e., the edge direction is colinear with the net direction of the flow or the applied force within that edge. (Any edge orientation suffices when $c(j) = 0$.) The basic graph includes all edges that are directed into the unit cell, as well as those edges internal to the cell. Edges entering cell $\mathbf{I}$ from $\mathbf{I}'$ are

assigned the macroscopic jump vector,

$$\mathbf{R}(j) = \mathbf{R}_{\mathbf{I}} - \mathbf{R}_{\mathbf{I}'}, \tag{2.3}$$

corresponding to the discrete "Darcy-scale" vector displacement from cell $\mathbf{I}'$ to cell $\mathbf{I}$. In this case, $\mathbf{R}(1) = \mathbf{R}(4) = \hat{\mathbf{x}} l_x$, whereas $\mathbf{R}(2) = \mathbf{R}(3) = \mathbf{0}$.

Two additional graphs are required to complete the graphical model. The entire infinitely extended periodic medium is captured by the global graph, Γ_g. This graph, which will be used to formulate the master equation, is formed by translations of the basic graph through its basic lattice [1, 12]. The local graph, Γ_l, is constructed by contracting all homologous vertices of the basic graph and removing the edges between them. This graph, depicted in Figure 2(b), is invariant to the choice of unit cell [1]. It will be employed in the moment scheme pursued later.

In order to cast the subsequent Taylor–Aris paradigm into an efficient matrix form, it is necessary to introduce several other graph theoretical entities [8] and transport matrices [12]. Let m denote the number of edges and n the number of nodes on the local graph. The graph connectivity is captured by the $n \times m$ incidence matrix $\mathbf{D}$, whose elements are defined as follows:

$$D_{ij} \overset{\text{def.}}{=} \begin{cases} 1 & \text{if edge } j \text{ has its terminal vertex in node } i, \\ -1 & \text{if edge } j \text{ has its initial vertex in node } i, \\ 0 & \text{otherwise.} \end{cases} \tag{2.4}$$

This matrix can be decomposed into a pair of $n \times m$ matrices,

$$\mathbf{D} = \Pi^{(+)} - \Pi^{(-)}, \tag{2.5}$$

where the nonzero entries of $\Pi^{(+)}$ are the positive entries in $\mathbf{D}$, and the nonzero entries of $\Pi^{(-)}$ constitute the absolute values of the negative entries in $\mathbf{D}$.

Collect the edge transport parameters into the pair of $m \times m$ diagonal matrices,

$$\mathbf{c} = c(j)\delta(i,j), \qquad \mathbf{d} = d(j)\delta(i,j), \tag{2.6}$$

and represent the respective nodal volumes and reaction rates by the pair of $n \times n$ diagonal matrices,

$$\mathbf{v} = v(i)\delta(i,j), \qquad \mathbf{k} = k(i)\delta(i,j). \tag{2.7}$$

Finally, define $\mathbf{R}$ as the $m \times 3$ matrix whose m rows are composed of the macroscopic jump vectors $\mathbf{R}(j)$.

2.3. Probability density on the graph. Consider the conditional reactive-probability density, $P_r(\mathbf{I},i,t \mid i_0) \geq 0$, that the solute "molecule" (particle) being tracked is instantaneously present in cell $\mathbf{I}$ and situated at vertex i at time t, given that the particle was initially introduced into cell $\mathbf{I}_0 = \mathbf{0}$ and vertex i_0 at time $t = 0$. This probability density necessarily obeys the normalized "conservation" equation

$$\sum_{i \in \Gamma g} P_r(\mathbf{I},i,t \mid i_0) = \begin{cases} 0, & t < 0, \\ 1, & t = 0, \\ < 1, & t > 0. \end{cases} \tag{2.8}$$

The last inequality arises from the attenuation of the total amount of solute present in the system at time $t > 0$ caused by its disappearance via chemical reaction or

irreversible adsorption. Indeed, after sufficient time has elapsed, the amount of solute remaining in the system, and hence its probability density, would be expected to be completely depleted (corresponding to $P_r(\mathbf{I},i,t\,|\,i_0)=0$ for all $(\mathbf{I},i)$), a fact which will be subsequently confirmed.

For the case of a reactive solute traversing the network, the reactive probability density is governed by the following convection-diffusion-reaction master equation at each node i on the global graph Γ_g:

$$\begin{aligned} v(i)\frac{dP_r(\mathbf{I},i,t\,|\,i_0)}{dt} &= \delta(\mathbf{I})\delta(i,i_0)\delta(t) - k(i)v(i)P_r(\mathbf{I},i,t\,|\,i_0) \\ &\quad + \sum_{\substack{j\in\Omega^+(i)\\ j=\{i',i\}}} c(j)P_r(\mathbf{I}',i',t\,|\,i_0) + d(j)\begin{bmatrix} P_r(\mathbf{I}',i',t\,|\,i_0) \\ -P_r(\mathbf{I},i,t\,|\,i_0)\end{bmatrix} \\ &\quad - \sum_{\substack{j\in\Omega^-\\ j=\{i,i'\}}} c(j)P_r(\mathbf{I},i,t\,|\,i_0) + d(j)\begin{bmatrix} P_r(\mathbf{I},i,t\,|\,i_0) \\ -P_r(\mathbf{I}',i',t\,|\,i_0)\end{bmatrix}, \end{aligned} \tag{2.9}$$

with $\delta(\mathbf{I})$ and $\delta(i,i_0)$ Kronecker delta functions, $\delta(t)$ the Dirac delta function, and with $j=\{a,b\}$ denoting an edge whose initial vertex is a and whose terminal vertex is b. Proper interpretations of all but one of the terms appearing in (2.9) are as discussed in [12]. The new term, $k(i)v(i)P_r(\mathbf{I},i,t\,|\,i_0)$, accounts for the CSTFR model of solute depletion, with P_r identified with the volumetric solute concentration (i.e., solute mass per unit volume).

In order to assure that subsequent infinite sums converge [cf. (3.3), (3.11)–(3.12)], all moments of the reactive-probability density are required to decay faster than algebraically, i.e.,

$$|\mathbf{R}|^m P_r(\mathbf{I},i,t\,|\,i_0)\to 0 \quad \text{as} \quad |\mathbf{I}|\to\infty. \tag{2.10}$$

3. Moment scheme.

3.1. Local moments. In order to ultimately arrive at the desired paradigm for computing the macrotransport parameters $\bar{K}^*$, $\bar{\mathbf{U}}^*$, and $\bar{\mathbf{D}}^*$, the generalized moment scheme proposed by Dungan, Shapiro, and Brenner [14] will be adapted to the master equation (2.9) governing solute transport on the graph. In this context, define the "*nonreactive*" solute probability density,

$$P(\mathbf{I},i,t\,|\,i_0) \overset{\text{def.}}{=} \frac{\exp\left(\bar{K}t\right)}{A(i_0)} P_r(\mathbf{I},i,t\,|\,i_0), \tag{3.1}$$

where the time- and position-independent reaction velocity constant $\bar{K}$ (defined globally on the network scale) and vertex field $A(i)$ (defined locally on the unit cell scale) will be determined later. Rescaling the probability density in accord with (3.1) will shortly result in a new transport problem (3.2) which only accounts for those species still present in the system at time t. Consequently, the rescaled problem is not adversely affected by very fast reaction rates which serve to deplete much of the solute mass before satisfying the long-time criteria. Rather, the impact of the fast reaction on the total amount of solute still present in the system at time t will be captured by the (reaction rate) eigenvalues $\bar{K}$, whereas the spatial dependence of the reaction will be reflected in the field $A(i)$, whose physical significance will be discussed in the context of (3.12).

The master equation governing P is derived by substituting (3.1) into (2.9) to obtain

$$\begin{aligned} v(i)\frac{dP(\mathbf{I},i,t\,|\,i_0)}{dt} =& \frac{\delta(\mathbf{I})\delta(i,i_0)\delta(t)}{A(i_0)} + \left[\bar{K}-k(i)\right] v(i)P\left(\mathbf{I},i,t\,|\,i_0\right) \\ &+ \sum_{\substack{j\in\Omega^+(i)\\ j=\{i',i\}}} c(j)P(\mathbf{I}',i',t\,|\,i_0) + d(j)\left[\begin{array}{c} P(\mathbf{I}',i',t\,|\,i_0) \\ -P(\mathbf{I},i,t\,|\,i_0) \end{array}\right] \\ &- \sum_{\substack{j\in\Omega^-\\ j=\{i,i'\}}} c(j)P(\mathbf{I},i,t\,|\,i_0) + d(j)\left[\begin{array}{c} P(\mathbf{I},i,t\,|\,i_0) \\ -P(\mathbf{I}',i',t\,|\,i_0) \end{array}\right]. \end{aligned} \tag{3.2}$$

Define the nonreactive local moment as the m-adic,

$$\mathbf{P}_m(i,t\,|\,i_0) \overset{\text{def.}}{=} \sum_{\mathbf{I}} \mathbf{R}_{\mathbf{I}}^m P\left(\mathbf{I},i,t\,|\,i_0\right), \tag{3.3}$$

where $\mathbf{R}_{\mathbf{I}}^m \equiv \mathbf{R}_{\mathbf{I}}\mathbf{R}_{\mathbf{I}}\cdots\mathbf{R}_{\mathbf{I}}$ (m-times); the triple sum,

$$\sum_{\mathbf{I}} \overset{\text{def.}}{=} \sum_{I_1=-\infty}^{\infty}\sum_{I_2=-\infty}^{\infty}\sum_{I_3=-\infty}^{\infty}, \tag{3.4}$$

is taken over all unit cells. The equation governing the local moments $\mathbf{P}_m(i)$ (with time and the initial condition suppressed therein for notational simplicity) is obtained upon multiplying (3.2) by $\mathbf{R}_{\mathbf{I}}^m$ and summing over $\mathbf{I}$, thereby obtaining

$$v(i)\frac{d\mathbf{P}_m(i)}{dt} = \frac{\delta(m,0)\delta(i,i_0)\delta(t)}{A(i_0)} + L\left[\mathbf{P}_m\left(i\right)\right] + \Gamma_m(i), \tag{3.5}$$

where $\delta\left(m,0\right)$ is a Kronecker delta function. In the latter, the vertex operator L operating on an arbitrary field $\psi(i)$ is defined as

$$\begin{aligned} L\left[\psi(i)\right] \overset{\text{def.}}{=}& \sum_{\substack{j\in\Omega^+(i)\\ j=\{i',i\}}} c(j)\psi\left(i'\right) + d(j)\left[\psi(i')-\psi(i)\right] \\ &- \sum_{\substack{j\in\Omega^-\\ j=\{i,i'\}}} c(j)\psi(i) + d(j)\left[\psi(i)-\psi(i')\right] + \left[\bar{K}-k(i)\right] v(i)\psi(i). \end{aligned} \tag{3.6}$$

The first few m-adics $\Gamma_m(i)$ appearing in (3.5) possess the respective forms

$$\Gamma_0(i) = 0, \tag{3.7}$$

$$\Gamma_1(i) = \sum_{\substack{j\in\Omega^+(i)\\ j=\{i',i\}}} \left[c(j)+d(j)\right]\mathbf{R}\left(j\right)P_0(i') - \sum_{\substack{j\in\Omega^-\\ j=\{i,i'\}}} d(j)\mathbf{R}(j)P_0(i'), \tag{3.8}$$

$$\Gamma_2(i) = 2\ \text{sym}\left\{\begin{array}{l} \displaystyle\sum_{\substack{j\in\Omega^+(i)\\ j=\{i',i\}}} \left[c(j)+d(j)\right]\left[\frac{1}{2}\mathbf{R}\left(j\right)\mathbf{R}\left(j\right)P_0(i') + \mathbf{R}\left(j\right)\mathbf{P}_1(i')\right] \\ \quad + \displaystyle\sum_{\substack{j\in\Omega^-\\ j=\{i,i'\}}} d(j)\left[\frac{1}{2}\mathbf{R}\left(j\right)\mathbf{R}\left(j\right)P_0(i') - \mathbf{R}\left(j\right)\mathbf{P}_1(i')\right] \end{array}\right\}, \tag{3.9}$$

where, for an arbitrary dyad $\mathbf{XY}$ (with $\mathbf{X}$ and $\mathbf{Y}$ vectors),

$$\operatorname{sym}(\mathbf{XY}) \stackrel{\text{def.}}{=} \frac{1}{2}(\mathbf{XY} + \mathbf{YX}). \tag{3.10}$$

3.2. Total moments. Define the respective unweighted and weighted *nonreactive* total moments,

$$\mathbf{M}'_m(t \mid i_0) \stackrel{\text{def.}}{=} \sum_{i \in \Gamma_l} v(i)\mathbf{P}_m(i, t \mid i_0), \tag{3.11}$$

$$\mathbf{M}_m(t \mid i_0) \stackrel{\text{def.}}{=} \sum_{i \in \Gamma_l} v(i)A(i)\mathbf{P}_m(i, t \mid i_0). \tag{3.12}$$

In the latter, the node-based field $A(i)$ [cf. (3.1)] arises from the necessity for introducing into macrotransport theory a fictitious initial condition [14], whose significance and defining equation will be established shortly. It is possible to choose the constant $\bar{K}$ appearing in the definition (3.1) such that M'_0 is conserved for sufficiently long-times [36] (see (4.7)). However, transients arising from the initial placement, i_0, of the particle (within cell $\mathbf{I}_0 = \mathbf{0}$) persist for long times, longer than the time required for the asymptotic theory of Taylor–Aris to constitute an accurate global representation of the transport phenomena. By way of example, consider the classical case of solute entrainment in a flow between parallel, reactive plates [9]. If the initial solute pulse is near the plates, then much of the solute mass only samples the slow moving streamlines near the wall before being depleted. In contrast, if the pulse is placed in the midplane, then the solute mass will have the opportunity to sample many streamlines before being depleted. This residual transient thereby impacts nontrivially upon the network-scale transport processes [7] and is captured by the fictitious initial condition field $A(i)$. To properly correct for such transients, as was done in the original derivation [14], we will derive a difference equation for the fictitious initial condition $A(i)$ (in place of the literal initial condition). This scheme insures that M_0 too is conserved for all times, thereby allowing a conventional Taylor–Aris moment analysis [14], involving the use of weighted moments. It will be shown that the rates of change of the weighted and unweighted moments differ only by exponentially small temporal terms, at least for sufficiently long times. As a consequence, the distinction between the two types of total moments, $\mathbf{M}'_m$ and $\mathbf{M}_m$, defined above proves irrelevant in the final macrotransport results.

The differential equation governing M_0 is derived by forming the product of $A(i)$ and (3.5) (with $m = 0$ and (3.7)), and summing over $i \in \Gamma_l$ to obtain

$$\begin{aligned}\frac{dM_0}{dt} = {} & \delta(t) + \sum_{i \in \Gamma_l} \left[\bar{K} - k(i)\right] v(i)A(i)P_0(i) \\ & + \sum_{\substack{j \in E\Gamma_l \\ j \in \Omega^+}} c(j)A(i)P_0(i') + d(j)A(i)\left[P_0(i') - P_0(i)\right] \\ & - \sum_{\substack{j \in E\Gamma_l \\ j \in \Omega^-}} c(j)A(i)P_0(i) + d(j)A(i)\left[P_0(i) - P_0(i')\right].\end{aligned} \tag{3.13}$$

Here and hereafter, the following compact summation notation will be employed:

$$\sum_{\substack{j\in E\Gamma_l\\ j\in\Omega^+}} \overset{\text{def.}}{=} \sum_{i\in V\Gamma_l} \sum_{\substack{j\in\Omega^+(i)\\ j=\{i',i\}}}, \qquad \sum_{\substack{j\in E\Gamma_l\\ j\in\Omega^-}} \overset{\text{def.}}{=} \sum_{i\in V\Gamma_l} \sum_{\substack{j\in\Omega^-(i)\\ j=\{i,i'\}}}. \tag{3.14}$$

The strong connectivity of the graph furnishes the pair of identities,

$$\sum_{\substack{j\in E\Gamma_l\\ j\in\Omega^+}} \epsilon(j)\phi(i') = \sum_{\substack{j\in E\Gamma_l\\ j\in\Omega^-}} \epsilon(j)\phi(i), \tag{3.15}$$

$$\sum_{\substack{j\in E\Gamma_l\\ j\in\Omega^+}} \epsilon(j)\phi_1(i')\phi_2(i) = \sum_{\substack{j\in E\Gamma_l\\ j\in\Omega^-}} \epsilon(j)\phi_1(i)\phi_2(i'), \tag{3.16}$$

where $\phi_k(i)$ and $\epsilon(j)$ are, respectively, node- and edge-based quantities. With use of these identities, (3.13) may be reformulated as

$$\begin{aligned}\frac{dM_0}{dt} &= \delta(t) + \sum_{\substack{j\in E\Gamma_l\\ j\in\Omega^+}} P_0(i)d(j)\left[A(i') - A(i)\right] + \sum_{i\in\Gamma_l}\left[\bar{K} - k(i)\right]v(i)A(i)P_0(i)\\ &\quad + \sum_{\substack{j\in E\Gamma_l\\ j\in\Omega^-}} P_0(i)\left[c(j) + d(j)\right]\left[A(i') - A(i)\right].\end{aligned} \tag{3.17}$$

In order that M_0 be conserved for all times, the summations appearing on the right-hand side of (3.17) must vanish; explicitly,

$$\begin{aligned}&\sum_{\substack{j\in\Omega^+(i)\\ j=\{i',i\}}} d(j)\left[A(i') - A(i)\right] + \sum_{\substack{j\in\Omega^-(i)\\ j=\{i,i'\}}} \left[c(j) + d(j)\right]\left[A(i') - A(i)\right]\\ &\qquad + \left[\bar{K} - k(i)\right]v(i)A(i) = 0.\end{aligned} \tag{3.18}$$

Equation (3.18), governing $A(i)$, may be restated in compact form as

$$\left\{\mathbf{k} - \mathbf{v}^{-1}\cdot\left[\mathbf{D}\cdot\mathbf{d} - \Pi^{(-)}\cdot\mathbf{c}\right]\cdot\mathbf{D}^{\dagger}\right\}\cdot\mathbf{A} = \bar{K}\mathbf{A}, \tag{3.19}$$

where $\mathbf{A}$ is an $n\times 1$ column vector whose elements are the fictitious initial nodal conditions embodied in $A(i)$ $(i = 1, 2, \ldots, n)$. Equation (3.19) constitutes an eigenvalue problem for simultaneously computing the eigenvalues $\bar{K}$ and eigenvectors $A(i)$. The scheme for identifying the one, physically relevant eigenvalue $\bar{K}$, as well as the required normalization of the corresponding physically relevant eigenvector $\mathbf{A}$, will be specified in the following section.

With use of (3.18), temporal integration of (3.17) demonstrates that

$$M_0 = \begin{cases} 0, & t < 0,\\ 1, & t \geq 0,\end{cases} \tag{3.20}$$

whereupon M_0 is indeed seen to be conserved for all times (independently of i_0). A generic equation governing the weighted total moments may also be derived with use of (3.18). To do so, multiply (3.5) by $A(i)$, sum over $i\in\Gamma_l$, and use (3.15) and (3.16), thereby obtaining the expression

$$\frac{d\mathbf{M}_m}{dt} = \delta(m,0)\delta(t) + \sum_{i\in\Gamma_l} A(i)\Gamma_m(i). \tag{3.21}$$

4. Asymptotic, long-time limits. The following section furnishes the asymptotic, long-time limits of the first few local and total moments of the nonreactive probability density. By "long-time" is meant that the residence time, t_R, of the solute in the network is long compared with the diffusion time scale; that is, $t_R \gg l^2/D_m$, where l denotes a characteristic linear dimension of the unit cell (typically the magnitude of a macroscopic jump vector, $|\mathbf{R}(j)|$) and D_m is the molecular diffusivity of the solute [9]. A further criterion imposed upon the definition of long-time behavior will be established later [cf. (4.4)].

4.1. Zero-order moments. For sufficiently long times, the zeroth-order local moment (3.3) assumes the asymptotic form

$$P_0(i,t\,|\,i_0) \approx P_0^{\infty}(i) + \exp \tag{4.1}$$

for all i_0, where "exp" denotes temporal terms that are exponentially small for sufficiently long times. The asymptotic probability density, $P_0^{\infty}(i)$, is independent of time as well as of the initial local position, i_0. The validity of (4.1) has been established [35, 37] via the use of eigenfunction expansions.

Substitute (4.1) into (3.5), set $m = 0$, and use (3.7) to obtain the difference equation governing $P_0^{\infty}(i)$, namely,

$$L\,[P_0^{\infty}(i)] = 0. \tag{4.2}$$

The latter may be recast into the compact matrix form,

$$\left\{\mathbf{k} - \mathbf{v}^{-1}\cdot\mathbf{D}\cdot\left[(\mathbf{c}+\mathbf{d})\cdot\left(\Pi^{(-)}\right)^{\dagger} - \mathbf{d}\cdot\left(\Pi^{(+)}\right)^{\dagger}\right]\right\}\cdot\mathbf{P} = \bar{K}\mathbf{P}, \tag{4.3}$$

where $\mathbf{P}$ is the $n\times 1$ column vector composed of the asymptotic probability densities, $P_0^{\infty}(i)$ $(i = 1, 2, \ldots, n)$. Similar to (3.19), (4.3) constitutes an eigenvalue problem posed for $P_0^{\infty}(i)$ and $\bar{K}$. The eigenvalue with the smallest real part (corresponding to the slowest decaying mode of the full solution) is identified as the effective reaction rate $\bar{K}^*$ [14, 35, 37].[4] For all physical circumstances, the eigenvalue possessing the smallest real part is pure real [35, 37]. Moreover, the solution of this eigenvalue problem furnishes a second criterion quantifying what is meant by the phrase "long-time behavior." Upon denoting the second smallest eigenvalue of (4.3) as $\bar{K}_1$, we require the residence time to satisfy the inequality

$$t_R \gg \left(\bar{K}^*\right)^{-1} - \bar{K}_1^{-1}, \tag{4.4}$$

whereupon the effective transport process is dominated by the eigenvalue with the smallest real part, $\bar{K}^*$.

As shown below, the eigenvalue problems posed for $P_0^{\infty}(i)$ (4.2) and $A(i)$ (3.18) are adjoint. Thus, let $\bar{K}_P$ and $\bar{K}_A$, respectively, be eigenvalues of (4.2) and (3.18). Upon multiplying (3.18) by $P_0^{\infty}(i)$, (4.2) by $A(i)$, and summing both results over $i \in \Gamma_l$, we see that $\bar{K}_P = \bar{K}_A$. Consequently, the appropriate fictitious initial condition $A(i)$ is the eigenvector of (3.18) corresponding to the eigenvalue $\bar{K} = \bar{K}^*$.

[4] By way of example, Batycky, Edwards, and Brenner [7] illustrate the dominance of the slowest decaying mode by comparing the (asymptotic) macrotransport solution for a nonadiabatic unsteady heat transfer process with its exact trigonometric function expansion. Moreover, their analysis clearly illustrates the necessity for incorporating the notion of a fictitious initial condition into effective-medium models, such as in the present macrotransport model.

The eigenvalue problems governing $P_0^\infty(i)$ and $A(i)$ only specify each of these two fields to within arbitrary, constant multipliers. These multipliers may be uniquely determined by applying the normalization conditions [14], namely,

$$\sum_{i\in\Gamma_l} v(i)P_0^\infty(i) = 1, \tag{4.5}$$

$$\sum_{i\in\Gamma_l} v(i)A(i)P_0^\infty(i) = 1. \tag{4.6}$$

To verify that the weighted and unweighted zeroth-order total moments are indistinguishable at long times, substitute the asymptotic solution (4.1) into (3.11), together with the normalization condition (4.5). This demonstrates that our choice of $\bar{K}$ and $A(i_0)$ conserves M_0' for long times, at least to within exponentially small terms; explicitly,

$$M_0' \approx 1 + \exp \tag{4.7}$$

for all i_0. Moreover, the ability to formulate consistent results for M_0, (3.20) and M_0', (4.7) verifies the change in variables (3.1), thereby confirming our prior assertion that the solute is eventually depleted completely at each and every node, and hence throughout the network as a whole.

4.2. First-order moments.

4.2.1. Mean velocity. The mean velocity of the reactive tracer through the network is determined from knowledge of the asymptotic limit of the rate of growth of the first total moment via the generic expression [14],

$$\bar{\mathbf{U}}^* = \lim_{t\to\infty} \frac{d\mathbf{M}_1}{dt}. \tag{4.8}$$

Substitute (3.8) into (3.21), set $m = 1$, and use (3.16) and (4.1) to obtain

$$\bar{\mathbf{U}}^* = \sum_{\substack{j\in E\Gamma_l \\ j\in\Omega^+}} c(j)\mathbf{R}(j)\,A(i)P_0^\infty(i') + d(j)\mathbf{R}(j)\left[A(i)P_0^\infty(i') - A(i')P_0^\infty(i)\right]. \tag{4.9}$$

4.2.2. Derivation of the B-equation. Subject to a posteriori verification, assume the following trial solution for the first-order local moment:

$$\mathbf{P}_1(i) \approx P_0^\infty(i)\left[\bar{\mathbf{U}}^* t + \mathbf{B}(i)\right] + \exp, \tag{4.10}$$

where $\mathbf{B}(i)$ is a node-based field to be determined. Substitution of (4.10) into (3.12), together with the choice $m = 1$, furnishes the weighted first-order total moment,

$$\mathbf{M}_1 \approx \bar{\mathbf{U}}^* t + \sum_{i\in\Gamma_l} v(i)A(i)P_0^\infty(i)\mathbf{B}(i) + \exp. \tag{4.11}$$

The difference equation governing $\mathbf{B}(i)$ is derived by substituting the trial solution (4.10) into (3.5), setting $m = 1$, and using (3.8). Elimination of time-dependent terms via (4.2), and reactive terms via the product of $\mathbf{B}(i)$ with (4.2), eventually yields

$$\begin{aligned}\sum_{\substack{j\in\Omega^+(i) \\ j=\{i',i\}}} &[c(j)+d(j)]\,P_0^\infty(i')\left[\mathbf{B}(i') - \mathbf{B}(i)\right] \\ &- \sum_{\substack{j\in\Omega^-(i) \\ j=\{i,i'\}}} d(j)P_0^\infty(i')\left[\mathbf{B}(i) - \mathbf{B}(i')\right] = v(i)P_0^\infty(i)\bar{\mathbf{U}}^* - \alpha(i),\end{aligned} \tag{4.12}$$

with $\alpha(i)$ the node-based vector,

$$\alpha(i) = \sum_{\substack{j\in\Omega^+(i)\\ j=\{i',i\}}} [c(j)+d(j)]\,\mathbf{R}(j)\,P_0^\infty(i') - \sum_{\substack{j\in\Omega^-(i)\\ j=\{i,i'\}}} d(j)\mathbf{R}(j)P_0^\infty(i'). \tag{4.13}$$

Equation (4.12) defines the **B**-field only to within an arbitrary additive constant vector [9], whose value ultimately proves irrelevant when computing the dispersivity [cf. (4.32)]. Consequently, the resulting degree of freedom may be utilized so as to conveniently allow an arbitrary reference node, say i^*, to be chosen such that $\mathbf{B}(i^*) = \mathbf{0}$. With the latter specification, the $(n-1)$ equations generated by (4.12) for $i \neq i^*$ suffice to determine the remaining vectors $\mathbf{B}(i)$.[5]

Subsequent calculations [cf. (4.32)] necessitate introducing the edge-based vector field,

$$\mathbf{b}(j) \overset{\text{def.}}{=} \mathbf{B}(i) - \mathbf{B}(i'), \quad \{j \in \Omega^+(i)\}, \tag{4.14}$$

defined such that edge j has its initial vertex at i' and its terminal vertex at i. The m vectors, $\mathbf{b}(j)$, may be computed from the solution of the $(n-1)$ equations generated by (4.12), together with the m definitions from (4.14). Alternatively, a difference equation may be derived for $\mathbf{b}(j)$ by substituting (4.14) into (4.12), so as to obtain

$$\sum_{\substack{j\in\Omega^-(i)\\ j=\{i,i'\}}} d(j)P_0^\infty(i')\mathbf{b}(j) - \sum_{\substack{j\in\Omega^+(i)\\ j=\{i',i\}}} [c(j)+d(j)]\,P_0^\infty(i')\mathbf{b}(j) = v(i)P_0^\infty(i)\bar{\mathbf{U}}^* - \alpha(i). \tag{4.15}$$

However, the n equations contained in (4.15) generally prove insufficient to solve for all m vectors $\mathbf{b}(j)$, since, for all but the most trivial networks, $m > n$. Consequently, it is necessary to augment the (nonsquare) coefficient matrix by noting that the sum of the **b** vectors vanishes along any cycle of the graph [12],

$$\sum_{j\in\text{cycle}} \mathbf{b}(j) = \mathbf{0}. \tag{4.16}$$

Superposition of the n equations provided by (4.15), together with the $(m-n)$-independent cycles chosen from (4.16), completely specifies the **b** vectors.

To reformulate the **B**-equations in matrix form, define the pair of $n \times m$ conditioned connectivity matrices,

$$\tilde{\Pi}_{ij}^{(+)} \overset{\text{def.}}{=} \begin{cases} P_0^\infty(i') & \text{if edge } j \text{ is directed from } i' \text{ to } i,\\ 0 & \text{otherwise;}\end{cases} \tag{4.17}$$

$$\tilde{\Pi}_{ij}^{(-)} \overset{\text{def.}}{=} \begin{cases} P_0^\infty(i') & \text{if edge } j \text{ is directed from } i \text{ to } i',\\ 0 & \text{otherwise.}\end{cases} \tag{4.18}$$

With use of the latter, the (nonsquare) equation set (4.15) governing **b** adopts the form

$$\left[\tilde{\Pi}^{(-)}\cdot\mathbf{d} - \tilde{\Pi}^{(+)}\cdot(\mathbf{c}+\mathbf{d})\right]\cdot\mathbf{b} = \mathbf{v}\cdot\mathbf{P}\cdot\bar{\mathbf{U}}^* - \left[\tilde{\Pi}^{(+)}\cdot(\mathbf{c}+\mathbf{d}) - \tilde{\Pi}^{(-)}\cdot\mathbf{d}\right]\cdot\mathbf{R}. \tag{4.19}$$

[5] In the nonreactive network theory [12], the cocycle space was invoked to provide a formal mechanism for choosing the reference node i^*. While this technique remains valid for the present reactive case, subsequent simplifications of the **B**-equations render the utility of such a formalism moot.

Conversion between $\mathbf{b}$ and $\mathbf{B}$ is accomplished via the transformation

$$\mathbf{b} = \mathbf{D}^{\dagger} \cdot \mathbf{B}, \tag{4.20}$$

whereupon (4.12) adopts the matrix form,

$$\left[\tilde{\Pi}^{(-)} \cdot \mathbf{d} - \tilde{\Pi}^{(+)} \cdot (\mathbf{c}+\mathbf{d})\right] \cdot \mathbf{D}^{\dagger} \cdot \mathbf{B} = \mathbf{v} \cdot \mathbf{P} \cdot \bar{\mathbf{U}}^{*} - \left[\tilde{\Pi}^{(+)} \cdot (\mathbf{c}+\mathbf{d}) - \tilde{\Pi}^{(-)} \cdot \mathbf{d}\right] \cdot \mathbf{R}. \tag{4.21}$$

The time-independence of (4.12) and (4.15) confirms a posteriori the assumed trial solution (4.10) for $\mathbf{P}_1$ as well as the resulting expression (4.11) for $\mathbf{M}_1$. Consequently, the unweighted first-order total moment may be computed from (3.11) (with $m = 1$) together with (4.5) and (4.10), yielding

$$\mathbf{M}_1' \approx \bar{\mathbf{U}}^{*} t + \bar{\mathbf{B}} + \text{exp}, \tag{4.22}$$

where the time- and position-independent vector $\bar{\mathbf{B}}$ is of the form

$$\bar{\mathbf{B}} = \sum_{i \in \Gamma_l} v(i) P_0^{\infty}(i) \mathbf{B}(i). \tag{4.23}$$

Differentiation of (4.11) and (4.22) with respect to time reveals that the temporal rates of change of $\mathbf{M}_1$ and $\mathbf{M}_1'$ differ only by exponentially small terms at long times.

4.3. Second-order moments. The difference equation governing the weighted second-order total moment, $\mathbf{M}_2$, is derived from (3.21) with $m = 2$, upon making use of (3.9), (3.16), (4.1), (4.9), and (4.10), thereby obtaining

$$\begin{aligned} \frac{d\mathbf{M}_2}{dt} \approx{}& 2\bar{\mathbf{U}}^{*}\bar{\mathbf{U}}^{*} t + 2 \text{ sym} \left\{ \sum_{\substack{j \in E\Gamma_l \\ j \in \Omega^{+}}} [c(j) + d(j)]\, A(i) P_0^{\infty}(i') \begin{bmatrix} \frac{1}{2}\mathbf{R}(j)\mathbf{R}(j) \\ +\mathbf{R}(j)\mathbf{B}(i') \end{bmatrix} \right\} \\ &+ 2 \text{ sym} \left\{ \sum_{\substack{j \in E\Gamma_l \\ j \in \Omega^{+}}} \left\{ d(j) A(i') P_0^{\infty}(i) \begin{bmatrix} \frac{1}{2}\mathbf{R}(j)\mathbf{R}(j) \\ -\mathbf{R}(j)\mathbf{B}(i) \end{bmatrix} \right\} \right\} + \text{exp}. \end{aligned} \tag{4.24}$$

Subject to a posteriori verification, assume a trial solution for the second-order local moment, $\mathbf{P}_2$, of the form

$$\mathbf{P}_2(i) \approx P_0^{\infty}(i) \left\{ \bar{\mathbf{U}}^{*}\bar{\mathbf{U}}^{*} t^2 + 2 \text{ sym} \left[\bar{\mathbf{U}}^{*}\mathbf{B}(i)\right] t + 2\bar{\mathbf{D}}^{*} t + \mathbf{H}(i) \right\} + \text{exp}, \tag{4.25}$$

with the constant dyadic $\bar{\mathbf{D}}^{*}$ and dyadic field $\mathbf{H}(i)$ to be determined forthwith.

To compute $\bar{\mathbf{D}}^{*}$, form the weighted second-order total moment from (3.12), with order $m = 2$, and (4.25), and differentiate the resulting expression with respect to time, so as to obtain

$$\frac{d\mathbf{M}_2}{dt} \approx 2\bar{\mathbf{U}}^{*}\bar{\mathbf{U}}^{*} t + 2\bar{\mathbf{D}}^{*} + 2 \text{ sym} \sum_{i \in \Gamma_l} v(i) P_0^{\infty}(i) A(i) \bar{\mathbf{U}}^{*} \mathbf{B}(i) + \text{exp}. \tag{4.26}$$

The summation appearing in (4.26) may be simplified by forming the product of (4.12) with $A(i)\mathbf{B}(i)$, and subsequently summing the result over $i \in \Gamma_l$, thereby yielding

(4.27)
$$2\ \mathrm{sym}\left[\sum_{i\in\Gamma_l} v(i)P_0^{\infty}(i)A(i)\bar{\mathbf{U}}^{*}\mathbf{B}(i)\right] = 2\ \mathrm{sym}\left[\sum_{i\in\Gamma_l}\alpha(i)A(i)\mathbf{B}(i)\right] + 2\ \mathrm{sym}\,(\mathbf{E})\,,$$

where $\mathbf{E}$ is the constant dyadic

$$\begin{aligned}\mathbf{E} = &\sum_{\substack{j\in E\Gamma_l\\ j\in\Omega^{+}}} [c(j)+d(j)]\,P_0^{\infty}(i')A(i)\mathbf{B}(i)\,[\mathbf{B}\,(i')-\mathbf{B}(i)]\\ &-\sum_{\substack{j\in E\Gamma_l\\ j\in\Omega^{-}}} d(j)P_0^{\infty}(i')A(i)\mathbf{B}(i)\,[\mathbf{B}(i)-\mathbf{B}(i')]\,.\end{aligned} \tag{4.28}$$

The dyadic $\mathbf{E}$ may itself be simplified upon multiplying (4.2) by $A(i)\mathbf{B}(i)\mathbf{B}(i)$ and (3.18) by $P_0^{\infty}(i)\mathbf{B}(i)\mathbf{B}(i)$, summing both results over $i \in \Gamma_l$ with use of (3.16), and forming their difference, so as to obtain the expression

(4.29)
$$\sum_{\substack{j\in E\Gamma_l\\ j\in\Omega^{+}}} \{[c(j)+d(j)]\,A(i)P_0^{\infty}(i') - d(j)A(i')P_0^{\infty}(i)\}\,[\mathbf{B}(i)\mathbf{B}(i)-\mathbf{B}(i')\mathbf{B}(i')] = \mathbf{0}.$$

Upon adding the null result (4.29) to (4.28), and using the definition (4.14), the symmetric portion of $\mathbf{E}$ is found to possess the form

$$2\ \mathrm{sym}\,(\mathbf{E}) = -\sum_{\substack{j\in E\Gamma_l\\ j\in\Omega^{+}}} \{[c(j)+d(j)]\,A(i)P_0^{\infty}(i') + d(j)A(i')P_0^{\infty}(i)\}\,\mathbf{b}(j)\mathbf{b}(j). \tag{4.30}$$

Equation (4.27) may be further simplified by using (4.13) jointly with the identity (3.16) to show that

$$\sum_{i\in\Gamma_l}\alpha(i)A(i)\mathbf{B}(i) = \sum_{\substack{j\in E\Gamma_l\\ j\in\Omega^{+}}} \left\{ \begin{array}{c} [c(j)+d(j)]\,\mathbf{R}(j)A(i)P_0^{\infty}(i')\mathbf{B}(i) \\ -d(j)\mathbf{R}(j)A(i')P_0^{\infty}(i)\mathbf{B}(i') \end{array} \right\}. \tag{4.31}$$

Upon comparing (4.24) with our trial solution (4.26), and making use of (4.27), (4.30), and (4.31), as well as the identity (3.16), there results the expression

$$\bar{\mathbf{D}}^{*} = \frac{1}{2}\sum_{\substack{j\in E\Gamma_l\\ j\in\Omega^{+}}} \{c(j)A(i)P_0^{\infty}(i') + d(j)\,[A(i)P_0^{\infty}(i') + A(i')P_0^{\infty}(i)]\}\,\tilde{\mathbf{b}}(j)\tilde{\mathbf{b}}(j), \tag{4.32}$$

where the vector $\tilde{\mathbf{b}}(j)$ is defined as

$$\tilde{\mathbf{b}}(j) \overset{\text{def.}}{=} \mathbf{R}(j) - \mathbf{b}(j). \tag{4.33}$$

Moreover, we see that $\bar{\mathbf{D}}^{*}$ represents the solute dispersivity dyadic, inasmuch as $\bar{\mathbf{D}}^{*}$ may also calculated from its definition [14], namely,

$$\bar{\mathbf{D}}^{*} = \frac{1}{2}\lim_{t\to\infty}\frac{d}{dt}\left(\mathbf{M}_2 - \mathbf{M}_1\mathbf{M}_1\right). \tag{4.34}$$

The latter is seen to accord with the result (4.32) upon use of (4.11) and (4.26).

Equation (4.32) enforces an equality between the trial solution $\mathbf{M}_2$, (4.26), and its derived formula (4.24). Consequently, a posteriori verification of (4.25) is completed by deriving a solvable difference equation for $\mathbf{H}(i)$ [37]. To do so, substitute the trial solution (4.25) into (3.5), with $m = 2$, and use (3.9). Removing the time-dependent terms via (4.2) and (4.12), and subsequently substituting for the reaction term upon multiplying (4.2) by $\mathbf{H}(i)$, ultimately furnishes the governing equation for $\mathbf{H}(i)$, namely,

(4.35)
$$\sum_{\substack{j\in\Omega^+(i)\\ j=\{i',i\}}} [c(j)+d(j)]\,P_0^\infty(i')\,[\mathbf{H}(i')-\mathbf{H}(i)] - \sum_{\substack{j\in\Omega^-(i)\\ j=\{i,i'\}}} d(j)P_0^\infty(i')\,[\mathbf{H}(i)-\mathbf{H}(i')] = \beta(i),$$

with $\beta(i)$ the symmetric forcing function

$$\beta(i) = 2\ \mathrm{sym}\left\{\begin{array}{l} v(i)P_0^\infty(i)\left[\bar{\mathbf{U}}^*\mathbf{B}(i)+\bar{\mathbf{D}}^*\right] \\ -\displaystyle\sum_{\substack{j\in\Omega^+(i)\\ j=\{i',i\}}} [c(j)+d(j)]\,P_0^\infty(i)\left[\frac{1}{2}\mathbf{R}(j)\mathbf{R}(j)+\mathbf{R}(j)\mathbf{B}(i')\right] \\ -\displaystyle\sum_{\substack{j\in\Omega^-(i)\\ j=\{i,i'\}}} d(j)P_0^\infty(i')\left[\frac{1}{2}\mathbf{R}(j)\mathbf{R}(j)-\mathbf{R}(j)\mathbf{B}(i')\right] \end{array}\right\}. \tag{4.36}$$

Inasmuch as the structure of (4.35) is identical to that of (4.12), (4.35) will possess a solution if (4.12) itself possesses a solution. As was the case in the original development [36] of this moment technique, computing $\mathbf{H}$ proves unnecessary. Indeed, we have already derived formulas for all the relevant macrotransport parameters without prior knowledge of $\mathbf{H}$. Rather, the demonstrated existence of the latter time-independent, solvable equation (4.35) simply completes the a posteriori verification of (4.25).

The latter verification permits computing $\mathbf{M}_2'$ by substituting (4.25) into (3.11), choosing $m = 2$, and invoking (4.5) and (4.23). Thereby, one obtains

$$\mathbf{M}_2' \approx \bar{\mathbf{U}}^*\bar{\mathbf{U}}^* t^2 + 2\ \mathrm{sym}\left(\bar{\mathbf{B}}\bar{\mathbf{U}}^*\right) t + 2\bar{\mathbf{D}}^* t + \sum_{i\in\Gamma_l} v(i)P_0^\infty(i)\mathbf{H}(i) + \mathrm{exp.} \tag{4.37}$$

Differentiation of the latter with respect to time, followed by subsequent comparison of the resulting expression with (4.26), reveals that the time rates of change of $\mathbf{M}_2$ and $\mathbf{M}_2'$ differ only by exponentially small terms at long times.

5. Example.

5.1. Kinematics. In the following detailed example, the general paradigm developed above is applied to the network depicted in Figure 3. This medium may be envisioned as being composed of a pair of infinitely extended parallel rows of wells, where the wells are connected via thin capillary tubes in the manner indicated in the figure. The centroids of the wells are separated by a distance l. When solute is present in well a, it is assumed to be depleted by a chemical reaction at the uniform rate k. The capillaries connecting these wells possess respective cross-sectional areas A and lengths λl ($\lambda < 1$). Transport occurs within all channels by molecular diffusion, quantified by the diffusion coefficient D_m. Application of the externally applied force F

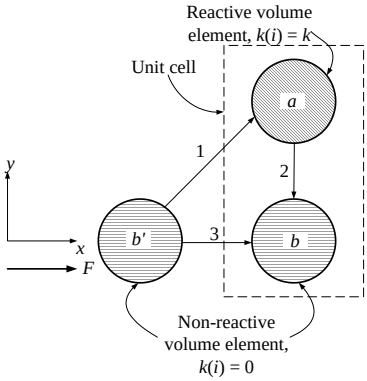


FIG. 3. *Basic graph of a model reactive porous medium. The unit cell, indicated by the dashed box, consists of two nodes, labeled a and b, connected by edges* $j = \{1, 2, 3\}$. *A reactive solute molecule possessing molecular diffusivity* D_m *is assumed not to react when present in subvolume element* $v(b)$, *owing, say, to the absence of a catalyst there, and to be consumed at the rate* k *when present in subvolume element* $v(a)$, *owing, say, to the presence of a catalyst. Application of an externally applied force of magnitude* F *in the* x*-direction gives rise to deterministic solute transport exclusively through edge* 3.

gives rise to solute transport in the x-direction. The solute is assumed to be point-size, whereupon no contribution arises from capillary-scale Taylor–Aris dispersion owing to the absence of a solvent velocity field.

The present example possesses an alternate interpretation as a model of solute transport via bulk flow (through edge 3) with periodic sites for (potentially irreversible) adsorption to the walls. Transport through edge 1 corresponds to diffusive transport from the bulk to adsorption site, while reversibly adsorbed solutes are returned to the bulk flow by diffusion through edge 2. (The depletion "reaction" corresponds to irreversible adsorbtion.) In such a model, the volume of node a, when scaled with the volumetric flow rate in edge 3, corresponds to the (average) residence time of the reversible adsorption process, while the volume of node b represents the volume of a period of the bulk channel containing one adsorption site. In essence, this model is equipollent with our previous analysis of entropic trapping [13], where the retention in the traps is analogous to the present irreversible adsorption process.

The periodic unit cell, indicated by the dashed box, consists of the pair of nodes, $i = \{a, b\}$, characterized by the parameters

$$\mathbf{v} = \tau_0 \begin{bmatrix} \phi_a & 0 \\ 0 & \phi_b \end{bmatrix}, \qquad \mathbf{k} = \begin{bmatrix} k & 0 \\ 0 & 0 \end{bmatrix}, \tag{5.1}$$

where $\tau_0 = v(a) + v(b)$ is the (accessible) volume of the unit cell, and ϕ_a and ϕ_b are the volume fractions of nodes a and b, respectively. The nodes are connected by a trio of edges, $j = \{1, 2, 3\}$, whose edge transport rates (2.6) and macroscopic jump vector are, respectively, given by

$$\mathbf{c} = \frac{D_m F A}{kT} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{d} = \frac{D_m A}{\lambda l} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{R} = l \begin{bmatrix} \hat{\mathbf{x}} \\ 0 \\ \hat{\mathbf{x}} \end{bmatrix}, \tag{5.2}$$

with kT the Boltzmann factor.

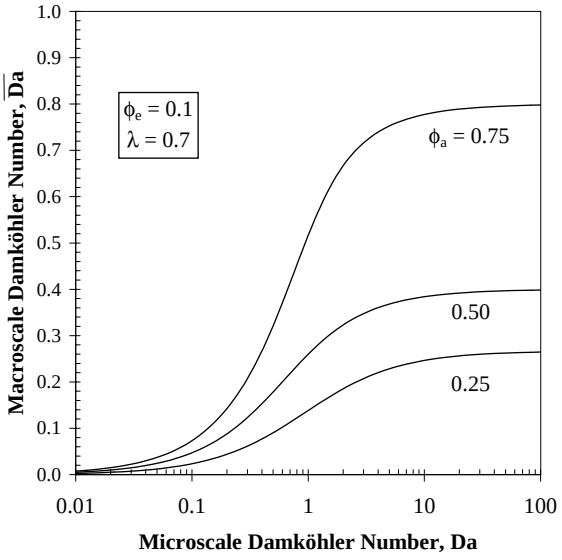


FIG. 4. *Plot of the macroscale Damköhler number, $\overline{Da}$, as a function of the microscale Damköhler number, Da, for several values of the volume fraction of the reactive well, ϕ_a, and for the specified geometric attributes shown in the inset.*

In what follows, it is useful to define the dimensionless parameters

$$\phi_e \overset{\text{def.}}{=} \frac{Al}{\lambda\tau_0}, \quad \mathrm{Da} \overset{\text{def.}}{=} \frac{kl^2}{D_m}, \quad \mathrm{Pe} \overset{\text{def.}}{=} \frac{Fl}{kT}, \tag{5.3}$$

which, respectively, correspond to the volume fraction of the edges, and the microscale Damköhler and Peclét numbers.

5.2. Macrotransport solution. With the geometrical and phenomenological microscale transport data now specified, the eigenvalue problem (4.3) may be rendered in dimensionless form as

$$\begin{bmatrix} 2\frac{\phi_e}{\phi_a} + \mathrm{Da} & -2\frac{\phi_e}{\phi_a} \\ -2\frac{\phi_e}{\phi_b} & 2\frac{\phi_e}{\phi_b} \end{bmatrix} \begin{bmatrix} P_0^\infty(a) \\ P_0^\infty(b) \end{bmatrix} = \frac{\bar{K}l^2}{D_m} \begin{bmatrix} P_0^\infty(a) \\ P_0^\infty(b) \end{bmatrix}. \tag{5.4}$$

Upon defining the macroscopic Damköhler number as

$$\overline{\mathrm{Da}} \overset{\text{def.}}{=} \frac{\bar{K}^* l^2}{D_m}, \tag{5.5}$$

the solution of the eigenvalue problem (5.4) reveals that

$$\overline{\mathrm{Da}} = \frac{\phi_e}{\phi_a\phi_b} + \frac{\mathrm{Da}}{2} - \left[\left(\frac{\phi_e}{\phi_b} - \mathrm{Da}\right)^2 + \frac{\phi_e}{\phi_a}\mathrm{Da} + \frac{\phi_e^2}{\phi_a^2\phi_b}(1+\phi_a)\right]^{\frac{1}{2}}. \tag{5.6}$$

From inspection, we see that $\overline{\mathrm{Da}}$ vanishes with Da; likewise, $\overline{\mathrm{Da}}$ approaches infinity linearly as Da approaches infinity. (Of course, this corresponds to the uninteresting limit where all of the solute mass is depleted instantaneously [5].) Figure 4 displays

numerical values of $\overline{\mathrm{Da}}$ (as a function of Da) for several different values of ϕ_a. $\overline{\mathrm{Da}}$ is seen to increase with Da (the apparent asymptotes appearing at Da ≈ 10 being artifacts of the semilog plot). Owing to the fact that molecular diffusion is the sole mechanism for transporting solute into the reactive well, the overall reaction rate is much slower than that prevailing in well a. By increasing the volume of the reactive well, thereby increasing the solute residence time therein, $\overline{\mathrm{Da}}$ will increase monotonically, all other things being equal.

With use of (4.5), the normalized (dimensionless) eigenvectors corresponding to the smallest eigenvalue (5.6) are, respectively,

$$P_0^{\infty}(a)\tau_0 = \frac{\phi_a - \gamma}{\phi_a\,(1-\gamma)}, \tag{5.7}$$

$$P_0^{\infty}(b)\tau_0 = (1-\gamma)^{-1}, \tag{5.8}$$

where γ denotes the following combination of dimensionless parameters:

$$\gamma \equiv \frac{\phi_a \phi_b}{2\phi_e}\overline{\mathrm{Da}}. \tag{5.9}$$

Substitution of (5.6) into (3.19), together with explicitly incorporating the normalization condition (4.6) in the first row (in lieu of the equation corresponding to $i = a$), furnishes the following matrix equation for the $A(i)$:

$$\begin{bmatrix} \phi_a P_0^{\infty}(a)\tau_0 & \phi_b P_0^{\infty}(b)\tau_0 \\ 2\phi_e & \phi_b\overline{\mathrm{Da}} - 2\phi_e \end{bmatrix} \begin{bmatrix} A(a) \\ A(b) \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \tag{5.10}$$

Solution of (5.10) yields the respective fictitious initial conditions,

$$A(a) = \beta\,(1-\gamma)\,(\phi_a - \gamma)\,, \tag{5.11}$$

$$A(b) = \phi_a \beta\,(1-\gamma)\,, \tag{5.12}$$

where β is the following combination of dimensionless parameters:

$$\beta^{-1} \equiv \phi_a - 2\phi_a\gamma + \gamma^2. \tag{5.13}$$

It is readily verified that the solutions (5.11)–(5.12) satisfy (3.18) for $i = a$. Moreover, $A(a) = A(b) = 1$ in the nonreactive limit, Da$\rightarrow 0$, as would be expected.

Armed with knowledge of $A(i)$ and $P_0^{\infty}(i)$, the mean velocity may be calculated from the summation (4.9), yielding

$$\bar{\mathbf{U}}^* = \hat{\mathbf{x}} D_m A \left\{ \mathrm{Pe} A(b) P_0^{\infty}(b) + \lambda^{-1}\left[A(a)P_0^{\infty}(b) - A(b)P_0^{\infty}(a)\right]\right\}. \tag{5.14}$$

The latter result may be simplified and rendered dimensionless via use of (5.7)–(5.8) and (5.11)–(5.12), yielding

$$\bar{\mathbf{U}}^* = \hat{\mathbf{x}}\left(\frac{D_m F}{kT}\right)\hat{U}^*, \tag{5.15}$$

in which $\hat{U}^*$ is the dimensionless scalar coefficient (i.e., speed)

$$\hat{U}^* = \lambda\beta\phi_a\phi_e. \tag{5.16}$$

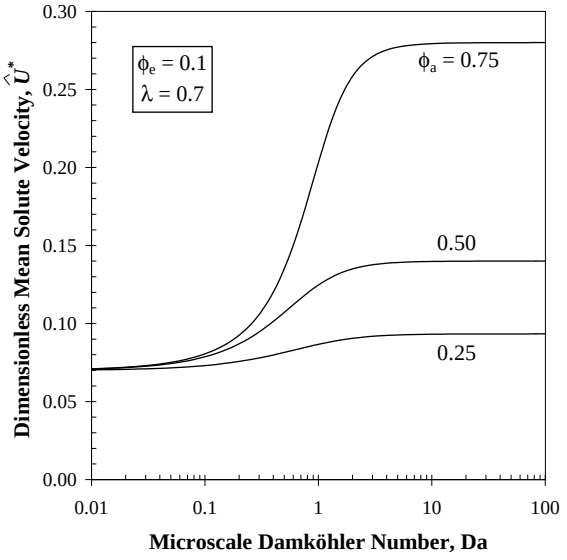


FIG. 5. *Plot of the dimensionless mean solute velocity, $\hat{U}^*$, as a function of the microscale Damköhler number, Da, for several values of the volume fraction of the reactive well, ϕ_a, and for the specified geometric attributes shown in the inset.*

The latter result reduces to $\hat{U}^* = \lambda\phi_e$ in the nonreactive Da$\rightarrow$ 0 limit, in accord with the mean velocity computed from nonreactive network theory [12]. The dimensionless mean velocity is plotted for nonzero values of Da in Figure 5. Since the solute is able to sample the tortuous diffusion path through well a, the (dimensionless) speed $\hat{U}^*$ is less than unity. The mean velocity increases with increasing reaction rate, since the solute entering well a is then depleted at a greater rate, thereby reducing its contribution to the overall transport rate. Similarly, increasing the residence time in well a causes $\hat{U}^*$ to increase. In the limit of infinite reaction rate, one would expect that no solute entering well a could contribute to the overall solute velocity, thereby leading to the value $\hat{U}^* = 1$. This contrasts with a naïve limit of (5.16), which would seem to imply that $\hat{U}^* = 0$. The latter incorrect limit derives from the singularity of (5.4) at Da$\rightarrow \infty$. While not essential to this illustrative example, the proper limiting behavior could be analyzed by rescaling the problem, i.e., solving for the ratio $\bar{K}^*/k$.

The existence of numerous candidates ((4.12), (4.15), (4.19), and (4.21)) for computing the **B**-field makes the choice of its solution protocol flexible. Upon noting that the edge subsets $j = \{1, 2\}$ and $j = \{3\}$ are independent cycles on the local graph, it follows from (4.16) that

$$\mathbf{b}(1) = -\mathbf{b}(2), \tag{5.17}$$

$$\mathbf{b}(3) = \mathbf{0}. \tag{5.18}$$

Use of (4.15) with $i = a$ furnishes the algebraic equation,

$$\phi_e P_0^\infty(b)\tau_0 \frac{\mathbf{b}(2)}{l} - \phi_e P_0^\infty(b)\tau_0 \frac{\mathbf{b}\,(1)}{l} = \phi_a P_0^\infty(a)\tau_0 \mathrm{Pe}\bar{\mathbf{U}}^* \left(\frac{kT}{D_m F}\right) - \hat{\mathbf{x}}\phi_e P_0^\infty(b)\tau_0. \tag{5.19}$$

Together with (5.7)–(5.8), (5.16), and (5.17), this furnishes the solution

$$\frac{\mathbf{b}(1)}{l} = \frac{\hat{\mathbf{x}}}{2}\left[1 - \mathrm{Pe}\lambda\beta\phi_a\left(\phi_a - \gamma\right)\right]. \tag{5.20}$$

It is readily verified that (5.17), (5.18), and (5.20) satisfy (4.15) with $i = b$.

From (4.33), $\tilde{\mathbf{b}}(j) = \hat{\mathbf{x}}\tilde{b}(j)l$, wherein the dimensionless scalar coefficients $\tilde{b}(j)$ possess the respective functional forms,

$$\tilde{b}(1) = 1/2\left[1 + \mathrm{Pe}\lambda\beta\phi_a\left(\phi_a - \gamma\right)\right], \tag{5.21}$$

$$\tilde{b}(2) = 1/2\left[1 - \mathrm{Pe}\lambda\beta\phi_a\left(\phi_a - \gamma\right)\right], \tag{5.22}$$

$$\tilde{b}(3) = 1. \tag{5.23}$$

The dispersivity is calculated from (4.32) as

$$\bar{\mathbf{D}}^* = \hat{\mathbf{x}}\hat{\mathbf{x}}\frac{DAl}{2\lambda}\left\{\begin{array}{c}\left[A(a)P_0^{\infty}(b) + A(b)P_0^{\infty}(a)\right]\left[\tilde{b}^2(1) + \tilde{b}^2(2)\right] \\ + \left(\lambda\mathrm{Pe} + 2\right)A(b)P_0^{\infty}(b)\tilde{b}^2(3)\end{array}\right\}, \tag{5.24}$$

which, with use of (5.7)–(5.8), (5.11)–(5.12), and (5.21)–(5.23), ultimately furnishes the dispersivity dyadic,

$$\bar{\mathbf{D}}^* = \hat{\mathbf{x}}\hat{\mathbf{x}}D_m\hat{D}^*, \tag{5.25}$$

with $\hat{D}^*$ the dimensionless scalar dispersivity,

$$\hat{D}^* = \frac{\phi_e\beta}{2}\left[3\phi_a - \gamma + \phi_a\lambda\mathrm{Pe} + \left(\phi_a\lambda\beta\right)^2\left(\phi_a - \gamma\right)^3\mathrm{Pe}^2\right]. \tag{5.26}$$

In the nonreactive limit,

$$\lim_{\mathrm{Da}\to 0}\hat{D}^* = \frac{\phi_e}{2}\left[3 + \lambda\mathrm{Pe} + \left(\lambda\phi_a\right)^2\mathrm{Pe}^2\right], \tag{5.27}$$

which is identical to the dispersivity calculated directly from the nonreactive theory [12].

Figure 6 portrays the dispersivity $\hat{D}^*$ as a function of Da for several different values of Pe. As is typically the case [9], the dispersivity increases with increasing Peclét number, all other things being equal. The latter effect owes its origin to the "mechanical" dispersion [12] arising from the "delay time" introduced by the tortuous path through node a. The magnitude of latter effect, which measures the "spread" between one solute particle which takes the tortuous path $b' \to a \to b$ and a second particle which takes the path $b' \to b$, depends upon the rate of convection and increases with increasing Pe.

For circumstances wherein $\mathrm{Da} < 1$, the dispersivity gradually increases from its nonreactive value, (5.27). Indeed, normalizing (5.26) for $\hat{D}^*$ with the nonreactive value (5.27) collapses the data for $\mathrm{Da} < 1$ onto a relatively thin band. This increase arises from the increase in the mean velocity over the same range of Da. Explicitly, the dispersivity is a measure of the deviations from the mean solute motion. As Da increases linearly, $\hat{U}^*$ increases dramatically (see Figure 5). Consequently, the mechanical dispersion caused by transport through a increases. Moreover, this increase more than offsets the decrease in dispersion which occurs from solute depletion.

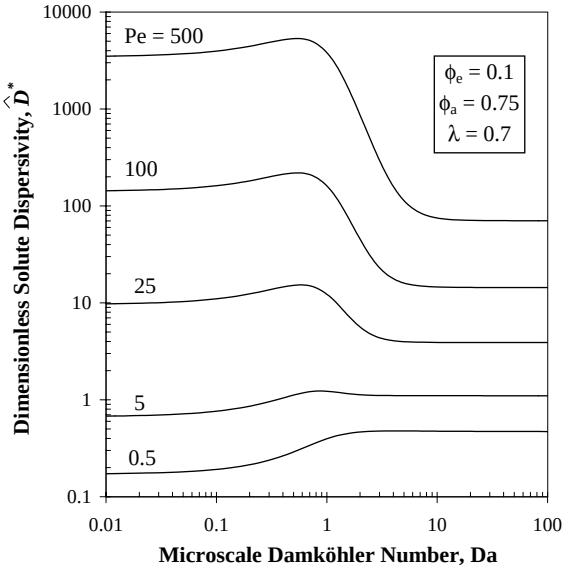


FIG. 6. *Plot of the dimensionless dispersivity, $\hat{D}^*$, as a function of the microscale Damköhler number, Da, for several values of the Peclét number, Pe, and for the specified geometric attributes shown in the inset.*

At Da $= 1$, the dispersion either levels off or undergoes a precipitous drop. When $\hat{D}^* < 1$ at Da $= 1$, the dispersion begins approaching its asymptotic value of $\hat{D}^* = 1$. The latter corresponds to the case where the only source of dispersion is from node b, since all solute molecules entering node a are depleted. When $\hat{D}^* > 1$ at Da $= 1$, the dispersivity decreases since the increasing reaction rate (at a fixed convection rate) serves to diminish the mechanical dispersion arising from the tortuous path through a. As Da $\to \infty$, the only contributions to $\hat{D}^*$ would be expected to arise from molecular diffusion in the nonreactive channels and mechanical dispersion resulting from the mixing process in well b. As was the case with the mean velocity, the singular nature of this limit prevents recovery of the proper limiting behavior directly from (5.26).

Acknowledgment. We are grateful to Dr. Sangtae Kim of Eli Lilly & Company for his encouragement in our microfluidic analyses.

REFERENCES

[1] P. M. ADLER AND H. BRENNER, *Transport processes in spatially periodic capillary networks. I. Geometrical description and linear flow hydrodynamics.*, PhysicoChem. Hydrodyn., 5 (1984), pp. 245–268.

[2] V. ALVARADO, H. T. DAVIS, AND L. E. SCRIVEN, *Effects of pore-level reaction on dispersion in porous media*, Chem. Eng. Sci., 52 (1997), pp. 2865–2881.

[3] J. S. ANDRADE, D. A. STREET, Y. SHIBUSA, S. HAVLIN, AND H. E. STANLEY, *Diffusion and reaction in percolating pore networks*, Phys. Rev. E, 55 (1997), pp. 772–777.

[4] B. E. AVILES AND M. D. LEVAN, *Network models for nonuniform flow and adsorption in fixed beds*, Chem. Eng. Sci., 46 (1991), pp. 1935–1944.

[5] V. BALAKOTAIAH AND H.-C. CHANG, *Dispersion of chemical solutes in chromatographs and reactors*, Philos. Trans. R. Soc. Lond. Ser. A, 351 (1995), pp. 39–75.

[6] V. BALAKOTAIAH AND S. M. S. DOMMETI, *Effective models for packed-bed catalytic reactors*, Chem. Eng. Sci., 54 (1999), pp. 1621–1638.

[7] R. P. Batycky, D. A. Edwards, and H. Brenner, *On the need for fictitious initial conditions in effective medium theories of transient nonconservative transport phenomena. Some elementary unsteady-state heat conduction examples*, Chem. Eng. Comm., 152–153 (1996), pp. 173–187.
[8] B. Bollobas, *Graph Theory: An Introductory Course*, Springer–Verlag, New York, 1979.
[9] H. Brenner and D. A. Edwards, *Macrotransport Processes*, Butterworth–Heinemann, Boston, 1993.
[10] R. I. Cukier, *Diffusion controlled processes among stationary reactive sinks: Effective medium approach*, J. Chem. Phys., 78 (1983), pp. 2573–2578.
[11] R. I. Cukier, *Effective medium theory of rate processes among stationary reactive sinks with the radiation boundary condition*, J. Phys. Chem., 87 (1983), pp. 582–586.
[12] K. D. Dorfman and H. Brenner, *Generalized Taylor-Aris dispersion in discrete spatially periodic networks. Microfluidic applications*, Phys. Rev. E, 65 (2002), p. 021103.
[13] K. D. Dorfman and H. Brenner, *Modeling DNA electrophoresis in microfluidic entropic trapping devices*, Biomed. Microdev., 4 (2002), pp. 237–244.
[14] S. R. Dungan, M. Shapiro, and H. Brenner, *Convective-diffusive-reactive Taylor dispersion processes in particulate multiphase systems*, Proc. R. Soc. Lond. Ser. A, 429 (1990), pp. 639–671.
[15] D. A. Edwards, M. Shapiro, and H. Brenner, *Dispersion and reaction in two-dimensional model porous media*, Phys. Fluids A. 5 (1993), pp. 837–848.
[16] T. R. Ginn, *Stochastic-convective transport with nonlinear reactions and mixing: Finite streamtube ensemble formulation for multicomponent reaction systems with intrastreamtube dispersion*, J. Contam. Hydrol., 47 (2001), pp. 1–28.
[17] M. P. Hollewand and L. F. Gladden, *Modeling of diffusion and reaction in porous catalysts using a random 3-dimensional network model*, Chem. Eng. Sci., 47 (1992), pp. 1761–1770.
[18] S. C. Jakeway, A. J. de Mello, and E. L. Russell, *Miniaturized total analysis systems for biological analysis*, Fresenius J. Anal. Chem., 366 (2000), pp. 525–539.
[19] K. Jensen, *Microreaction engineering — Is smaller better?*, Chem. Eng. Sci., 56 (2001), pp. 293–303.
[20] M. Krishnan, V. Namasivayam, R. S. Lin, R. Pal, and M. A. Burns, *Microfabricated reaction and separation systems*, Curr. Opin. Biotechnol., 12 (2001), pp. 92–98.
[21] J. Kruger, *Effective medium theory for diffusion-controlled reactions among stationary perfect sinks*, Physica A, 166 (1990), pp. 206–219.
[22] J. Kruger, *Effective medium theory of diffusion and chemical-reaction in the presence of stationary sinks*, Physica A, 169 (1990), pp. 393–406.
[23] J. P. Kutter, *Current developments in electrophoretic and chromatographic separation methods on microfabricated devices*, Trac-Trends Anal. Chem., 19 (2000), pp. 352–363.
[24] M. Leitzelement, P. Maj, J. A. Dodds, and J. L. Greffe, *Deep bed filtration in a network of random tubes*, in Solid-Liquid Separation, J. Gregory, ed., Ellis Horwood, Chichester, U.K., 1984, pp. 273–296.
[25] G. Y. Li and H. Rabitz, *A general lumping analysis of a reaction system coupled with diffusion*, Chem. Eng. Sci., 46 (1991), pp. 2041–2053.
[26] K. Mattern and B. U. Felderhof, *Rate of diffusion-controlled reactions in a random array of spherical sinks*, Physica A, 143 (1987), pp. 1–20.
[27] R. Mauri, *Dispersion, convection, and reaction in porous-media*, Phys. Fluids, 3 (1991), pp. 743–756.
[28] K. N. Mehta, M. C. Tiwari, and K. D. P. Nigam, *Effect of permeability and chemical-reaction on laminar dispersion of a solute*, Int. J. Eng. Fluid Mech., 1 (1988), pp. 351–364.
[29] M. Muthukumar, *Concentration dependence of diffusion controlled processes among static traps*, J. Chem. Phys., 76 (1982), pp. 2667–2671.
[30] D. Pal, *Effect of chemical reaction on the dispersion of a solute in a porous medium*, Appl. Math. Model., 23 (1999), pp. 557–566.
[31] S. H. Park and Y. G. Kim, *The effect of chemical-reaction on effective diffusivity within biporous catalysts.* I. *Theoretical development*, Chem. Eng. Sci., 39 (1984), pp. 523–531.
[32] S. D. Rege and H. S. Fogler, *A network model for deep bed filtration of solid particles and emulsion drops*, AIChE J., 34 (1988), pp. 1761–1772.
[33] D. Ryan, R. G. Carbonell, and S. Whitaker, *Effective diffusivities for catalyst pellets under reactive conditions*, Chem. Eng. Sci., 35 (1980), pp. 10–16.
[34] M. Sahimi, G. Gavalas, and T. T. Tsotsis, *Statistical and continuum models of fluid-solid reactions in porous media*, Chem. Eng. Sci., 45 (1990), pp. 1443–1502.
[35] M. Shapiro and H. Brenner, *Taylor dispersion of chemically reactive species: Irreversible first-order reactions in bulk and on boundaries*, Chem. Eng. Sci., 41 (1986), pp. 1417–1433.

[36] M. Shapiro and H. Brenner, *Chemically reactive generalized Taylor dispersion phenomena*, AIChE J., 33 (1987), pp. 1155–1167.
[37] M. Shapiro and H. Brenner, *Dispersion of a chemically reactive solute in a spatially periodic model of a porous medium*, Chem. Eng. Sci., 43 (1988), pp. 551–571.
[38] B. J. Suchomel, B. M. Chen, and M. B. Allen, *Network model of flow, transport and biofilm effects in porous media*, Transp. Porous Media, 30 (1998), pp. 1–23.
[39] N. Wakad and Y. Nardse, *Effective diffusivities and dead-end pores*, Chem. Eng. Sci., 29 (1973), pp. 1304–1306.
[40] L. Zhang and N. A. Seaton, *The application of continuum equations to diffusion and reaction in pore networks*, Chem. Eng. Sci., 49 (1994), pp. 41–50.

SIAM J. APPL. MATH.
Vol. 63, No. 3, pp. 987–1028

REALIZABLE (AVERAGE STRESS, AVERAGE STRAIN) PAIRS IN A PLATE WITH HOLES*

G. W. MILTON†, S. K. SERKOV†, AND A. B. MOVCHAN‡

Abstract. Here a complete characterization is given of the set of all possible (average stress, average strain) pairs that can exist in a plate containing a fixed volume fraction f of holes. Specifically, for a given average stress, the range of values the average strain takes as the microgeometry is varied (while keeping f fixed) is determined. It is shown that multiple rank laminate materials suffice to generate all possible values of the average strain. When the microgeometry is restricted to be a periodic array of holes, with only one hole per unit cell, the average strain takes a smaller range of values as the hole shape is varied. A certain necessary condition for optimality must be satisfied if the hole is such that the average strain is on the boundary of this range. Numerical results are obtained for the range in the limit where the holes are well separated and occupy a small volume fraction. Optimal hole shapes, associated with average strains on the boundary of the range of admissible values, are identified. Analytical expressions are obtained for certain optimal holes, called critical holes, for which the tangential stress is zero along the smooth portions of their boundary.

Key words. planar elasticity, composites, bounds, optimal microstructures

AMS subject classifications. 74Q20, 74P10

PII. S0036139901395717

1. Introduction.

The objective of this paper is to cast light on the realizable (average stress, average strain) pairs that can occur in planar two phase composites, when one phase is void and the other phase is isotropic. Specifically, for a given fixed applied stress $\boldsymbol{\sigma}_*$ we ask: What is the range $R(f, \boldsymbol{\sigma}_*)$ of values that the average strain $\boldsymbol{\epsilon}_*$ takes as the microgeometry is varied over all possible configurations that have a fixed volume fraction f of the void phase? Also, what microgeometries are associated with average strains $\boldsymbol{\epsilon}_*$ that lie on or near the boundary of the range $R(f, \boldsymbol{\sigma}_*)$? Putting it another way, how should one punch out a fixed area fraction of holes in a stressed plate to obtain an extreme average strain $\boldsymbol{\epsilon}_*$?

Characterizing the range $R(f, \boldsymbol{\sigma}_*)$ and identifying microgeometries associated with the boundary is important for structural design problems. Consider, for example, the following optimal design problem, analogous to one treated by Gibiansky, Lurie, and Cherkaev (1988) for conductivity. A planar body contains a fixed total volume fraction p of voids in an isotropic plate material. The body is subject to given displacement boundary conditions. How should the voids be configured so that the net force on a given segment of the boundary is maximized (or minimized)? This is a tricky computational problem, since the answer might be that it is best to have infinitely many voids distributed in some microstructure. A solution is to consider the relaxed problem, where one allows for composites in the body, and not just plate or void phase. Thus one looks for the best macroscopic stress field $\boldsymbol{\sigma}_*(\mathbf{x})$ and associated macroscopic strain field $\boldsymbol{\epsilon}_*(\mathbf{x})$ compatible with the displacement boundary conditions

*Received by the editors September 26, 2001; accepted for publication (in revised form) August 28, 2002; published electronically February 25, 2003. This research was supported by the EPSRC through grant GR/L41950; by the National Science Foundation through grants DMS-9402763, DMS-9803748, and DMS-0108626; and by the Universities of Bath, Liverpool, and Utah.
http://www.siam.org/journals/siap/63-3/39571.html

†Department of Mathematics, University of Utah, Salt Lake City, UT, 84112 (milton@math.utah.edu, serkov@math.utah.edu).

‡Department of Mathematical Sciences, University of Liverpool, Liverpool, L69 3BX, United Kingdom (abm@maths.liv.ac.uk).

such that the relation $\boldsymbol{\epsilon}_*(\mathbf{x}) \in R(f(\mathbf{x}), \boldsymbol{\sigma}_*(\mathbf{x}))$ holds everywhere in the body for some choice of the function $f(\mathbf{x})$ whose average over the body is p, with $1 \geq f(\mathbf{x}) \geq 0$ for all $\mathbf{x}$. Once this problem is solved, one introduces voids into the body so that the microstructure in the vicinity of each point $\mathbf{y}$ has a local volume fraction $f(\mathbf{x})$ of voids and is such that application to it of a local average stress $\boldsymbol{\sigma}_*(\mathbf{y})$ produces the local average strain $\boldsymbol{\epsilon}_*(\mathbf{y})$.

The relation between the average strain $\boldsymbol{\epsilon}_*$ and average stress $\boldsymbol{\sigma}_*$ takes the form

$$\boldsymbol{\epsilon}_* = \boldsymbol{\mathcal{S}}_* \boldsymbol{\sigma}_*, \tag{1.1}$$

where $\boldsymbol{\mathcal{S}}_*$ is the effective compliance tensor which is a fourth order tensor. This relation can also be expressed in the equivalent form

$$\begin{pmatrix} \epsilon_{11}^* \\ \epsilon_{22}^* \\ \sqrt{2}\epsilon_{12}^* \end{pmatrix} = \mathbf{S}_* \begin{pmatrix} \sigma_{11}^* \\ \sigma_{22}^* \\ \sqrt{2}\sigma_{12}^* \end{pmatrix}, \tag{1.2}$$

where the ϵ_{ij}^* and σ_{ij}^* are the elements of the average strain $\boldsymbol{\epsilon}_*$ and average stress $\boldsymbol{\sigma}_*$ in Cartesian coordinates. In this representation the fourth order effective compliance tensor $\boldsymbol{\mathcal{S}}_*$ is represented by the symmetric 3×3 matrix $\mathbf{S}_*$.

This question of characterizing $R(f, \boldsymbol{\sigma}_*)$ is nontrivial because we do not know the range of values that the effective compliance matrix $\mathbf{S}_*$ can take as the microgeometry is varied. The set of all possible compliance matrices is known as the G-closure at constant volume fraction (see Cherkaev (2000) and references therein) and is denoted as $G_f U$, in which U represents the set of compliance matrices of the two component phases. If we use the elements of the $\mathbf{S}_*$ as coordinates, $G_f U$ is represented (for a fixed volume fraction f) by a set in a six-dimensional space. However, there is one degree of rotational invariance associated with this set, due to the fact that if we rotate the microgeometry, then the effective compliance matrix is transformed in the obvious manner. Due to this invariance we can represent $G_f U$ as a set in a five-dimensional space. In contrast, it requires only four dimensions to represent (for a fixed volume fraction f) all of the sets $R(f, \boldsymbol{\sigma}_*)$ that are generated as $\boldsymbol{\sigma}_*$ varies. To see this we can assume without loss of generality (by choosing the coordinate axes and normalizing the stress appropriately) that the tensor $\boldsymbol{\sigma}_*$ takes the form

$$\boldsymbol{\sigma}_* = \begin{pmatrix} \sigma & 0 \\ 0 & 1 \end{pmatrix} \quad \text{with} \quad |\sigma| \leq 1. \tag{1.3}$$

For each value of σ the range $R(f, \boldsymbol{\sigma}_*)$ can be represented as a set in a three-dimensional space with ϵ_{11}^*, ϵ_{22}^*, and ϵ_{12}^* as coordinates. Thus the entire collection of sets $R(f, \boldsymbol{\sigma}_*)$ obtained as σ is varied can be represented by a single set in a four-dimensional space with σ, ϵ_{11}^*, ϵ_{22}^*, and ϵ_{12}^* as coordinates. Due to the reduction in dimensionality it is anticipated that the task of characterizing all the sets $R(f, \boldsymbol{\sigma}_*)$ should be easier than that of characterizing the G-closure at constant volume fraction.

For linear conductivity the analogous problem has been completely solved for a d-dimensional composite containing an arbitrary number n of anisotropic or anisotropic conducting phases, mixed in fixed proportions by Tartar (1995), following earlier work of Raĭtum (1983) and Murat and Tartar (1985) for the case of composites containing two isotropic phases. For a given fixed applied electric field $\mathbf{e}_*$, the problem is to find the range $R(f_1, \ldots, f_n, \mathbf{e}_*)$ of values that the average current field $\mathbf{j}_*$ takes as the microstructure is varied (allowing for arbitrary spatial variations in the orientation of

each phase), while keeping the volume fractions $f_1, \ldots, f_n$ of the phases fixed. For fixed values of the volume fraction, this range turned out to be simply a sphere in the space with the components j_1^*, j_2^*,$\ldots$,j_d^* of $\mathbf{j}_*$ as coordinates. (It can alternatively be represented as a set in a two-dimensional space with the invariants $\mathbf{j}_* \cdot \mathbf{e}_*/(\mathbf{e}_* \cdot \mathbf{e}_*)$ and $\mathbf{j}_* \cdot \mathbf{j}_*/(\mathbf{e}_* \cdot \mathbf{e}_*)$ as coordinates.) Moreover, average currents on the boundary of the range $R(f_1, \ldots, f_n, \mathbf{e}_*)$ were associated with simple laminates of the n phases. The phases are oriented so that the axes with minimum conductivity are aligned parallel to the direction of lamination $\mathbf{n}$, and so the axes with maximum conductivity are aligned in another perpendicular direction, with the fields $\mathbf{j}_*$ and $\mathbf{e}_*$ both lying in the plane spanned by these two directions. As the angle between $\mathbf{n}$ and $\mathbf{e}_*$ is varied, the vector $\mathbf{j}_*$ ranges over the entire boundary of the sphere $R(f_1, \ldots, f_n, \mathbf{e}_*)$. Laminates appear naturally as the optimal composites since they are the best microgeometries for guiding current in desired directions. In some sense the optimal microstructures we seek in this paper can be regarded as the best microgeometries for guiding stress.

For nonlinear conductivity less is known. The case of composites of two isotropic phases, with one phase occupying a fixed volume fraction f, has been investigated by Milton and Serkov (2000). The set $R(f, \mathbf{e}_*)$ (of all possible average current fields $\mathbf{j}_*$ associated with a given applied electric field $\mathbf{e}_*$) is no longer a sphere, and there are other microstructures besides simple laminates which are needed to generate currents $\mathbf{j}_*$ on the boundary of $R(f, \mathbf{e}_*)$. Curiously, to generate the maximum current in the direction of $\mathbf{e}_*$, it is sometimes necessary to use a simple laminate with its layers oriented perpendicular (rather than parallel) to the field.

The structure of the present paper can be summarized as follows:

- Section 2 gives an essentially complete characterization of the range $R(f, \boldsymbol{\sigma}_*)$ of values that $\boldsymbol{\epsilon}_*$ takes as the microstructure varies over all conceivable configurations. Microgeometries corresponding to values of $\boldsymbol{\epsilon}_*$ on or near the boundary of $R(f, \boldsymbol{\sigma}_*)$ are identified.
- Section 3 considers periodic arrays of holes with only one simply connected hole per unit cell occupying a fixed volume fraction f. A condition for optimality is derived which must necessarily hold if $\boldsymbol{\epsilon}_*$ is at the boundary of its range of admissible values for this restricted class of microgeometry.
- Section 4 presents numerical results for the range of values of $\boldsymbol{\epsilon}_*$ for periodic arrays of well separated holes occupying a fixed infinitesimal volume fraction f. Hole shapes are identified which correspond to values of $\boldsymbol{\epsilon}_*$ at the boundary of its range of admissible values for this restricted class of microgeometry.
- Section 5 gives analytical expressions for the shape of certain holes, which we call critical holes. These have the property that the tangential stress is zero along all smooth portions of the boundary of the hole, for at least one loading. Such holes satisfy the optimality criterion, and the numerical evidence suggests that they are optimal holes.

Throughout the paper we will assume that the plate is in a state of plane stress (i.e., the plate contracts or expands in thickness so that the normal component of the stress on the plate surface remains zero). The formulae are easily adapted to the case of plane strain (where the thickness of the plate is constrained to be constant) by making the standard adjustments to the moduli entering the formulae.

The characterization of $R(f, \boldsymbol{\sigma}_*)$ given in section 2 can be generalized to three-dimensional porous media. The only difficult part is constructing porous "pentamode" materials for which the effective elasticity tensor has one finite nonzero eigenvalue and five zero eigenvalues. The structures given in Figure 5 of Milton and Cherkaev (1995)

provide a basis for constructing such pentamode materials. However, to ensure that one eigenvalue is nonzero, the junctions between the linkages have to be appropriately modified. The three-dimensional case will be analyzed in detail in a forthcoming paper.

2. The essentially complete characterization of $R(f, \boldsymbol{\sigma}_*)$. The characterization of $R(f, \boldsymbol{\sigma}_*)$ is a delicate question because strange effective behaviors can result when one phase is void (Khruslov (1978), Briane (1998), Briane and Mazliak (1998)). Additionally, nonlinear effects, such as the buckling of the layers in a laminate of the matrix and void phases, cannot be ignored. To avoid these technical issues, we will suppose that the void phase is not really void but has a nonzero isotropic elasticity matrix $\mathbf{C}_V$. Then the set $G_f U$ of all possible compliance matrices is well defined, and its associated set $G_f U \boldsymbol{\sigma}_*$ consisting of all strain tensors $\boldsymbol{\epsilon}_*$ such that $\boldsymbol{\epsilon}_* = \mathbf{S}_* \boldsymbol{\sigma}_*$ for some $\mathbf{S}_* \in G_f U$ is also well defined. We define $R(f, \boldsymbol{\sigma}_*)$ as the set which $G_f U \boldsymbol{\sigma}_*$ approaches in the limit as $\mathbf{C}_V$ approaches zero. (It remains to show that this set does not depend upon the path of isotropic positive definite fourth order tensors that $\mathbf{C}_V$ follows as it approaches zero, but the ensuing analysis suggests that this is the case.)

The main result of this section is the following theorem.

THEOREM 1. *The set $R(f, \boldsymbol{\sigma}_*)$ includes all strain tensors $\boldsymbol{\epsilon}_*$ such that*

$$\boldsymbol{\epsilon}_* : \boldsymbol{\sigma}_* > \lim_{\mathbf{C}_V \to 0} \left[\min_{\mathbf{S}_* \in G_f U} \boldsymbol{\sigma}_* : \mathbf{S}_* \boldsymbol{\sigma}_* \right], \tag{2.1}$$

in which the right-hand side (when divided by 2) can be identified with the lower bound on the elastic energy for which an explicit expression is available (Gibiansky and Cherkaev (1984), *Allaire and Kohn* (1993*a*, 1993*b*)*). Hence any tensor $\boldsymbol{\epsilon}_*$ for which $\boldsymbol{\epsilon}_* : \boldsymbol{\sigma}_*$ is less than the right-hand side of* (2.1) *is not contained in $R(f, \boldsymbol{\sigma}_*)$. When* $\det(\boldsymbol{\sigma}_*) \geq 0$*, the set $R(f, \boldsymbol{\sigma}_*)$ also contains those strain tensors $\boldsymbol{\epsilon}_*$ such that equality holds in* (2.1).

This theorem does not quite provide a complete characterization of $R(f, \boldsymbol{\sigma}_*)$, because when $\det(\boldsymbol{\sigma}_*) < 0$ it is still uncertain as to whether tensors $\boldsymbol{\epsilon}_*$ such that equality holds in (2.1) belong to $R(f, \boldsymbol{\sigma}_*)$.

In establishing Theorem 1, we skip many of the technical details. During much of the proof we treat the void phase as being truly void, and then at the end we discuss how the argument needs to be modified when one correctly takes the limit $\mathbf{C}_V \to 0$.

Let $\mathbf{S} = \mathbf{S}(E, \nu)$ be the compliance matrix of the original plate, which for plane stress is given by

$$\mathbf{S}(E, \nu) = \begin{pmatrix} 1/E & -\nu/E & 0 \\ -\nu/E & 1/E & 0 \\ 0 & 0 & (1+\nu)/E \end{pmatrix}, \tag{2.2}$$

where E and ν are the Young's modulus and Poisson's ratio of the isotropic plate material, respectively. Rather than characterizing the set $R(f, \boldsymbol{\sigma}_*)$ of possible average strains $\boldsymbol{\epsilon}_*$ associated with a given average stress, it is simpler to consider the equivalent problem of characterizing the set

$$D(f, \boldsymbol{\sigma}_*) = \frac{R(f, \boldsymbol{\sigma}_*)}{f} - \frac{\mathbf{S}\boldsymbol{\sigma}_*}{f} \tag{2.3}$$

of possible values of

$$\boldsymbol{\epsilon}_V = \frac{[\boldsymbol{\epsilon}_* - \mathbf{S}\boldsymbol{\sigma}_*]}{f} = [\mathbf{S}_*(E, \nu) - \mathbf{S}(E, \nu)] \frac{\boldsymbol{\sigma}_*}{f}, \tag{2.4}$$

when the average stress $\boldsymbol{\sigma}_*$ and the void volume fraction f are fixed. This tensor $\boldsymbol{\epsilon}_V$ can be regarded as the average strain within the void phase. To justify this interpretation, suppose that the void phase was not really void but had elasticity matrix $\mathbf{C}_V$. Let $\boldsymbol{\epsilon}_V$ and $\boldsymbol{\sigma}_V$ represent the average strain and stress in this phase, and let $\boldsymbol{\epsilon}_P$, $\boldsymbol{\sigma}_P$ represent the average strain and stress in the remaining plate. Then we have

$$\begin{aligned}\boldsymbol{\epsilon}_* &= f\boldsymbol{\epsilon}_V + (1-f)\boldsymbol{\epsilon}_P = f\boldsymbol{\epsilon}_V + (1-f)\mathbf{S}\boldsymbol{\sigma}_P \\ &= f\boldsymbol{\epsilon}_V + \mathbf{S}(\boldsymbol{\sigma}_* - f\boldsymbol{\sigma}_V) = f(\mathbf{I} - \mathbf{S}\mathbf{C}_V)\boldsymbol{\epsilon}_V + \mathbf{S}\boldsymbol{\sigma}_*.\end{aligned} \tag{2.5}$$

This relation implies that $\boldsymbol{\epsilon}_V$ is given by (2.4) in the limit $\mathbf{C}_V \to 0$.

The set $D(f, \boldsymbol{\sigma}_*)$ of possible average strains within the void phase must have a rather trivial dependence on E and ν. It has been established (Day et al. (1992), Cherkaev, Lurie, and Milton (1992), Zheng and Hwang (1996, 1997), Hu and Weng (2001)) that $E[\mathbf{S}_*(E,\nu) - \mathbf{S}(E,\nu)]$ is independent of the moduli E and ν of the plate. It follows that $E\boldsymbol{\epsilon}_V$ is independent of the moduli of the plate. Consequently the set $D(f, \boldsymbol{\sigma}_*)$ must be independent of ν and must rescale in proportion to $1/E$ as the Young's modulus E of the plate is varied.

Some elementary bounds on the set $D(f, \boldsymbol{\sigma}_*)$ follow immediately from optimal bounds on the elastic energy $\boldsymbol{\epsilon}_* : \boldsymbol{\sigma}_*/2$, due to Allaire and Kohn (1993a) (see also Gibiansky and Cherkaev (1984) and Allaire and Kohn (1993b) for related energy bounds in planar elasticity). Without loss of generality let us suppose that $\boldsymbol{\sigma}_*$ takes the form (1.3). Then their bounds imply

$$\boldsymbol{\epsilon}_V : \boldsymbol{\sigma}_* \geq B(\sigma), \quad \text{where} \quad B(\sigma) = \frac{(1+|\sigma|)^2}{(1-f)E} \tag{2.6}$$

for any $\boldsymbol{\epsilon}_V \in D(f, \boldsymbol{\sigma}_*)$. Here we will establish that any $\boldsymbol{\epsilon}_V$ satisfying this bound as an inequality can be identified with the average strains within the void phase of a particular multiple rank laminate composite. The only open question is the minor point of whether those $\boldsymbol{\epsilon}_V$ satisfying the bound as an equality are realizable. We can answer this affirmatively only when σ is nonnegative.

First let us consider the case when $\sigma > 0$. It is well known that second rank laminates attain the energy bound. These second rank laminates are obtained by laminating the void phase with the plate phase, and then laminating this structure (on a much larger length scale, with a different direction of lamination) with the plate phase again. In an appropriately chosen second rank laminate subject to the applied stress $\boldsymbol{\sigma}_*$, the average strain $\boldsymbol{\epsilon}_V$ in the void phase is such that (2.6) holds as an equality. Our aim is to show the stronger result that, for every $\boldsymbol{\epsilon}_V$ such that (2.6) holds as an equality, there exists a second rank laminate such that $\boldsymbol{\epsilon}_V$ is the average strain in the void phase. Geometrically the set of $\boldsymbol{\epsilon}_V$ such that (2.6) holds as an equality represents a plane in the space with ϵ^*_{11}, ϵ^*_{22}, and ϵ^*_{12} as coordinates, and we want to show that every point on this plane corresponds to some second rank laminate. Second rank laminates are special, in that their associated effective elasticity matrix $\mathbf{C}_* = \mathbf{S}_*^{-1}$ has one zero eigenvalue. Consequently two different average strains can produce the same average stress. They are unimode materials in the sense of having one easy mode of deformation (Milton and Cherkaev, 1995). A line of average strains (rather than a single average strain) is associated with a given average stress $\boldsymbol{\sigma}_*$. This line lies in the plane $\boldsymbol{\epsilon}_V : \boldsymbol{\sigma}_* = B(\sigma)$. The orientation of the line in this plane varies as one changes the structure of the second rank laminate, while still attaining the bound.

We will now prove that the structure can be adjusted so that the line passes through any desired point in the plane $\boldsymbol{\epsilon}_V : \boldsymbol{\sigma}_* = B(\sigma)$.

Explicit formulae are available for the effective elasticity or compliance tensor of second rank laminates (Francfort and Murat (1986), Gibiansky and Cherkaev (1987)), but we do not need these formulae in the ensuing analysis. Consider a simple laminate of the plate phase and the void phase, mixed in proportions $(1-p)$ and p, with the average strain in the void phase being $\boldsymbol{\epsilon}_V$ and with the interfaces being oriented parallel to a unit vector $\mathbf{t}_1$. The stress $\boldsymbol{\sigma}_P^{(1)}$ and strain $\boldsymbol{\epsilon}_P^{(1)}$ in the plate phase must be

$$\boldsymbol{\sigma}_P^{(1)} = E(\mathbf{t}_1 \cdot \boldsymbol{\epsilon}_V \mathbf{t}_1)\mathbf{t}_1 \otimes \mathbf{t}_1, \quad \boldsymbol{\epsilon}_P^{(1)} = \mathbf{S}\boldsymbol{\sigma}_P^{(1)} \tag{2.7}$$

to ensure compatibility of stresses and strains at each interface. (Here $\mathbf{t}_1 \otimes \mathbf{t}_1$ denotes the rank-1 matrix $\mathbf{t}_1\mathbf{t}_1^T$, in which the row vector $\mathbf{t}_1^T$ is the transpose of the column vector $\mathbf{t}_1$.) The resulting average stress and strain in the entire laminate are therefore $\boldsymbol{\sigma}_*^{(1)} = p\boldsymbol{\sigma}_P^{(1)}$ and $\boldsymbol{\epsilon}_*^{(1)} = p\boldsymbol{\epsilon}_P^{(1)} + (1-p)\boldsymbol{\epsilon}_V$. Next suppose that this laminate has been layered (on a much larger length scale) with new layers of the plate phase, in proportions $(1-q)$ and q, with the new interfaces being oriented parallel to a unit vector $\mathbf{t}_2$. The stress $\boldsymbol{\sigma}_P^{(2)}$ and strain $\boldsymbol{\epsilon}_P^{(2)}$ in the new layers of the plate phase must be

$$\boldsymbol{\sigma}_P^{(2)} = \boldsymbol{\sigma}_*^{(1)} + E(1-p)(\mathbf{t}_2 \cdot \boldsymbol{\epsilon}_V \mathbf{t}_2)\mathbf{t}_2 \otimes \mathbf{t}_2, \quad \boldsymbol{\epsilon}_P^{(2)} = \mathbf{S}\boldsymbol{\sigma}_P^{(2)} \tag{2.8}$$

to ensure compatibility of the stresses $\boldsymbol{\sigma}_P^{(2)}$ and $\boldsymbol{\sigma}_*^{(1)}$ and compatibility of the associated strains, $\mathbf{S}\boldsymbol{\sigma}_P^{(2)}$ and $\boldsymbol{\epsilon}_*^{(1)}$. The total average stress in this rank-2 laminate is therefore

$$\boldsymbol{\sigma}_* = (1-f)E[c_1(\mathbf{t}_1 \cdot \boldsymbol{\epsilon}_V \mathbf{t}_1)\mathbf{t}_1 \otimes \mathbf{t}_1 + c_2(\mathbf{t}_2 \cdot \boldsymbol{\epsilon}_V \mathbf{t}_2)\mathbf{t}_2 \otimes \mathbf{t}_2], \tag{2.9}$$

where $f = (1-p)(1-q)$ is the volume fraction of void in the second rank laminate and c_1 and c_2 are the nonnegative lamination parameters

$$c_1 = \frac{p}{1-f} = \frac{p}{p+q-pq}, \qquad c_2 = \frac{q(1-p)}{1-f} = \frac{q(1-p)}{p+q-pq} \tag{2.10}$$

introduced by Tartar (1985), satisfying $c_1 + c_2 = 1$. For any fixed value of f, $c_1 = 1 - c_2$ can take any value between 0 and 1. Now if $\boldsymbol{\epsilon}_V$, $\mathbf{t}_1$, and $\mathbf{t}_2$ are such that

$$\mathbf{t}_1 \cdot \boldsymbol{\epsilon}_V \mathbf{t}_1 = \mathbf{t}_2 \cdot \boldsymbol{\epsilon}_V \mathbf{t}_2 = \frac{1+\sigma}{(1-f)E}, \tag{2.11}$$

then $\boldsymbol{\sigma}_*$ will have the desired trace $1+\sigma$, and

$$\boldsymbol{\epsilon}_V : \boldsymbol{\sigma}_* = (1-f)E[c_1(\mathbf{t}_1 \cdot \boldsymbol{\epsilon}_V \mathbf{t}_1)^2 + c_2(\mathbf{t}_1 \cdot \boldsymbol{\epsilon}_V \mathbf{t}_1)^2] = \frac{(1+\sigma)^2}{(1-f)E}. \tag{2.12}$$

In other words, the bound (2.6) will be satisfied as an equality. (Note that $\boldsymbol{\sigma}_*$ given by (2.9) necessarily has eigenvalues of the same sign when (2.11) holds.) To ensure that $\boldsymbol{\sigma}_*$ is diagonal we choose

$$c_1 = \frac{-\cos\theta_2 \sin\theta_2}{\cos\theta_1 \sin\theta_1 - \cos\theta_2 \sin\theta_2}, \qquad c_2 = \frac{\cos\theta_1 \sin\theta_1}{\cos\theta_1 \sin\theta_1 - \cos\theta_2 \sin\theta_2}, \tag{2.13}$$

where θ_1 and θ_2 are the angles of the layers, in terms of which

$$\mathbf{t}_1 = \begin{pmatrix} \sin\theta_1 \\ \cos\theta_1 \end{pmatrix}, \quad \mathbf{t}_2 = \begin{pmatrix} \sin\theta_2 \\ \cos\theta_2 \end{pmatrix}. \tag{2.14}$$

Both c_1 and c_2 will be positive if we take θ_1 between 0 and $\pi/2$, and θ_2 between $\pi/2$ and π. To ensure that the diagonal elements of $\boldsymbol{\sigma}_*$ are σ and 1 we take

$$\tan\theta_2 = -\frac{\sigma}{\tan\theta_1}. \tag{2.15}$$

Next note that any $\boldsymbol{\epsilon}_V$ satisfying equality in (2.6) can be represented in the form

$$\boldsymbol{\epsilon}_V = \frac{(1+\sigma)\mathbf{I}}{(1-f)E} + \mathbf{A} \quad \text{with} \quad \mathbf{A} = \begin{pmatrix} a & b \\ b & -a\sigma \end{pmatrix} \tag{2.16}$$

for some choice of the constants a and b. The condition (2.11) then requires that $\mathbf{t}_1$ and $\mathbf{t}_2$ be chosen with

$$\mathbf{t}_1 \cdot \mathbf{A}\mathbf{t}_1 = \mathbf{t}_2 \cdot \mathbf{A}\mathbf{t}_2 = 0. \tag{2.17}$$

This will be satisfied if $\tan\theta_2$ is given by (2.15), and if $\tan\theta_1$ is the unique nonnegative root of the quadratic

$$a(\tan\theta_1)^2 + 2b\tan\theta_1 - a\sigma = 0, \tag{2.18}$$

which always has such a root. By choosing the parameters according to (2.13), (2.15), and (2.18), we have obtained a laminate such that $\boldsymbol{\epsilon}_V$ is the average strain in the void phase and $\boldsymbol{\sigma}_*$ given by (1.3) is the average stress. In the special case when $a = 0$ we have $\theta_1 = 0$ and $\theta_2 = \pi$, and (2.13) is not sufficient to determine c_1 and c_2. (The tensor $\boldsymbol{\sigma}_*$ given by (2.9) is always diagonal.) In this case we choose $c_1 = \sigma/(1+\sigma)$ and $c_2 = 1/(1+\sigma)$ to ensure that the diagonal elements of $\boldsymbol{\sigma}_*$ are σ and 1. In another special case when $b = 0$ the equations are solved with $\tan\theta_1 = -\tan\theta_2 = \sqrt{\sigma}$ and $c_1 = c_2 = 1/2$.

When $\sigma = 0$ a simple laminate of the void and plate phases with its layers parallel to the x_2-axis suffices to generate all values of $\boldsymbol{\epsilon}_V$ on the boundary of $D(f, \boldsymbol{\sigma}_*)$. The effective elasticity matrix $\mathbf{C}_* = \mathbf{S}_*^{-1}$ of such a laminate has two zero eigenvalues. It is a bimode material in the sense of having two easy modes of deformation. A plane of average strains (rather than a single average strain) is associated with a given average stress $\boldsymbol{\sigma}_*$. It is easy to check that the associated plane of values of $\boldsymbol{\epsilon}_V$ is

$$\boldsymbol{\epsilon}_V : \boldsymbol{\sigma}_* = \frac{1}{(1-f)E}, \tag{2.19}$$

which, according to (2.6), is the boundary of $D(f, \boldsymbol{\sigma}_*)$.

When $\sigma < 0$ it is known that the energy bound is attained by rank-2 laminates with $\theta_1 = 0$ and $\theta_2 = \pi$. Indeed, by choosing these values of θ_1 and θ_2 and

$$\mathbf{t}_1 \cdot \boldsymbol{\epsilon}_V \mathbf{t}_1 = -\mathbf{t}_2 \cdot \boldsymbol{\epsilon}_V \mathbf{t}_2 = -\frac{1-\sigma}{(1-f)E}, \quad c_1 = \frac{-\sigma}{1-\sigma}, \quad c_2 = \frac{1}{1-\sigma}, \tag{2.20}$$

we see that $\boldsymbol{\sigma}_*$ given by (2.9) is diagonal with elements σ and 1, and that the bound (2.6) is achieved. In contrast to the case when $\sigma > 0$, these are the only rank-2 laminates which attain the bound. The constraint (2.20) on $\boldsymbol{\epsilon}_V$ forces it to have the form

$$\boldsymbol{\epsilon}_V = \begin{pmatrix} -k & c \\ c & k \end{pmatrix}, \quad \text{where} \quad k = \frac{1-\sigma}{(1-f)E}, \tag{2.21}$$

which generates only one line on the boundary of $D(f,\boldsymbol{\sigma}_*)$ as c is varied. What saves us is the fact that there are many other laminate structures which come close to attaining the bound.

Suppose we want to find a multiple rank laminate for which the average stress $\boldsymbol{\epsilon}_V$ within the void phase is close to boundary of $D(f,\boldsymbol{\sigma}_*)$. We can express $\boldsymbol{\epsilon}_V$ in the form

$$\boldsymbol{\epsilon}_V = \begin{pmatrix} a-k-\delta & b \\ b & k+\delta-a\sigma \end{pmatrix}, \tag{2.22}$$

where δ is a small positive parameter, and a and b can have any real values.

Now it was established by Milton and Cherkaev (1995) (see also section 30.7 in Milton (2002)) that for any given tensor $\boldsymbol{\sigma}_*$ with negative determinant there exists a rank-4 laminate which is bimode and supports the stress $\boldsymbol{\sigma}_*$. The associated plane of values of $\boldsymbol{\epsilon}_V$ necessarily takes the form

$$\boldsymbol{\epsilon}_V : \boldsymbol{\sigma}_* = (1-\sigma)h, \tag{2.23}$$

where h is a fixed constant with $h > k$. This plane is parallel to the boundary of $D(f,\boldsymbol{\sigma}_*)$. Let us assume that δ is small enough so that $h > k+\delta$. Our aim is to show that, by laminating together (on a very large length scale) this bimode material with the second rank laminate achieving the energy bound, we can attain the desired value (2.22) of $\boldsymbol{\epsilon}_V$.

In general the set $R(f,\boldsymbol{\sigma}_*)$ has the property that

$$\boldsymbol{\epsilon}_*^C = p\boldsymbol{\epsilon}_*^A + (1-p)\boldsymbol{\epsilon}_*^B \in R(f,\boldsymbol{\sigma}_*) \tag{2.24}$$

for all $p \in (0,1)$, whenever $\boldsymbol{\epsilon}_*^A$ and $\boldsymbol{\epsilon}_*^B$ are such that

$$\boldsymbol{\epsilon}_*^A \in R(f,\boldsymbol{\sigma}_*), \quad \boldsymbol{\epsilon}_*^B \in R(f,\boldsymbol{\sigma}_*), \quad \text{and} \quad \mathbf{t}\cdot\boldsymbol{\epsilon}_*^A\mathbf{t} = \mathbf{t}\cdot\boldsymbol{\epsilon}_*^B\mathbf{t} \tag{2.25}$$

for some choice of $\mathbf{t} \neq 0$. The condition that $\boldsymbol{\epsilon}_*^A$ and $\boldsymbol{\epsilon}_*^B$ lie in $R(f,\boldsymbol{\sigma}_*)$ implies that there exist composites A and B of the two phases, with the void phase occupying the volume fraction f in each composite, such that if an average stress $\boldsymbol{\sigma}_*$ is applied to these composites, then the resulting average strains will be $\boldsymbol{\epsilon}_*^A$ and $\boldsymbol{\epsilon}_*^B$, respectively. The last condition in (2.25) ensures compatibility of the average strains. The associated composite C is a laminate of the two composites A and B, laminated (on a length scale much larger than the microstructure of composites A and B) in proportions p and $1-p$, respectively, with its layers being parallel to $\mathbf{t}$. When an average stress $\boldsymbol{\sigma}_*^C$ is applied to this laminate, the elasticity equations are solved with the stress field within composites A and B having the same average value of $\boldsymbol{\sigma}_*$ resulting in an average strain of $\boldsymbol{\epsilon}_*^C$ in composite C.

When mapped to the set $D(f,\boldsymbol{\sigma}_*)$ (which is just $R(f,\boldsymbol{\sigma}_*)$ shifted and rescaled) this property implies

$$\boldsymbol{\epsilon}_V^C = p\boldsymbol{\epsilon}_V^A + (1-p)\boldsymbol{\epsilon}_V^B \in D(f,\boldsymbol{\sigma}_*) \tag{2.26}$$

for all $p \in (0,1)$, whenever $\boldsymbol{\epsilon}_V^A$ and $\boldsymbol{\epsilon}_V^B$ are such that

$$\boldsymbol{\epsilon}_V^A \in D(f,\boldsymbol{\sigma}_*), \quad \boldsymbol{\epsilon}_V^B \in D(f,\boldsymbol{\sigma}_*), \quad \text{and} \quad \mathbf{t}\cdot\boldsymbol{\epsilon}_V^A\mathbf{t} = \mathbf{t}\cdot\boldsymbol{\epsilon}_V^B\mathbf{t} \tag{2.27}$$

for some choice of $\mathbf{t} \neq 0$. We want to find $\boldsymbol{\epsilon}_V = \boldsymbol{\epsilon}_V^A$ of the form (2.21) and $\boldsymbol{\epsilon}_V = \boldsymbol{\epsilon}_V^B$ satisfying (2.23), such that $\boldsymbol{\epsilon}_V^C$ matches $\boldsymbol{\epsilon}_V$ in (2.22). First note that if the last

condition in (2.27) holds, then necessarily

$$0 = \mathbf{t} \cdot (\boldsymbol{\epsilon}_V^C - \boldsymbol{\epsilon}_V^A)\mathbf{t} = \mathbf{t} \cdot \begin{pmatrix} \delta - a & c - b \\ c - b & a\sigma - \delta \end{pmatrix} \mathbf{t}. \tag{2.28}$$

Letting θ denote the angle of the layers in the final lamination, so that $\mathbf{t} = (\sin\theta, \cos\theta)$, this last condition becomes

$$(\delta - a)(\tan\theta)^2 + 2(c - b)\tan\theta + a\sigma - \delta = 0. \tag{2.29}$$

We are free to select any value of θ such that $\tan\theta$ is nonzero and finite, and we set

$$c = b + \frac{(a - \delta)(\tan\theta)^2 + \delta - a\sigma}{2\tan\theta} \tag{2.30}$$

so that (2.29) is satisfied. Having found c and $\mathbf{t}$, we choose

$$\boldsymbol{\epsilon}_V^B = \frac{(\boldsymbol{\epsilon}_V^C - p\boldsymbol{\epsilon}_V^A)}{(1 - p)} \quad \text{with} \quad p = \frac{h - k - \delta}{h - k} \tag{2.31}$$

so that (2.26) holds, so that $\boldsymbol{\epsilon}_V = \boldsymbol{\epsilon}_V^B$ satisfies (2.23), and so that the last condition in (2.27) holds. The average strain in the void phase in the resulting composite then matches (2.22).

So far, for each positive, zero, or negative value of σ, we have found composites realizing values of $\boldsymbol{\epsilon}_V$ such that (2.6) is satisfied, or almost satisfied, as an equality. It remains to find composites realizing the remaining values of $\boldsymbol{\epsilon}_V$ satisfying the inequality (2.6). These composites have average strain in the void phase matching any $\boldsymbol{\epsilon}_V$ in the interior of $D(f, \boldsymbol{\sigma}_*)$. Given any α between 1 and $1/f$, consider the set of composites we have found that have values of $\boldsymbol{\epsilon}_V$ on or close to the boundary of $D(\alpha f, \boldsymbol{\sigma}_*)$. These have a volume fraction αf of the void phase, and are such that

$$\boldsymbol{\epsilon}_V : \boldsymbol{\sigma}_* \approx \frac{(1 + |\sigma|)^2}{(1 - \alpha f)E}. \tag{2.32}$$

Now we can reduce the volume fraction of the void phase down to f in each of these composites by inserting islands of the plate phase into the void spaces. Inside these islands both the stress and the strain will be zero. The stress field and hence the average stress field $\boldsymbol{\sigma}_*$ will remain unchanged. The displacement field $\mathbf{u}(\mathbf{x})$ in the original part of the plate phase (excluding the new islands) will also remain invariant. Since $\mathbf{u}(\mathbf{x})/\|\mathbf{x}\|$ approaches $\boldsymbol{\epsilon}_*\mathbf{x}/\|\mathbf{x}\|$ as $\|\mathbf{x}\| \to \infty$, it follows that the average strain will remain unchanged when we insert the islands. The average strain $\boldsymbol{\epsilon}_V$ in the void phase, given by (2.4), will be multiplied by the factor α because the volume fraction of void is reduced by the factor α. In this way we obtain composites, with the void phase occupying the volume fraction f, which realize or come close to realizing any $\boldsymbol{\epsilon}_V$ satisfying

$$\boldsymbol{\epsilon}_V : \boldsymbol{\sigma}_* = \alpha \frac{(1 + |\sigma|)^2}{(1 - \alpha f)E}. \tag{2.33}$$

As α is varied from 1 to $1/f$, this plane of values of $\boldsymbol{\epsilon}_V$ sweeps across the entire range allowed by (2.6).

There is one technical point which remains. That is, the results seem very sensitive to perturbations. In particular, a small change in the structure of a bimode material

may cause the zero eigenvalues of its effective elasticity matrix to become nonzero, and the associated set of values of $\boldsymbol{\epsilon}_V$, for a given applied stress $\boldsymbol{\sigma}_*$, changes from a plane of values to just a single value. Also, to rigorously justify our analysis, we should really take the void phase to have a nonzero Young's modulus E_V (so that it is not void) and then take the limit as $E_V \to 0$. But for any $E_V > 0$ the effective elasticity matrix will have nonzero eigenvalues, and so there will be just a single value of $\boldsymbol{\epsilon}_V$ for a given $\boldsymbol{\sigma}_*$ in any unimode or bimode material. It looks like our analysis falls apart.

However, the key observation to make is that while these perturbations make large changes to the effective compliance matrix $\mathbf{S}_*$, they make small changes to the effective elasticity matrix $\mathbf{C}_*$. For a given fixed applied strain $\boldsymbol{\epsilon}_*$ let us consider the range $S(f, \boldsymbol{\epsilon}_*)$ of values that the average stress $\boldsymbol{\sigma}_*$ takes as the microgeometry is varied over all possible configurations that have a fixed volume fraction f of the void phase. When the structure of a material is slightly perturbed, the value of $\boldsymbol{\sigma}_*$ will not change by much. Characterizing all the sets $S(f, \boldsymbol{\epsilon}_*)$ and characterizing all the sets $R(f, \boldsymbol{\sigma}_*)$ are clearly equivalent problems. The energy bounds imply

$$\boldsymbol{\epsilon}_* : \boldsymbol{\sigma}_* \geq \boldsymbol{\sigma}_* : \boldsymbol{\mathcal{S}}\boldsymbol{\sigma}_* + \frac{f(|\lambda_1| + |\lambda_2|)^2}{(1-f)E}, \tag{2.34}$$

where λ_1 and λ_2 are the eigenvalues of $\boldsymbol{\sigma}_*$ and $\boldsymbol{\mathcal{S}}$ is the fourth order compliance tensor of the plate (represented by the matrix $\mathbf{S}$ given by (2.2)). The set of $\boldsymbol{\sigma}_*$ consistent with these bounds for a given value of $\boldsymbol{\epsilon}_*$ does not have a simple geometrical interpretation in the space with σ_{11}^*, σ_{22}^*, and σ_{12}^* as coordinates, although its intersection of its boundary with the region $\det \boldsymbol{\sigma}_* \geq 0$ is an ellipsoidal surface. To find a microstructure associated with a given $\boldsymbol{\sigma}_*$ in this set, we use the same construction as before, that is, we look for the microstructure associated with $\boldsymbol{\epsilon}_*$ when the average stress $\boldsymbol{\sigma}_*$ is applied. Now, however, $\boldsymbol{\sigma}_*$ will be only slightly perturbed if we make E_V very small and positive. Since the structure of the material associated with a given $\boldsymbol{\sigma}_*$ changes continuously as $\boldsymbol{\sigma}_*$ is varied in the regions $\det \boldsymbol{\sigma}_* > 0$ and $\det \boldsymbol{\sigma}_* < 0$, we can realize any $\boldsymbol{\sigma}_*$ satisfying the strict inequality (2.34) for sufficiently small values of E_V, provided that $\det \boldsymbol{\sigma}_* \neq 0$. If $\det \boldsymbol{\sigma}_* = 0$ and E_V is fixed but very small, then we consider the microstructures A and B associated with $\boldsymbol{\sigma}_*^A$ and $\boldsymbol{\sigma}_*^B$, where

$$\boldsymbol{\sigma}_*^A = \boldsymbol{\sigma}_* + \delta \mathbf{t} \otimes \mathbf{t}, \qquad \boldsymbol{\sigma}_*^B = \boldsymbol{\sigma}_* - \delta \mathbf{t} \otimes \mathbf{t}, \tag{2.35}$$

where δ is small and the unit vector $\mathbf{t}$ is chosen so that $\det \boldsymbol{\sigma}_*^A \neq 0$. Then $\det \boldsymbol{\sigma}_*^B = -\det \boldsymbol{\sigma}_*^A$ is also nonzero. By laminating together the microstructures A and B in equal proportions with the layer interfaces being oriented parallel to $\mathbf{t}$, we obtain a composite having the average stress $\boldsymbol{\sigma}_*$.

3. A necessary condition for optimality for a periodic array of holes. Although we have completely characterized the range $R(f, \boldsymbol{\sigma}_*)$ and identified optimal (or almost optimal) microgeometries, the solution is unsatisfactory from a practical viewpoint. Multiple rank laminates are difficult to construct, and when one phase is void we expect that buckling of the layers and contact between adjacent layers could be a serious problem. Instead of allowing all possible microgeometries, it makes sense to consider a restricted class of more realistic microgeometries, such as periodic composites comprised of a regular array of holes with one simply connected hole per unit cell. We focus on this problem since a periodic composite with only one hole per unit cell should be much easier to manufacture than a composite with multiple

(or perhaps infinitely many) holes per unit cell. The holes occupy a (possibly large) volume fraction f in the plate. We let $\widetilde{R}(f, \boldsymbol{\sigma}_*)$ denote the range of values that the average strain $\boldsymbol{\epsilon}_*$ takes as the boundary Γ of the hole in the unit cell is varied, while keeping f fixed. (The range $\widetilde{R}(f, \boldsymbol{\sigma}_*)$ will depend on our choice of primitive vectors of the unit cell.)

A first step towards determining $\widetilde{R}(f, \boldsymbol{\sigma}_*)$ is to find its convex hull. This is obtained from its Legendre transform

$$g(f, \boldsymbol{\sigma}_*^0, \boldsymbol{\sigma}_*) = \min_{\boldsymbol{\epsilon}_* \in \widetilde{R}(f, \boldsymbol{\sigma}_*)} \boldsymbol{\sigma}_*^0 : \boldsymbol{\epsilon}_*. \tag{3.1}$$

Knowing $g(f, \boldsymbol{\sigma}_*^0, \boldsymbol{\sigma}_*)$ for given values of f, $\boldsymbol{\sigma}_*^0$, and $\boldsymbol{\sigma}_*$ allows one to compute a tangent plane to the set $\widetilde{R}(f, \boldsymbol{\sigma}_*)$ (with the tangent plane having normal $\boldsymbol{\sigma}_*^0$). By varying $\boldsymbol{\sigma}_*^0$ and taking the envelope of the resulting family of tangent lines, one recovers the convex hull of $\widetilde{R}(f, \boldsymbol{\sigma}_*)$. When $\boldsymbol{\sigma}_*^0 = \boldsymbol{\sigma}_*$, the quantity $\boldsymbol{\sigma}_*^0 : \boldsymbol{\epsilon}_*/2$ is simply the elastic energy stored in the composite when it is subject to the average stress $\boldsymbol{\sigma}_*$, and so $g(f, \boldsymbol{\sigma}_*, \boldsymbol{\sigma}_*)/2$ represents a sharp bound on the elastic energy of a periodic array of holes, with the void phase occupying the volume fraction f. The bound (2.34) clearly implies

$$g(f, \boldsymbol{\sigma}_*, \boldsymbol{\sigma}_*) \geq \boldsymbol{\sigma}_* : \boldsymbol{\mathcal{S}} \boldsymbol{\sigma}_* + \frac{f(|\lambda_1| + |\lambda_2|)^2}{(1-f)E}, \tag{3.2}$$

where λ_1 and λ_2 are the eigenvalues of $\boldsymbol{\sigma}_*$. This inequality becomes an equality only when $\det \boldsymbol{\sigma}_* \geq 0$, and the hole shapes attaining the minimum value of the elastic energy are those found by Vigdergauz (1986, 1994, 1996, 1999) (see also Grabovsky and Kohn (1995)). These holes approach ellipses as the volume fraction f approaches zero. When $\det \boldsymbol{\sigma}_*$ is negative, Allaire and Aubry (1999) have proved that there is a definite gap between the left- and right-hand sides of the above equation. The assumption of only one hole per unit cell represents a significant restriction on the set of microstructures being considered. Cherkaev et al. (1998), following preliminary work of Cherkaev and Vigdergauz (1986), have shown that in the limit as $f \to 0$ the hole minimizing the elastic energy is almost rectangular in shape when $\det \boldsymbol{\sigma}_* < 0$.

Here we consider what happens when the boundary Γ of the hole is varied, while keeping f fixed, so as to minimize the bilinear form

$$\boldsymbol{\sigma}_*^0 : \boldsymbol{\mathcal{S}}_* \boldsymbol{\sigma}_*, \tag{3.3}$$

where $\boldsymbol{\sigma}_*^0$ and $\boldsymbol{\sigma}_*$ are fixed tensors. The value of the minimum is of course $g(f, \boldsymbol{\sigma}_*^0, \boldsymbol{\sigma}_*)$, the function we seek to find. It is conceivable that the minimum value is never achieved. The boundary of the hole could become increasingly convoluted as the minimum is approached. Alternatively, the boundary of the optimal shaped hole could touch itself along certain intervals, corresponding to slits or infinitesimally thin bridges. For the moment let us ignore these possibilities and assume that the minimum is achieved by a simply connected hole, with a boundary that does not touch itself and that additionally does not touch the boundary of the unit cell.

We will show that if the shape of hole is such that this bilinear form is minimized, then the following necessary condition for optimality holds:

$$\sigma_{tt}^0(\mathbf{x}) \sigma_{tt}(\mathbf{x}) = \text{constant on all smooth portions of } \Gamma. \tag{3.4}$$

Here, $\sigma^0_{tt}(\mathbf{x})$ and $\sigma_{tt}(\mathbf{x})$ are the tangential components of the periodic stress tensor fields $\boldsymbol{\sigma}^0(\mathbf{x})$ and $\boldsymbol{\sigma}(\mathbf{x})$ that are generated when average stresses $\boldsymbol{\sigma}^0_*$ and $\boldsymbol{\sigma}_*$, respectively, are applied to the composite. Thus the stress tensor field $\boldsymbol{\sigma}^0(\mathbf{x})$, defined to be zero within each hole, has average value

$$\langle \boldsymbol{\sigma}^0 \rangle = \boldsymbol{\sigma}^0_* \tag{3.5}$$

and satisfies the equilibrium equation

$$\nabla \cdot \boldsymbol{\sigma}^0(\mathbf{x}) = 0 \tag{3.6}$$

and the boundary condition

$$\boldsymbol{\sigma}^0(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x}) = 0, \quad \mathbf{x} \in \Gamma, \tag{3.7}$$

where $\mathbf{n}(\mathbf{x})$ is the outward normal to the surface Γ at the point $\mathbf{x}$. Additionally, this stress is related to the associated strain tensor field

$$\boldsymbol{\epsilon}^0(\mathbf{x}) = \frac{1}{2}(\nabla \mathbf{u}^0(\mathbf{x}) + (\nabla \mathbf{u}^0(\mathbf{x}))^T) \tag{3.8}$$

through the constitutive law

$$\boldsymbol{\epsilon}^0(\mathbf{x}) = \boldsymbol{\mathcal{S}} \boldsymbol{\sigma}^0(\mathbf{x}). \tag{3.9}$$

By deleting the superscript 0 in these formulae, one obtains the equations satisfied by the stress tensor field $\boldsymbol{\sigma}(\mathbf{x})$. By definition, the effective compliance tensor $\boldsymbol{\mathcal{S}}_*$ governs the relations

$$\boldsymbol{\epsilon}^0_* = \boldsymbol{\mathcal{S}}_* \boldsymbol{\sigma}^0_*, \qquad \boldsymbol{\epsilon}_* = \boldsymbol{\mathcal{S}}_* \boldsymbol{\sigma}_* \tag{3.10}$$

between the average strain, $\boldsymbol{\epsilon}^0_*$ or $\boldsymbol{\epsilon}_*$, and the average stress, $\boldsymbol{\sigma}^0_*$ or $\boldsymbol{\sigma}_*$. Since the local strains $\boldsymbol{\epsilon}^0(\mathbf{x})$ and $\boldsymbol{\epsilon}(\mathbf{x})$ are undefined within each hole, the statement that the average strains are $\boldsymbol{\epsilon}^0_*$ and $\boldsymbol{\epsilon}_*$ holds in the sense that

$$\mathbf{u}^0(\mathbf{x}) - \boldsymbol{\epsilon}^0_* \mathbf{x} \text{ and } \mathbf{u}(\mathbf{x}) - \boldsymbol{\epsilon}_* \mathbf{x} \text{ are periodic in } \mathbf{x}. \tag{3.11}$$

Now let Ω denote the region occupied by the plate within the unit cell, excluding the hole. Then from (3.10) we have the useful identity

$$\begin{aligned}
\int_\Omega \boldsymbol{\sigma}^0(\mathbf{x}) : \boldsymbol{\mathcal{S}} \boldsymbol{\sigma}(\mathbf{x}) d\mathbf{x} &= \int_\Omega \boldsymbol{\sigma}^0(\mathbf{x}) : \boldsymbol{\epsilon}(\mathbf{x}) d\mathbf{x} \\
&= \int_{\text{boundary of the cell}} \mathbf{n} \cdot \boldsymbol{\sigma}^0(\mathbf{x}) \mathbf{u}(\mathbf{x}) dl + \int_\Gamma \mathbf{n} \cdot \boldsymbol{\sigma}^0(\mathbf{x}) \mathbf{u}(\mathbf{x}) dl \\
&= \int_{\text{boundary of the cell}} \mathbf{n} \cdot \boldsymbol{\sigma}^0(\mathbf{x}) [\boldsymbol{\epsilon}_* \mathbf{x}] dl \\
&= \int_\Omega \boldsymbol{\sigma}^0(\mathbf{x}) : \boldsymbol{\epsilon}_* d\mathbf{x} \\
&= \boldsymbol{\sigma}^0_* : \boldsymbol{\mathcal{S}}_* \boldsymbol{\sigma}_*,
\end{aligned} \tag{3.12}$$

where we have used the fact that $\mathbf{n} \cdot \boldsymbol{\sigma}^0(\mathbf{x})$ vanishes on the boundary Γ and the fact that on opposite sides of the unit cell $\mathbf{n} \cdot \boldsymbol{\sigma}^0(\mathbf{x})$ takes opposite values, while $\mathbf{u}(\mathbf{x}) - \boldsymbol{\epsilon}_* \mathbf{x}$ takes the same value, implying that their scalar product integrates to zero.

To establish (3.4) we introduce a Lagrange multiplier c and look for the stationary points of

$$\boldsymbol{\sigma}_*^0 : \boldsymbol{\mathcal{S}}_* \boldsymbol{\sigma}_* + cf \tag{3.13}$$

as the boundary of the hole is varied. For simplicity, let us assume that the unit cell is square with sides of unit length.

Let $\lambda\gamma(\mathbf{x})$ be the perturbation of the interface in the normal direction; that is, the new boundary $\Gamma(\lambda)$ consists of points

$$\mathbf{y} = \mathbf{x} + \lambda\gamma(\mathbf{x})\mathbf{n}(\mathbf{x}), \quad \mathbf{x} \in \Gamma, \tag{3.14}$$

where $\mathbf{n}(\mathbf{x})$ is the outward normal to the original surface Γ at the point $\mathbf{x}$.

Our aim is to evaluate

$$\frac{d}{d\lambda}(\boldsymbol{\sigma}_*^0 : \boldsymbol{\mathcal{S}}_*(\lambda)\boldsymbol{\sigma}_* + cf(\lambda))\bigg|_{\lambda=0}. \tag{3.15}$$

The derivative of $f(\lambda)$ is simply

$$\frac{d}{d\lambda} f(\lambda)\bigg|_{\lambda=0} = \int_\Gamma \gamma(\mathbf{x}) dl. \tag{3.16}$$

Also from (3.12) we see that the remaining derivative in (3.15) splits into the sum of three terms:

$$\begin{aligned}\frac{d}{d\lambda}(\boldsymbol{\sigma}_*^0 : \boldsymbol{\mathcal{S}}_*(\lambda)\boldsymbol{\sigma}_*)\bigg|_{\lambda=0} &= \int_\Omega \frac{d\boldsymbol{\sigma}^0(\mathbf{x})}{d\lambda}\bigg|_{\lambda=0} : \boldsymbol{\mathcal{S}}\boldsymbol{\sigma}(\mathbf{x}) d\mathbf{x} + \int_\Omega \boldsymbol{\sigma}^0(\mathbf{x}) : \boldsymbol{\mathcal{S}} \frac{d\boldsymbol{\sigma}(\mathbf{x})}{d\lambda}\bigg|_{\lambda=0} d\mathbf{x} \\ &\quad + \int_\Gamma \boldsymbol{\sigma}^0(\mathbf{x}) : \boldsymbol{\mathcal{S}}\boldsymbol{\sigma}(\mathbf{x})\gamma(\mathbf{x}) dl,\end{aligned} \tag{3.17}$$

where the last term arises because the domain of integration $\Omega(\lambda)$ depends on λ. The first term can be equated with the derivative that we seek to find through a calculation similar to that given in (3.12):

$$\begin{aligned}\int_\Omega \frac{d\boldsymbol{\sigma}^0(\mathbf{x})}{d\lambda}\bigg|_{\lambda=0} : \boldsymbol{\mathcal{S}}\boldsymbol{\sigma}(\mathbf{x}) d\mathbf{x} &= \int_\Omega \frac{d\boldsymbol{\epsilon}^0(\mathbf{x})}{d\lambda}\bigg|_{\lambda=0} : \boldsymbol{\sigma}(\mathbf{x}) d\mathbf{x} \\ &= \int_{\text{boundary of the cell}} \mathbf{n} \cdot \boldsymbol{\sigma}(\mathbf{x}) \frac{d\mathbf{u}^0(\mathbf{x})}{d\lambda} dl + \int_\Gamma \mathbf{n} \cdot \boldsymbol{\sigma}(\mathbf{x}) \frac{d\mathbf{u}^0(\mathbf{x})}{d\lambda} dl \\ &= \int_{\text{boundary of the cell}} \mathbf{n} \cdot \boldsymbol{\sigma}(\mathbf{x}) \left[\frac{d\boldsymbol{\epsilon}_*^0}{d\lambda}\mathbf{x}\right] dl \\ &= \int_\Omega \boldsymbol{\sigma}(\mathbf{x}) \frac{d\boldsymbol{\epsilon}_*^0}{d\lambda} d\mathbf{x} \\ &= \frac{d}{d\lambda}(\boldsymbol{\sigma}_*^0 : \boldsymbol{\mathcal{S}}_*(\lambda)\boldsymbol{\sigma}_*)\bigg|_{\lambda=0}.\end{aligned} \tag{3.18}$$

Similarly, the second term can also be equated with this derivative, and so we deduce that

$$\frac{d}{d\lambda}(\boldsymbol{\sigma}_*^0 : \boldsymbol{\mathcal{S}}_*(\lambda)\boldsymbol{\sigma}_* + cf(\lambda))\bigg|_{\lambda=0} = \int_\Gamma (-\boldsymbol{\sigma}^0(\mathbf{x}) : \boldsymbol{\mathcal{S}}\boldsymbol{\sigma}(\mathbf{x}) + c)\gamma(\mathbf{x}) dl. \tag{3.19}$$

For this to be zero for all choices of $\gamma(\mathbf{x})$ we must have

$$\boldsymbol{\sigma}^0(\mathbf{x}) : \boldsymbol{\mathcal{S}}\boldsymbol{\sigma}(\mathbf{x}) = c \quad \text{on all smooth portions of } \Gamma. \tag{3.20}$$

But the only nonzero components of $\boldsymbol{\sigma}^0(\mathbf{x})$ and $\boldsymbol{\sigma}(\mathbf{x})$ on the boundary of the hole are the tangential components σ^0_{tt} and σ_{tt}. It follows that

$$\boldsymbol{\sigma}^0(\mathbf{x}) : \boldsymbol{\mathcal{S}}\boldsymbol{\sigma}(\mathbf{x}) = \sigma^0_{tt}[(\mathbf{t}\otimes\mathbf{t}) : \boldsymbol{\mathcal{S}}(\mathbf{t}\otimes\mathbf{t})]\sigma_{tt}, \tag{3.21}$$

where $\mathbf{t}$ is the unit vector tangential to the hole boundary. Finally, since $\boldsymbol{\mathcal{S}}$ is rotationally invariant, $(\mathbf{t}\otimes\mathbf{t}) : \boldsymbol{\mathcal{S}}(\mathbf{t}\otimes\mathbf{t})$ is a constant independent of the direction of $\mathbf{t}$. In conclusion, we arrive at (3.4).

4. Possible stress-strain pairs for a dilute array of holes. Here we consider a periodic array of holes in a plate having isotropic compliance matrix $\mathbf{S}(E,\nu)$. It is assumed there is only one hole per unit cell, occupying a very small volume fraction f of the unit cell. Additionally the hole is assumed to be well separated from the holes in adjoining cells, so that interaction effects can be neglected. Let $\widetilde{D}(\boldsymbol{\sigma}_*)$ denote the range of values that $\boldsymbol{\epsilon}_V$ takes as the hole shape is varied, in the limit $f \to 0$. Our aim is to find $\widetilde{D}(\boldsymbol{\sigma}_*)$ and to identify the optimal hole shapes that correspond to values of $\boldsymbol{\epsilon}_V$ on the boundary of $\widetilde{D}(\boldsymbol{\sigma}_*)$. In some sense these optimal holes have the greatest effect on the average strain when a given (small) volume fraction of them is inserted into a plate which is subject to the applied loading $\boldsymbol{\sigma}_*$.

To the first order in f the effective compliance matrix is given by

$$\mathbf{S}_* = \mathbf{S} + f\mathbf{E} + \mathcal{O}(f^2), \tag{4.1}$$

where the compliance polarizability matrix $\mathbf{E}$ is obtained by solving the planar elasticity equations for a single hole in an infinite plane, subject to uniform applied stresses. Associated with the matrix $\mathbf{E}$ is a fourth order compliance polarizability tensor $\boldsymbol{\mathcal{E}}$. (The tensor $f\boldsymbol{\mathcal{E}}$ has been called the hole compliance tensor or H-tensor by Kachanov (1993) and Shafiro and Kachanov (1999), and the inverse of $\mathbf{E}$ has been called the Pólya–Szegö matrix by Movchan and Serkov (1997).)

To a first approximation, the average strain $\boldsymbol{\epsilon}_V$ in the void phase, as defined by (2.4), is given by

$$\boldsymbol{\epsilon}_V \approx \mathbf{E}\boldsymbol{\sigma}_*. \tag{4.2}$$

Thus the problem of determining the set $\widetilde{D}(\boldsymbol{\sigma}_*)$ reduces to finding the range of values that $\mathbf{E}\boldsymbol{\sigma}_*$ takes as the hole shape is varied.

Now in the periodic array, $\mathbf{S}_* - \mathbf{S} \approx f\mathbf{E}$ should remain virtually unchanged when we insert into each hole an island of plate material that is connected to the surrounding plate by a very thin bridge of plate material (to ensure that the hole in the unit cell remains simply connected). This reduces the volume fraction of void f by some factor, and $\mathbf{E}$ must accordingly increase by the same factor. We deduce that

$$\lambda\boldsymbol{\epsilon}_V \in \widetilde{D}(\boldsymbol{\sigma}_*) \quad \text{for all } \lambda \geq 1, \quad \text{provided } \boldsymbol{\epsilon}_V \in \widetilde{D}(\boldsymbol{\sigma}_*). \tag{4.3}$$

This obviously implies that the set $\widetilde{D}(\boldsymbol{\sigma}_*)$ is unbounded in some directions.

Our goal is to numerically compute this set and to identify the hole shapes which correspond to values of $\boldsymbol{\epsilon}_V$ on the boundary of $\widetilde{D}(\boldsymbol{\sigma}_*)$. The approach we take is similar to one successfully used by Cherkaev et al. (1998) to find hole shapes which minimize the elastic energy under a given loading $\boldsymbol{\sigma}_*$. For values of $\boldsymbol{\sigma}_*$ with positive determinant they recovered the known result that the optimal hole is an ellipse,

while for values of $\boldsymbol{\sigma}_*$ with negative determinant they found that the optimal hole is almost rectangular in shape. The first step in these computations is to determine the compliance polarizability matrix $\mathbf{E}$ for a given shaped hole.

4.1. Compliance polarizability matrices for a dilute array of holes. For an arbitrary shaped hole Movchan and Serkov (1997) (see also Serkov (1998)) show how the calculation of $\mathbf{E}$ can be reduced to solving a system of linear equations. In this section we briefly summarize the main points of their analysis and also show how it can be used to obtain the elastic fields in the vicinity of each hole.

Let the shape of the hole be specified by the image of the exterior of the unit circle under the conformal map

$$z = \omega(\xi), \tag{4.4}$$

where $z = x_1 + ix_2$, and the complex variable ξ is associated with the unit circle. Without loss of generality we can assume that $\omega(\xi) \approx \xi$ when $|\xi|$ is very large. The complex valued function ω can be expanded in a Laurent series, and, after truncation, one has

$$z = \omega(\xi) = \xi + \sum_{n=1}^{N} c_{-n}\xi^{-n}, \tag{4.5}$$

with the c_i being complex coefficients that determine the shape of the hole. We consider the response to an applied stress (representing the stress at a large distance from the hole)

$$\boldsymbol{\sigma}_* = \begin{pmatrix} \sigma_{11}^* & \sigma_{12}^* \\ \sigma_{12}^* & \sigma_{22}^* \end{pmatrix} \tag{4.6}$$

as the loading coefficients σ_{11}^*, σ_{22}^*, and σ_{12}^* are varied. The solution to this planar elasticity problem is naturally expressed in terms of the Kolosov–Muskhelishvili complex potentials ϕ and ψ (Muskhelishvili (1953)). For plane stress they determine the two-dimensional displacement field $\mathbf{u}(\mathbf{x})$ through the relation

$$u_1 + iu_2 = \frac{[(3-\nu)\phi(z)/(1+\nu) - z\overline{\phi'(z)} - \overline{\psi(z)}]}{(2\mu)}, \tag{4.7}$$

in which ν is Poisson's ratio, μ is the shear modulus, and the overline denotes complex conjugation. The associated stress field $\boldsymbol{\sigma}(\mathbf{x})$ has components σ_{11}, σ_{22}, and σ_{12} given by

$$\sigma_{11} + \sigma_{22} = 4\,\mathrm{Re}[\phi'(z)], \quad \sigma_{22} - \sigma_{11} + 2i\sigma_{12} = 2[\overline{z}\phi''(z) + \psi'(z)]. \tag{4.8}$$

The potential $\phi(\xi)$ can be shown (see the formulae between (3.11) and (3.13) in Movchan and Serkov (1997)) to satisfy the integral equation

$$\phi(\xi) - \frac{1}{2\pi i}\int_{|t|=1} \frac{\omega(t)\overline{\phi'(t)}}{\overline{\omega'(t)}(t-\xi)}dt = \alpha\xi - \overline{\gamma}\xi^{-1}, \tag{4.9}$$

where the real coefficient α and the complex coefficient γ are determined from the loading coefficients:

$$\alpha = \frac{(\sigma_{11}^* + \sigma_{22}^*)}{4}, \quad \gamma = \frac{(\sigma_{22}^* - \sigma_{11}^* + 2i\sigma_{12}^*)}{2}. \tag{4.10}$$

Now the complex potential $\phi(\xi)$ can be expanded in a Laurent series in powers of ξ, and the coefficients of this series must depend linearly on the applied stress, that is, linearly on α and linearly on the real and imaginary parts of γ. Also when ξ is large, $\phi(\xi)$ approaches $\alpha\xi$. Accordingly, the potential $\phi(\xi)$ has the representation

$$\phi(\xi) = \alpha\xi - \overline{\gamma}\xi^{-1} - \alpha\sum_{n=1}^{\infty} a_n^{\alpha}\xi^{-n} - \mathrm{Re}(\gamma)\sum_{n=1}^{\infty} a_n^{\gamma}\xi^{-n} - \mathrm{Im}(\gamma)\sum_{n=1}^{\infty} a_n^{\tau}\xi^{-n}, \tag{4.11}$$

where the minus signs and the additional term $-\overline{\gamma}\xi^{-1}$ are introduced to simplify subsequent equations. This serves to define the three sets of coefficients a_n^j, $j = \alpha, \gamma, \tau$, each set being associated with a different loading. By substituting (4.11) back into (4.9) and using the residue theorem to evaluate the contour integrals, one finds that

$$a_n^{\alpha} = a_n^{\gamma} = a_n^{\tau} = 0 \quad \text{for all } n > N, \tag{4.12}$$

and that the remaining coefficients satisfy the system of linear equations

$$\begin{aligned}
a_m^{\alpha} - \sum_{k=1}^{N-m-1} \rho_{N-m-k-1} k \overline{a}_k^{\alpha} &= \rho_{N-m}, \qquad a_m^{\alpha} \in \mathbb{C}, \\
a_m^{\gamma} - \sum_{k=1}^{N-m-1} \rho_{N-m-k-1} k \overline{a}_k^{\gamma} &= \rho_{N-m-2}, \qquad a_m^{\gamma} \in \mathbb{C}, \\
a_m^{\tau} - \sum_{k=1}^{N-m-1} \rho_{N-m-k-1} k \overline{a}_k^{\tau} &= \rho_{N-m-2} i, \qquad a_m^{\tau} \in \mathbb{C},
\end{aligned} \tag{4.13}$$

for $m = 1, 2, 3, \ldots, N$, where $\rho_k = 0$ for $k < 0$, $\rho_0 = c_{-N}$, and

$$\rho_k = \frac{1}{k!}\frac{d^k}{d\xi^k}\left[\frac{\xi^{N+1} + \sum_{n=1}^{N} c_{-n}\xi^{N-n}}{1 - \sum_{n=1}^{N} n\overline{c}_{-n}\xi^{n+1}}\right]\Bigg|_{\xi=0} \tag{4.14}$$

for $k > 0$.

This system can be written in the matrix form

$$\mathbf{X}^{(i)} - \mathbf{A}\overline{\mathbf{X}^{(i)}} = \mathbf{B}^{(i)}, \qquad i = \alpha, \gamma, \tau, \tag{4.15}$$

where

$$\mathbf{X}^{(\alpha)} = \begin{pmatrix} a_1^{\alpha} \\ a_2^{\alpha} \\ \vdots \\ a_{N-2}^{\alpha} \\ a_{N-1}^{\alpha} \\ a_N^{\alpha} \end{pmatrix}, \quad \mathbf{X}^{(\gamma)} = \begin{pmatrix} a_1^{\gamma} \\ a_2^{\gamma} \\ \vdots \\ a_{N-2}^{\gamma} \\ a_{N-1}^{\gamma} \\ a_N^{\gamma} \end{pmatrix}, \quad \mathbf{X}^{(\tau)} = \begin{pmatrix} a_1^{\tau} \\ a_2^{\tau} \\ \vdots \\ a_{N-2}^{\tau} \\ a_{N-1}^{\tau} \\ a_N^{\tau} \end{pmatrix},$$

$$\mathbf{B}^{(\alpha)} = \begin{pmatrix} \rho_{N-1} \\ \rho_{N-2} \\ \vdots \\ \rho_2 \\ \rho_1 \\ \rho_0 \end{pmatrix}, \quad \mathbf{B}^{(\gamma)} = \begin{pmatrix} \rho_{N-3} \\ \rho_{N-4} \\ \vdots \\ \rho_0 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{B}^{(\tau)} = i\mathbf{B}^{(\gamma)} \tag{4.16}$$

are N-dimensional vectors and

$$\mathbf{A} = \begin{pmatrix} \rho_{N-3} & 2\rho_{N-4} & \cdots & (N-2)\rho_0 & 0 & 0 \\ \rho_{N-4} & 2\rho_{N-5} & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \rho_1 & 2\rho_0 & \cdots & 0 & 0 & 0 \\ \rho_0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 \end{pmatrix} \tag{4.17}$$

is an $N \times N$ matrix. It helps to rewrite (4.15) in the equivalent form

$$(\mathbf{I} - \mathbf{A}\overline{\mathbf{A}})\mathbf{X}^{(i)} = \mathbf{B}^{(i)} + \mathbf{A}\overline{\mathbf{B}^{(i)}}, \qquad i = \alpha, \gamma, \tau, \tag{4.18}$$

which may be solved directly for the unknown vectors $\mathbf{X}^{(i)}$.

From the solution for the coefficients a_n^j, $j = \alpha, \gamma, \tau$, we can find the potential $\phi(z)$ for an arbitrary applied stress and hence determine the tangential stresses on the boundary of the hole:

$$\sigma_{tt} = \text{Trace}(\boldsymbol{\sigma}) = 4\,\text{Re}[\Phi(\mathbf{x})] \ \text{ on } \Gamma, \ \text{ where } \Phi(\mathbf{x}) = \frac{\phi'(\xi)}{\omega'(\xi)}. \tag{4.19}$$

This is useful because in order to check if the optimality condition (3.4) is satisfied (or approximately satisfied) we need to determine stress fields at the boundary of the hole.

Also, once the potential $\phi(\xi)$ has been determined, the potential $\psi(\xi)$ can be found from the relation

$$\psi(\xi) = \gamma\xi - \alpha\xi^{-1} + \frac{1}{2\pi i}\int_{|t|=1} \frac{\overline{\omega(t)}\phi'(t)}{\omega'(t)(t-\xi)}dt, \tag{4.20}$$

which directly follows from the formulae between (3.11) and (3.13) in Movchan and Serkov (1997). By substituting the series expansion (4.11) for $\phi(\xi)$ in (4.9) and using the residue theorem to evaluate the contour integrals, one finds that $\psi(\xi)$ is given by the series expansion

$$\psi(\xi) = \gamma\xi - \alpha\xi^{-1} - \alpha\sum_{n=1}^{\infty} a_{-n}^{\alpha}\xi^{-n} - \text{Re}(\gamma)\sum_{n=1}^{\infty} a_{-n}^{\gamma}\xi^{-n} - \text{Im}(\gamma)\sum_{n=1}^{\infty} a_{-n}^{\tau}\xi^{-n}, \tag{4.21}$$

where the coefficients a_{-n}^{α}, a_{-n}^{γ}, and a_{-n}^{τ} are obtained by setting $m = -n$ in the formulae (4.13). It is curious that the series expansion for the potential $\phi(\xi)$ terminates at $n = N$, whereas the series expansion for $\psi(\xi)$ has an infinite number of terms. Having obtained the potentials $\phi(\xi)$ and $\psi(\xi)$, the displacement field $\mathbf{u}(\mathbf{x})$ around the hole can be computed using (4.7).

The compliance polarizability matrix $\mathbf{E}$ is determined by the area of the hole and by the far field behavior of the potentials ϕ and ψ in the z-plane. This far field behavior in the z-plane is in turn determined first by the far field behavior of ϕ and ψ in the ξ plane, which to leading order is governed by the six coefficients a_1^j and a_{-1}^j with $j = \alpha, \gamma, \tau$, and second by the dependence of z on ξ when $|\xi|$ is large, which to leading order is governed by the coefficient c_{-1}. Explicit calculation shows that the compliance polarizability matrix is given by the formula

$$\mathbf{E} = \frac{-1}{E(1 - \sum_{n=1}^{\infty} n|c_{-n}|^2)} \begin{pmatrix} -2\Omega + \Sigma - \frac{1}{2}\Xi & 2\Omega - \frac{1}{2}\Xi & \Lambda - 2\Theta \\ 2\Omega - \frac{1}{2}\Xi & -2\Omega - \Sigma - \frac{1}{2}\Xi & \Lambda + 2\Theta \\ \Lambda - 2\Theta & \Lambda + 2\Theta & 2\Upsilon \end{pmatrix}, \tag{4.22}$$

where E is the Young's modulus and

$$\begin{aligned}
&\Omega = 1 + \mathrm{Re}(a_1^\gamma), \qquad \Sigma = 2\,\mathrm{Re}(c_{-1}) + \mathrm{Re}(a_1^\alpha) + \mathrm{Re}(a_{-1}^\gamma),\\
&\Xi = 1 + \mathrm{Re}(a_{-1}^\alpha), \qquad \Upsilon = -2 + 2\,\mathrm{Im}(a_1^\tau),\\
&\Theta = \sqrt{2}\,\mathrm{Im}(a_1^\gamma), \qquad \Lambda = \sqrt{2}\,\mathrm{Im}(c_{-1}) + \sqrt{2}\,\mathrm{Im}(a_1^\alpha).
\end{aligned} \tag{4.23}$$

The expression (4.22) for the compliance polarizability matrix $\mathbf{E}$ differs from the one given in formula (4.8) of Movchan and Serkov (1997) by the prefactor of

$$V = \pi\left(1 - \sum_{n=1}^{\infty} n|c_{-n}|^2\right), \tag{4.24}$$

which represents the area of a hole which is the image of the unit circle $|\xi| = 1$ under the conformal map (4.5). This additional factor is needed to normalize with respect to the area of the hole so that (4.1) gives the correct estimate for the effective compliance tensor of a dilute periodic array of these holes.

4.2. Explicit formulae for elliptical holes. For elliptical holes the formulae (4.22) and the whole algorithm can be simplified. An ellipse of arbitrary eccentricity and arbitrary orientation can be generated by a conformal map of the form

$$z = \omega(\xi) = \xi + c_{-1}\xi^{-1}, \quad c_{-1} \in \mathbb{C}, \quad |c_{-1}| < 1. \tag{4.25}$$

By selecting

$$c_{-1} = \frac{1-r}{1+r}e^{2i\beta}, \quad \beta \in [0,\pi], \quad r > 0, \tag{4.26}$$

the boundary of the unit disc $|\xi| = 1$ is mapped onto the boundary of an ellipse with semiaxis lengths a and $b = 2-a$ having a desired ratio $r = b/a$, and with axes oriented at desired angles of β and $\beta + \pi/2$.

The formulae for coefficients Ω, Σ, Ξ, Υ, Θ, and Λ can be reduced to the form

$$\begin{aligned}
&\Omega = 1, \qquad \Xi = 2\left(1 + \left(\frac{1-r}{1+r}\right)^2\right), \qquad \Upsilon = -2,\\
&\Sigma = 4\frac{1-r}{1+r}\cos 2\beta, \qquad \Lambda = 2\sqrt{2}\frac{1-r}{1+r}\sin 2\beta, \qquad \Theta = 0,
\end{aligned} \tag{4.27}$$

and consequently the compliance polarizability matrix is

$$\mathbf{E} = \frac{-1}{Er}\begin{pmatrix} -(1+r+r^2) + (1-r^2)\cos 2\beta & r & \frac{1}{\sqrt{2}}(1-r^2)\sin 2\beta \\ r & -(1+r+r^2) - (1-r^2)\cos 2\beta & \frac{1}{\sqrt{2}}(1-r^2)\sin 2\beta \\ \frac{1}{\sqrt{2}}(1-r^2)\sin 2\beta & \frac{1}{\sqrt{2}}(1-r^2)\sin 2\beta & -\frac{1}{2}(1+r)^2 \end{pmatrix}. \tag{4.28}$$

Without loss of generality we can assume that $\boldsymbol{\sigma}_*$ has the form (1.3). Then straightforward calculations give the following analytical formulae for $\boldsymbol{\epsilon}_V = (\epsilon_{11}^V, \epsilon_{22}^V, \sqrt{2}\epsilon_{12}^V)$ when the applied stress $\boldsymbol{\sigma}_*$ is given by (1.3):

$$\begin{aligned}
\epsilon_{11}^V &= \frac{1}{Er}\left\{\sigma(r+1)^2 + \sigma(r^2-1)\cos 2\beta - r(\sigma+1)\right\},\\
\epsilon_{22}^V &= \frac{1}{Er}\left\{(r+1)^2 + (1-r^2)\cos 2\beta - r(\sigma+1)\right\},\\
\epsilon_{12}^V &= \frac{1}{2Er}(\sigma+1)(r^2-1)\sin 2\beta.
\end{aligned} \tag{4.29}$$

The formulae (4.28) and (4.29) were obtained by Kachanov (1993) as a limiting case of Eshelby's results for an ellipsoidal cavity.

In particular, when $\sigma = -1$ (which corresponds to a pure shear applied stress) we see that ϵ_{12}^V is necessarily zero for all orientations and eccentricities of the ellipse. However, as shown in Figure 10 (see section 5 below), ϵ_{12}^V is nonzero for other inclusion shapes. Thus, when $\sigma = -1$, the maximum value that $|\epsilon_{12}^V|$ attains as an inclusion is rotated can be taken as a measure of its nonellipticity.

Also when $\sigma = 0$ (which corresponds to an applied uniaxial stress), we see that $\epsilon_{11}^V = -1/E$ for all orientations and eccentricities of the ellipse.

4.3. Numerical results for arbitrary shaped holes. Without loss of generality we can take $E = 1$ and assume that $\boldsymbol{\sigma}_*$ has the form (1.3). The convex hull of the set $\widetilde{D}(\boldsymbol{\sigma}_*)$ can be computed from its Legendre transform, which is the minimum value (in the limit $N \to \infty$) of

$$\boldsymbol{\sigma}_*^0 : \boldsymbol{\epsilon}_V = \sigma_{11}^0 \epsilon_{11}^V + \sigma_{22}^0 \epsilon_{22}^V + 2\sigma_{12}^0 \epsilon_{12}^V = \begin{pmatrix} \sigma_{11}^0 \\ \sigma_{22}^0 \\ \sqrt{2}\sigma_{12}^0 \end{pmatrix} \mathbf{E}(c_{-1}, c_{-2}, \ldots, c_{-N}) \begin{pmatrix} \sigma \\ 1 \\ 0 \end{pmatrix} \tag{4.30}$$

as the complex coefficients c_{-1}, c_{-2}, ..., c_{-N}, representing the hole shape, are varied over their admissible range. One has to avoid those sets of coefficients such that the mapping $w(\xi)$ from the exterior of the unit circle onto its image is not one-to-one.

Under the 180° rotation $\mathbf{x} \to -\mathbf{x}$, the tensors $\boldsymbol{\sigma}_*^0$ and $\boldsymbol{\sigma}_*$ remain invariant. Since the bilinear form $\boldsymbol{\sigma}_*^0 : \boldsymbol{\mathcal{E}}\boldsymbol{\sigma}_*$ remains invariant under this rotation, we expect (unless there is symmetry breaking) that the hole shape minimizing this bilinear form should also be invariant under this rotation. Consequently we restrict our attention to conformal maps satisfying

$$\omega(\xi) = -\omega(-\xi), \tag{4.31}$$

or equivalently to complex sets of coefficients c_{-n} which are zero for all even values of n.

By varying $\boldsymbol{\sigma}_*^0$ we could in principle recover the convex hull of the set $\widetilde{D}(\boldsymbol{\sigma}_*)$. However, we decided to first focus on finding the convex hull of the projection of the set $\widetilde{D}(\boldsymbol{\sigma}_*)$ onto the $(\epsilon_{11}^V, \epsilon_{22}^V)$ plane. This projection, which we denote by $\widetilde{D}_1$, represents the set of possible values of $(\epsilon_{11}^V, \epsilon_{22}^V)$. Its convex hull is obtained by taking, say,

$$\boldsymbol{\sigma}_*^0 = \begin{pmatrix} \sigma^0 & 0 \\ 0 & 1 \end{pmatrix} \tag{4.32}$$

and computing the minimum value of the bilinear form (4.30) for varying values of σ^0. Our numerical results suggest that we can associate hole shapes with every point on the boundary of the convex hull. This indicates that $\widetilde{D}_1$ is itself convex, in which case it would coincide with its convex hull. In any case, the convex hull of $\widetilde{D}_1$ is itself important: it represents the set of possible values of $\boldsymbol{\epsilon}_V$ in a plate containing well-separated holes having a mixture of different shapes and sizes, with the holes occupying an infinitesimal average volume fraction f.

The downhill simplex method (see, for example, Press et al. (1986)) was initially used for the minimization, giving quite satisfactory results. Then to obtain more refined results we used the gradient flow method for the minimization as this was faster than the downhill simplex method. The question of uniqueness of the hole shape

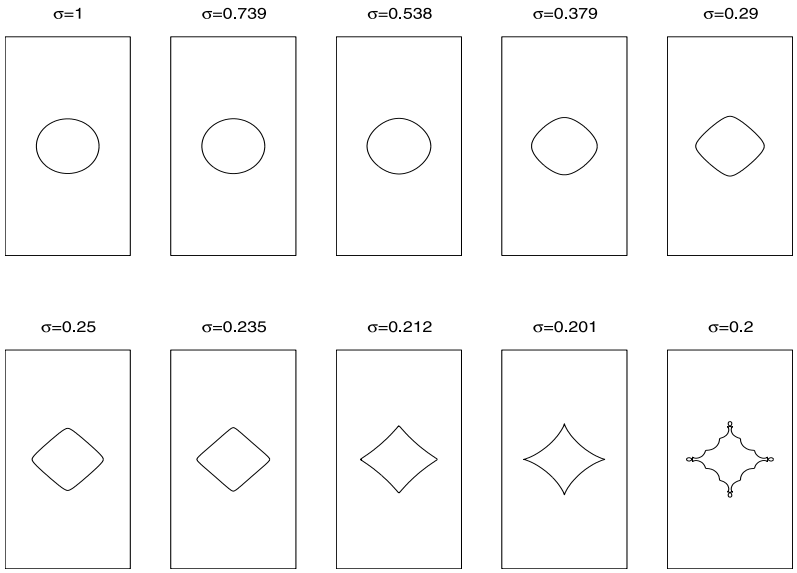


FIG. 1. *Evolution of the optimal hole shapes minimizing the bilinear form with* $\sigma^0 = 1/\sigma$. *The numerical method for finding the hole shapes breaks down in the last plot.*

minimizing the bilinear form and the question of whether the numerically obtained hole shape is a global (and not just local) minimizer of the bilinear form was not studied. Attention was paid to the constraint that the coefficients c_{-n} be admissible only at the end of the calculation. If the hole boundary turned out to self-intersect, we redid the calculation with a smaller step size, which sometimes (but not always) corrected the problem. In the numerical experiments described below, we took $N = 21$ and did the minimization over the 11 complex coefficients c_{-1}, $c_{-3}, \ldots$, c_{-21}. As a check, we also took $N = 23$ in a few test cases and found little change in the shape of the optimal hole. Since the tensors $\boldsymbol{\sigma}_*^0$ and $\boldsymbol{\sigma}_*$ remain invariant under the reflections $x_1 \to -x_1$ and $x_2 \to -x_2$, we expected that the optimal holes would also have this symmetry, i.e., that the minimizing coefficients c_{-n} would turn out to be real. This was found to be the case. After finding an optimal hole (for given values of $\boldsymbol{\sigma}_*^0$ and $\boldsymbol{\sigma}_*$), we calculated the associated value of $\boldsymbol{\epsilon}_V = \mathbf{E}\boldsymbol{\sigma}_*$ lying on the boundary of $\widetilde{D}(\boldsymbol{\sigma}_*)$ and the associated tangent plane to this set having normal $\boldsymbol{\sigma}_*^0$.

An additional simplification occurs in the special case when $\sigma^0 = 1/\sigma$. Then under a 90° rotation the tensor $\boldsymbol{\sigma}_*^0$ is transformed to $\boldsymbol{\sigma}_*/\sigma$, while $\boldsymbol{\sigma}_*^0$ is transformed to $\sigma\boldsymbol{\sigma}_*^0$. Thus the bilinear form $\boldsymbol{\sigma}_*^0 : \boldsymbol{\mathcal{E}}\boldsymbol{\sigma}_*$ remains invariant if we rotate the hole by 90° or if we reflect it about the axes. Unless there is symmetry breaking we expect that the hole shape minimizing this bilinear form should also be invariant under these transformations, that is, it should have square symmetry. Consequently we expect that the minimizing coefficients c_{-n} should all be real and zero unless $n+1$ is a multiple of 4. Figure 1 shows our numerical results for the evolution of optimal shapes as σ is increased. The case $\sigma = 1$ corresponds to a circular hole, which is known to minimize the elastic energy under hydrostatic loading. Figure 2 shows plots of the product $\sigma_{tt}^0(\mathbf{x})\sigma_{tt}(\mathbf{x})$ of the tangential components of the stress field around the boundary of the inclusion. One can see that the optimality criterion (3.4) is approximately satisfied, at least for values of σ which are not too small. The results

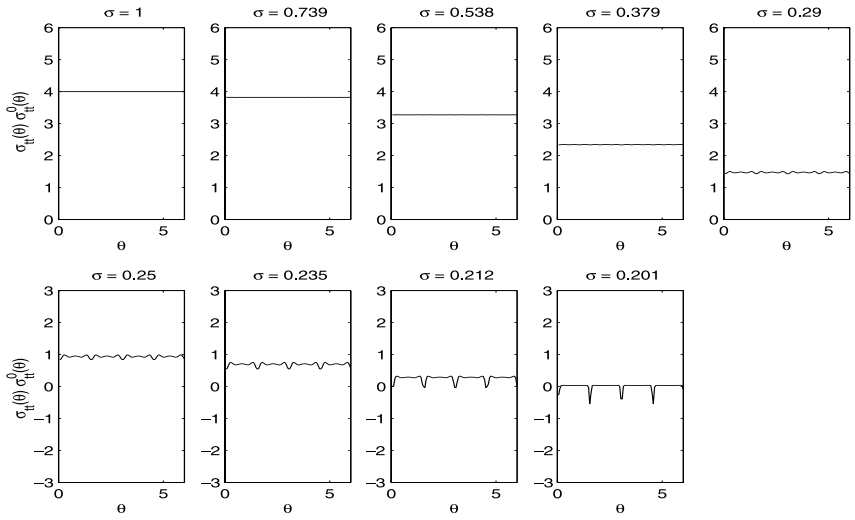


FIG. 2. *Product of the tangential stresses $\sigma^0_{tt}(\theta)$ and $\sigma_{tt}(\theta)$ along the hole boundary for the holes of Figure* 1 *as a function of the angle θ, where $\xi = e^{i\theta}$.*

indicate that the optimal shaped hole becomes nonconvex for sufficiently small σ and that its boundary develops cusps when $\sigma \approx 0.2$. In the next subsection we will provide an analytical formula for what we believe is the optimal hole shape when $\sigma = 0.2$. Below this critical value of σ the method fails; the algorithm produces a nonsensical self-intersecting boundary, as illustrated in the last plot of Figure 1.

In Figures 3–9 (parts (a)), the solid curves are numerical results representing a portion of the boundary of $\widetilde{D}_1$ for various values of σ. The dots represent values of $(\epsilon^V_{11}, \epsilon^V_{22})$ associated with elliptical holes of varying eccentricity and orientation. They are calculated using the analytical formulae (4.29). Parts (b) of Figures 3–9 display the shapes of the optimal holes corresponding to marked points on the solid curves.

For a hydrostatic load, with $\sigma = 1$, Figure 3(a) shows that elliptical shaped holes cover almost the entire set $\widetilde{D}_1$. However, among the ellipses, only the circular hole is optimal for this loading, and it minimizes the elastic energy. One can check that the optimality condition (3.4) is not satisfied for an elliptical hole unless it is a circle. Indeed the optimal shaped holes shown in diagrams A–D and F–L of Figure 3(b) are not exactly ellipses. For this particular loading, the set $\widetilde{D}_1$ is sufficient to allow us to determine the entire set $\widetilde{D}(\boldsymbol{\sigma}_*)$. This is because when $\boldsymbol{\sigma}_* = \mathbf{I}$, we can choose our coordinates so that $\boldsymbol{\epsilon}^V$ is diagonal. Thus the set $\widetilde{D}_1$ also represents the set of possible eigenvalue pairs (λ_1, λ_2) of the matrix $\boldsymbol{\epsilon}^V$.

When σ and σ^0 are both close to 1, the optimal hole is close to being circular in shape. Suppose that the boundary of the hole is the curve traced by $x_1 + ix_2 = r(\theta)e^{i\theta}$ as θ is varied. This serves to define $r(\theta)$, which parameterizes the hole shape. A straightforward but tedious perturbation analysis shows that when

$$\sigma = 1 + c\varepsilon, \quad \sigma^0 = 1 + c_0\varepsilon \tag{4.33}$$

(in which ε is a small parameter and c and c_0 are constants), the optimality condition (3.4) is satisfied when

$$r(\theta) = 1 + \varepsilon\left(\frac{c + c_0}{4}\right)\cos 2\theta + \varepsilon^2\left(\frac{c - c_0}{4}\right)^2\left[\frac{\cos 4\theta}{6} - \cos 2\theta\right] + \mathcal{O}(\varepsilon^3). \tag{4.34}$$

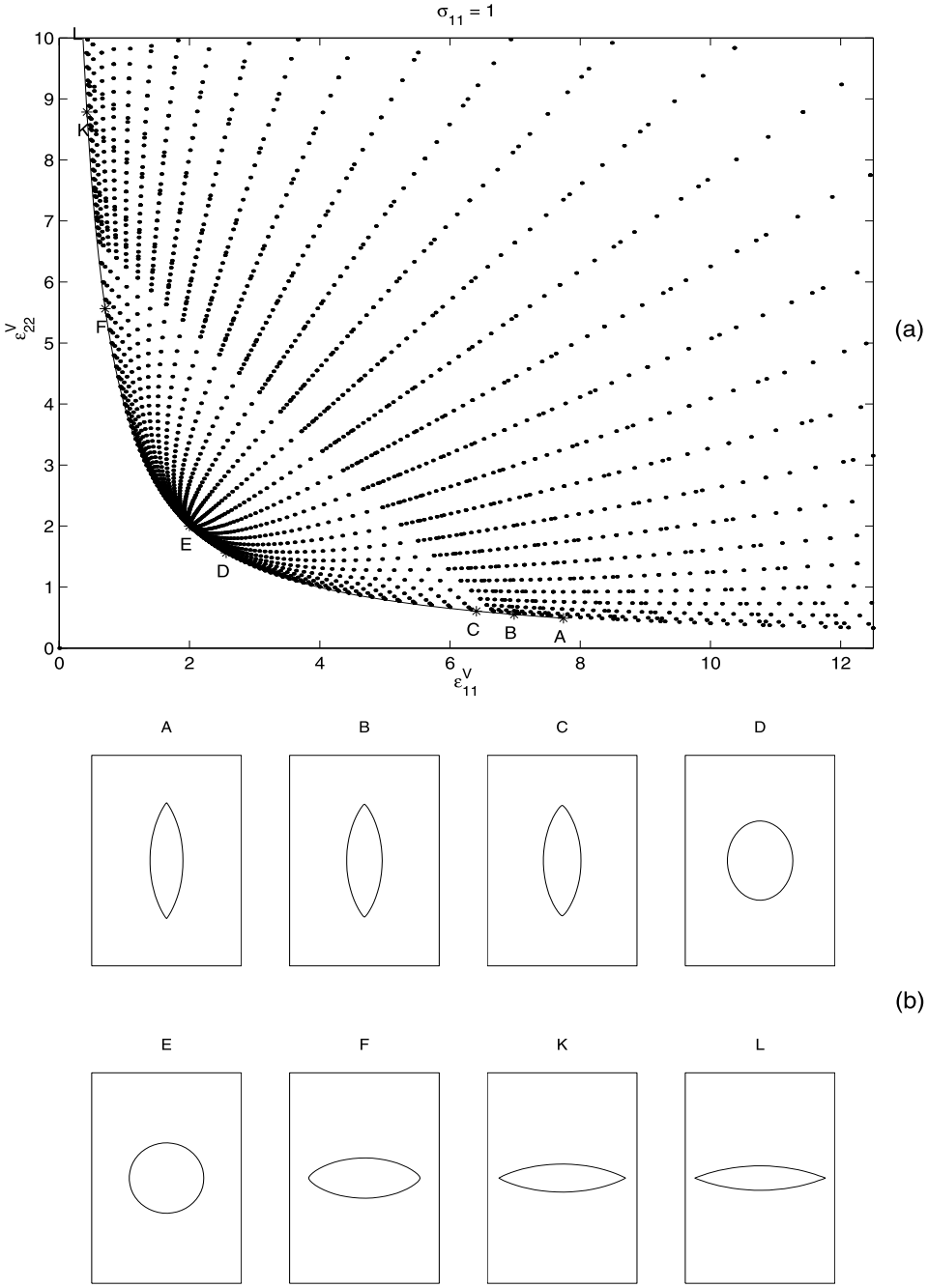


FIG. 3. (a) *The range of values of* $(\epsilon^V_{11}, \epsilon^V_{22})$ *for a periodic array of well separated holes under hydrostatic loading with* $\sigma = 1$. *The dots represent results for elliptical holes, while the solid line is the envelope of numerical results for the optimal shaped holes.* (b) *The shapes of the optimal holes associated with the various points A–L on the solid line.*

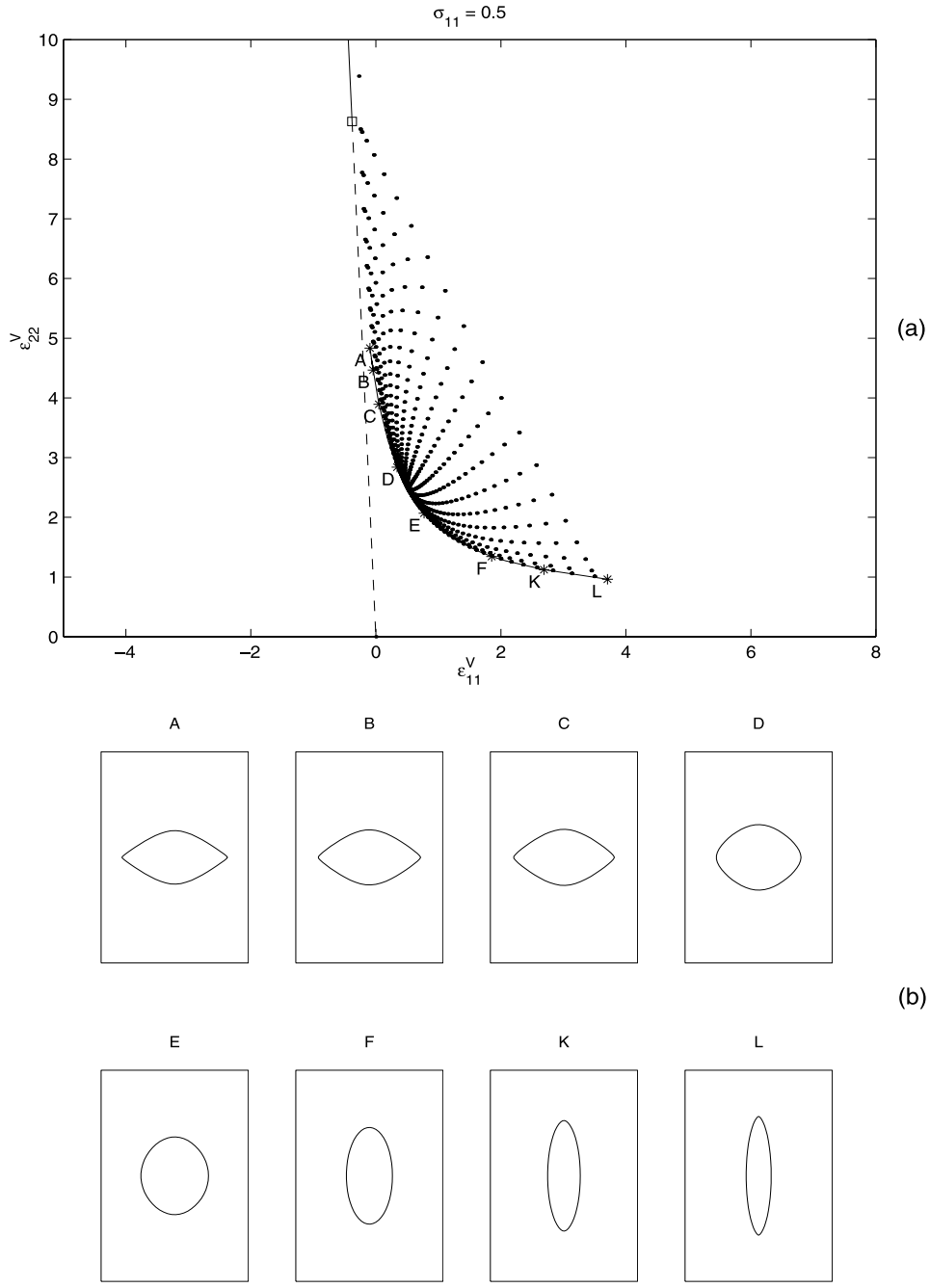


FIG. 4. *As for Figure* 3, *but for biaxial loading with* $\sigma = 0.5$. *The square in* (a) *corresponds to the critical hole of Figure* 12, *with* $\sigma_{tt}(\mathbf{x})$ *being zero on all smooth portions of the boundary. The adjoining straight solid line (whose extension, denoted by the dashed line, passes through the origin) is generated by inserting an island of plate material inside the critical hole.*

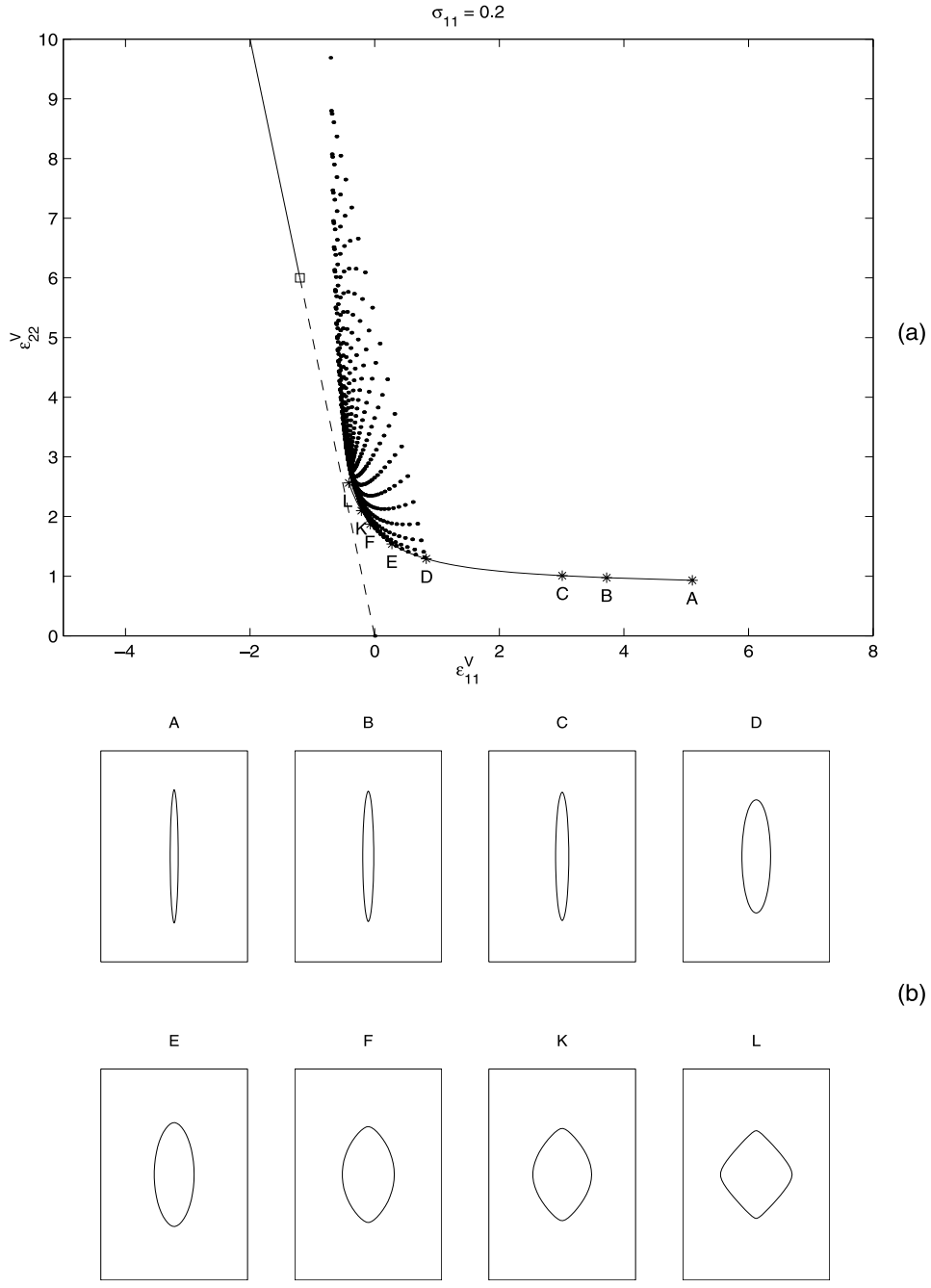


FIG. 5. *As for Figure* 4, *but with* $\sigma = 0.2$. *Both* $\sigma_{tt}(\mathbf{x})$ *and* $\sigma^0_{tt}(\mathbf{x})$ *are zero along the smooth portions of the boundary of the critical hole.*

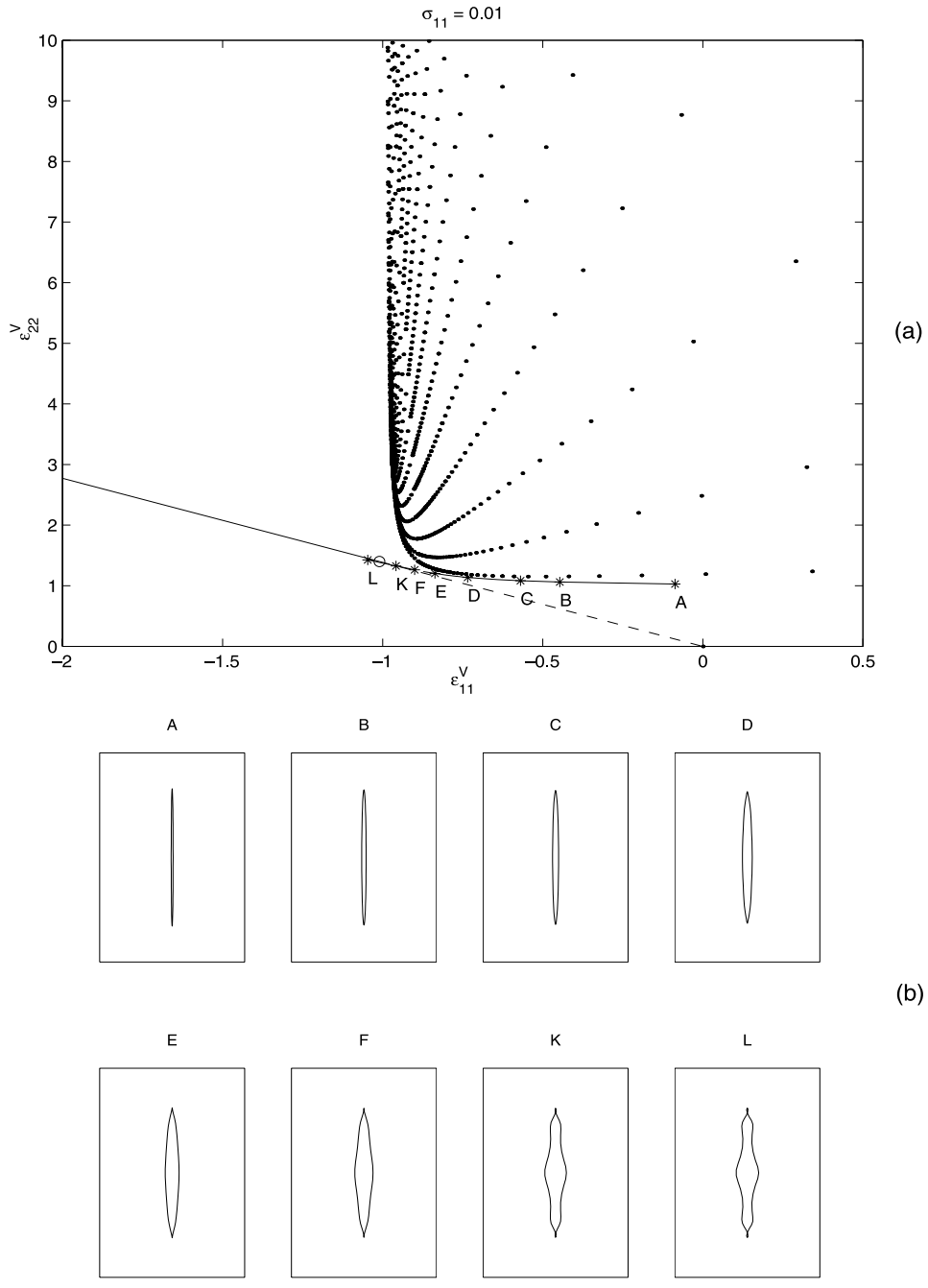


FIG. 6. *As for Figure* 3, *but for biaxial loading with* $\sigma = 0.01$. *The circle in* (a) *corresponds to the critical hole with* $\sigma^0_{tt}(\mathbf{x})$ *being zero on all smooth portions of the boundary. The adjoining straight solid line (whose extension, denoted by the dashed line, passes through the origin) is generated by inserting an island of plate material inside the critical hole. The numerically generated optimal holes corresponding to points K and L are questionable because they have self-intersections close to the corner points.*

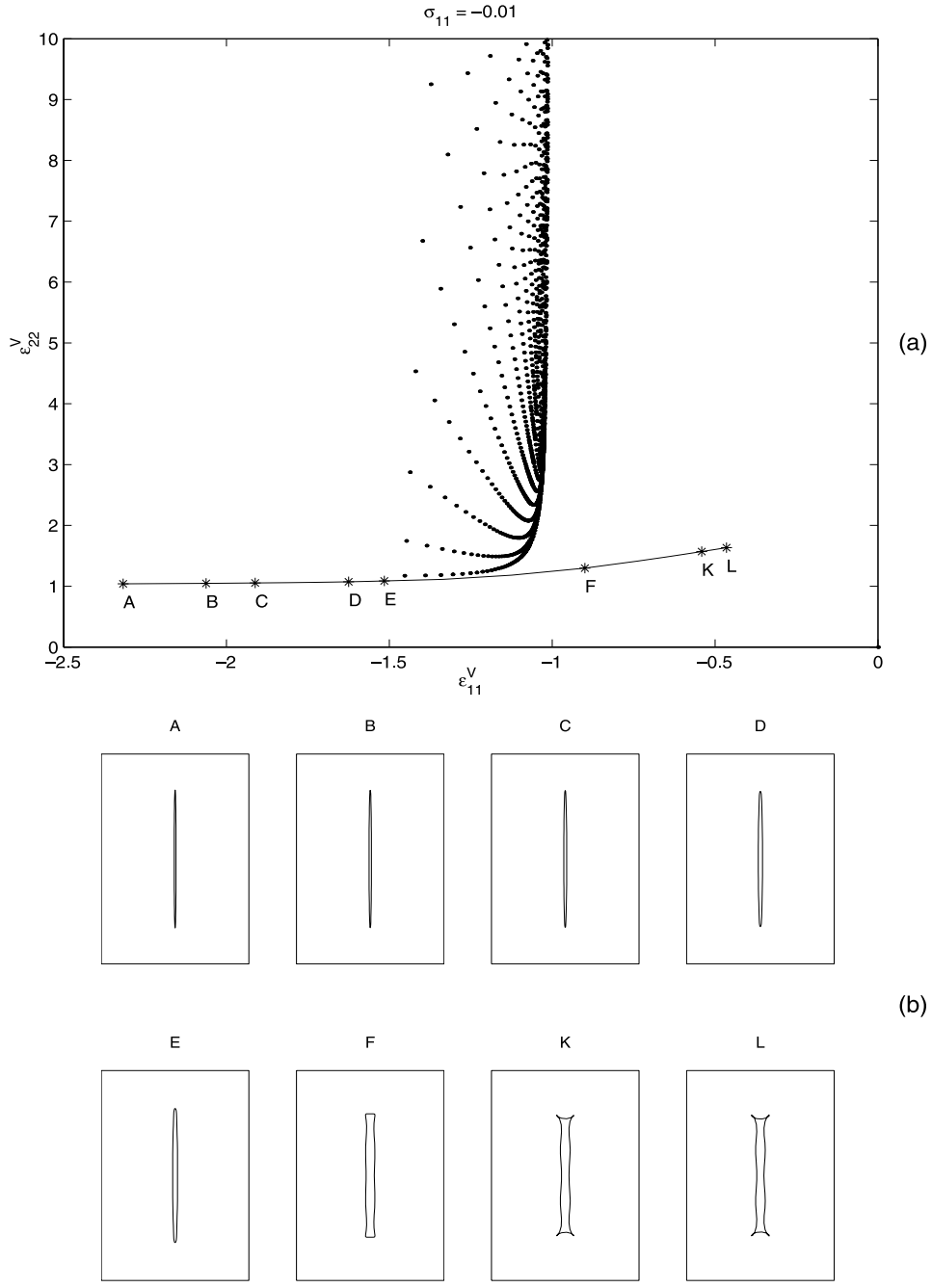


FIG. 7. *As for Figure* 3, *but for biaxial loading with* $\sigma = -0.01$. *The numerically generated optimal holes corresponding to points* K *and* L *are questionable because they have self-intersections close to the corner points.*

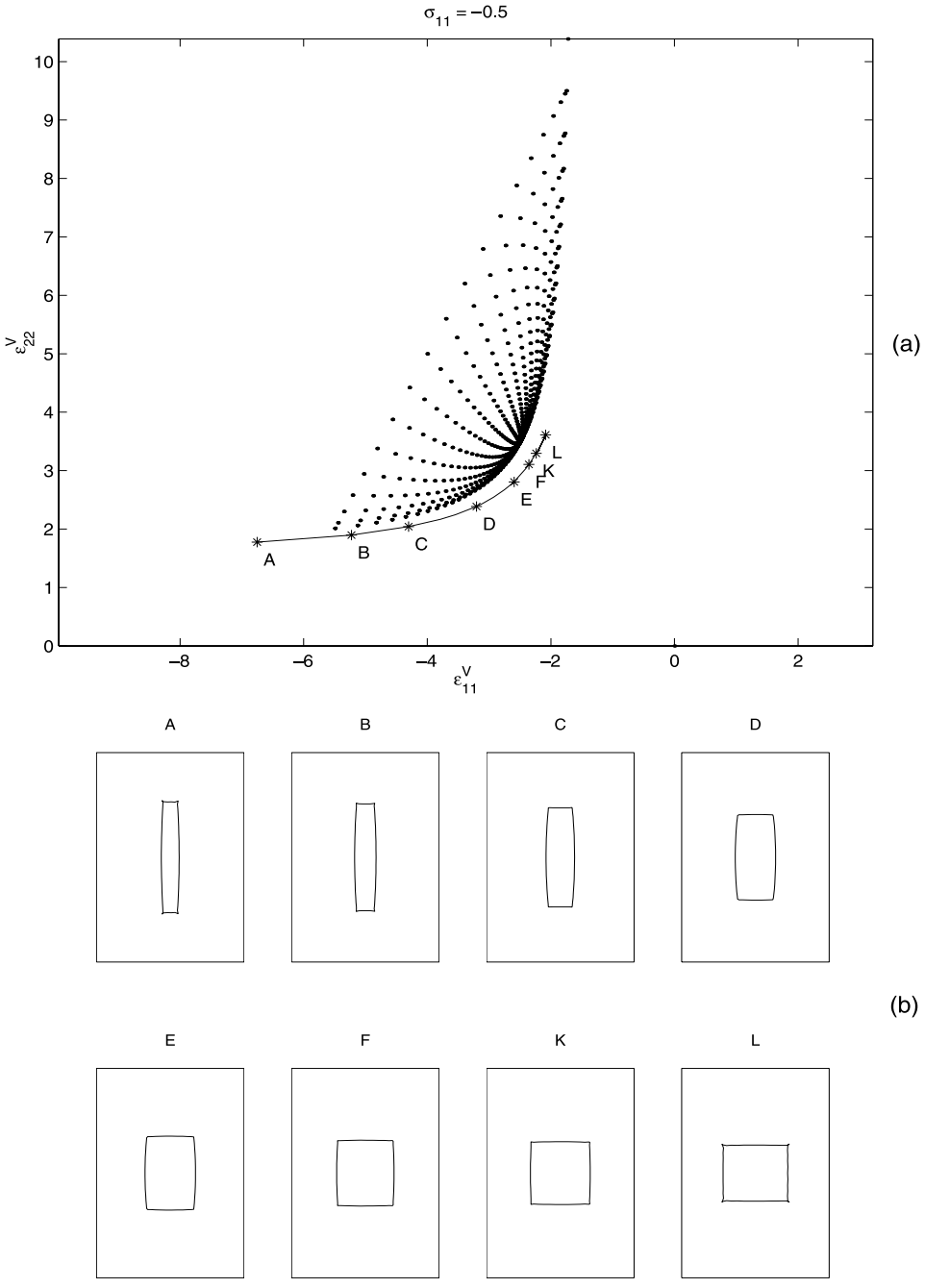


FIG. 8. *As for Figure* 3, *but for biaxial loading with* $\sigma = -0.5$.

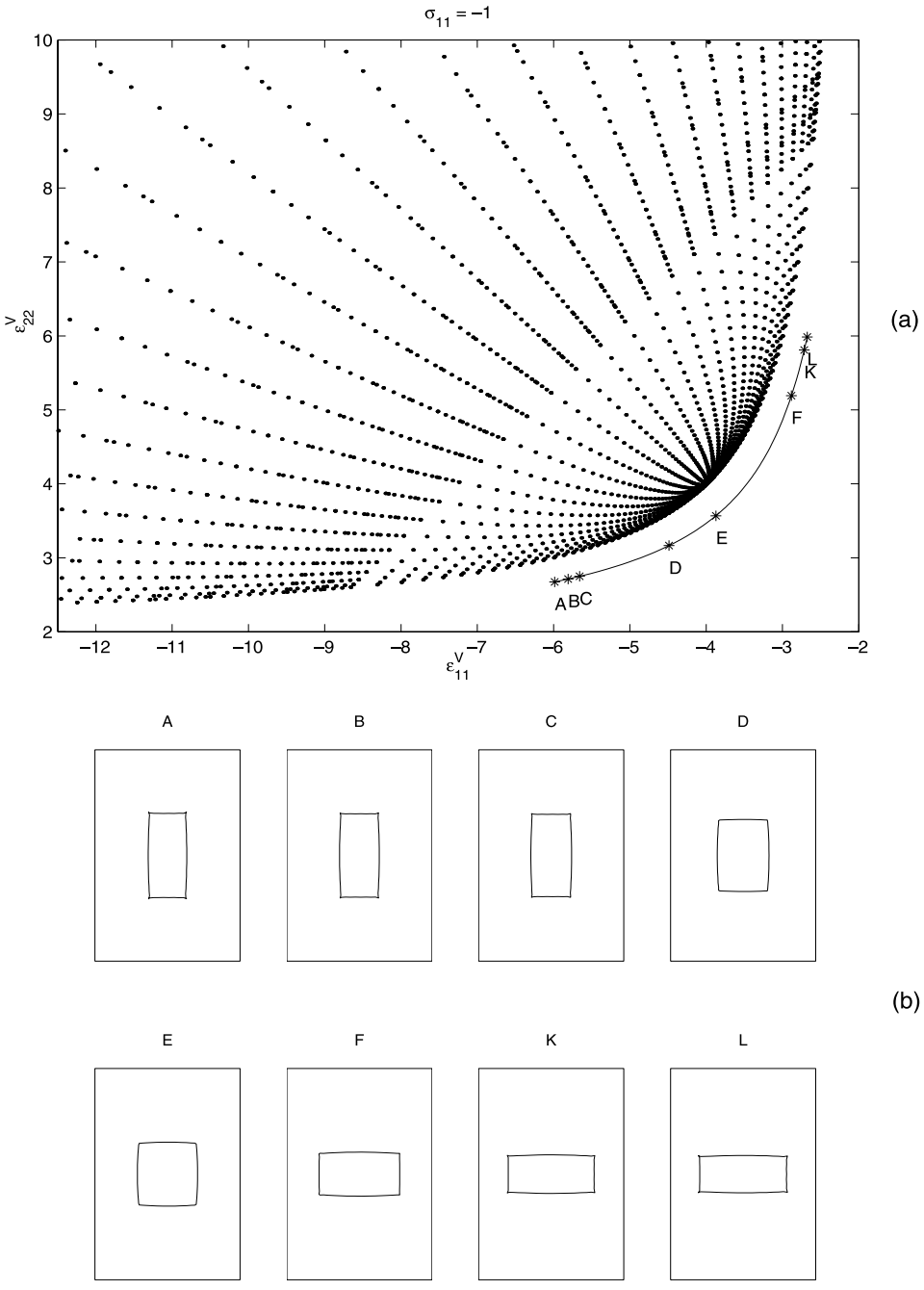


FIG. 9. *As for Figure* 3, *but for pure shear loading with* $\sigma = -1$.

This perturbation analysis and the numerical results suggest that the optimal hole is smooth when ε is sufficiently small, and that it is elliptical or circular to the first order in ε but nonelliptical to the second order in ε (unless $c_0 = c$, in which case the bilinear form is the energy and is minimized by elliptical holes for all $\varepsilon > -1/c$). When $c_0 = -c + c^2\varepsilon$, which corresponds to $\sigma^0 \approx 1/\sigma$, the terms involving $\cos 2\theta$ cancel, and the resulting hole has fourfold symmetry, in agreement with Figure 1.

When σ is less than 1, there are values of $\boldsymbol{\epsilon}_*$ associated with elliptical holes such that $\lambda\boldsymbol{\epsilon}_*$ is not associated with any elliptical hole for some $\lambda > 1$ (but is associated with an elliptical hole with an island in it). In particular, as illustrated in Figures 4, 5, 6, for $\sigma = 0.5$, $\sigma = 0.2$, and $\sigma = 0.01$, respectively, ϵ_{11}^V can be negative for elliptical holes but cannot be arbitrarily large and negative. It follows from (4.29) that $\epsilon_{11}^V \geq \sigma - 1$ for all elliptical holes in a plate with $E = 1$, with the bound being achieved with $\beta = \pi/2$ in the limit $r \to \infty$. Thus the values of $(\epsilon_{11}^V, \epsilon_{22}^V)$ associated with elliptical holes no longer adequately cover $\widetilde{D}_1$. Accordingly, as illustrated in the diagrams A–L of Figures 5(b) and 6(b), the optimal shaped holes are sometimes quite different from elliptical in shape.

For $\sigma = 0$ the algorithm did not produce any reliable results. This case may be rather singular. When σ is infinitesimal and positive, it is possible to find an elliptical hole with an island in it (connected to the surrounding plate by a thin bridge) matching any given value of $(\epsilon_{11}^V, \epsilon_{22}^V)$ with $\epsilon_{11}^V + \epsilon_{22}^V > 0$ and $\epsilon_{22}^V > 1$. Unless $\epsilon_{11}^V \approx -1$, the required hole will be very elongated in shape and aligned at a slight angle to the x_2 axis. When σ is infinitesimal and negative, it is possible to find an elliptical hole matching any given value of $(\epsilon_{11}^V, \epsilon_{22}^V)$ with $\epsilon_{11}^V < -1$ and $\epsilon_{22}^V > 1$. Again unless $\epsilon_{11}^V \approx -1$, the required ellipse will be very elongated in shape and aligned at a slight angle to the x_2 axis. When $\sigma = 0$, the bound on the elastic energy implies that $\epsilon_{22}^V \geq 1$ for any shaped hole.

For negative values of σ the optimal shaped holes change to being almost rectangular in shape. This is illustrated in Figures 7, 8, and 9, for $\sigma = -0.01$, $\sigma = -0.5$, and $\sigma = -1$, respectively.

We also obtained some results for the projection of the set $\widetilde{D}(\boldsymbol{\sigma}_*)$ onto the $(\sigma\epsilon_{11}^V + \epsilon_{22}^V, \epsilon_{12}^V)$ plane. (The prefactor of σ is added to ϵ_{11}^V so that nontrivial projections are obtained for both positive and negative values of σ.) This projection, which we denote as $\widetilde{D}_2$, represents the set of possible values of $(\sigma\epsilon_{11}^V + \epsilon_{22}^V, \epsilon_{12}^V)$. The convex hull of this projection is obtained by taking

$$\boldsymbol{\sigma}_*^0 = \begin{pmatrix} \sigma & \delta \\ \delta & 1 \end{pmatrix} \tag{4.35}$$

and varying δ. For the hydrostatic load $\sigma = 1$ the results for $\widetilde{D}_1$ in Figure 3(a) (representing the set of possible eigenvalue pairs of the matrix $\boldsymbol{\epsilon}_*$) can be used to determine $\widetilde{D}_2$. The optimal holes associated with the boundary of $\widetilde{D}_2$ are obtained by rotating the optimal holes associated with the boundary of $\widetilde{D}_1$. For the loading with $\sigma = -1$ the tensors $\boldsymbol{\sigma}_*$ and $\boldsymbol{\sigma}_*^0$ are transformed under a 90° rotation to $-\boldsymbol{\sigma}_*$ and $-\boldsymbol{\sigma}_*^0$, respectively, leaving the bilinear form $\boldsymbol{\sigma}_*^0 : \boldsymbol{\mathcal{E}}\boldsymbol{\sigma}_*$ invariant. Unless there is symmetry, breaking the optimal hole should also have this symmetry; that is, the minimizing complex coefficients c_{-n} should be zero unless $n + 1$ is a multiple of 4. Numerical results for $\widetilde{D}_2$ are presented in Figure 10(a). As discussed in section 4.2, elliptical holes necessarily have $\epsilon_{12}^V = 0$ when $\sigma = -1$. The optimal shaped holes are roughly rotated squares, as illustrated in the diagrams A–L of Figure 10(b). The shape changes slightly from one diagram to the next, and is not just a rotation of the

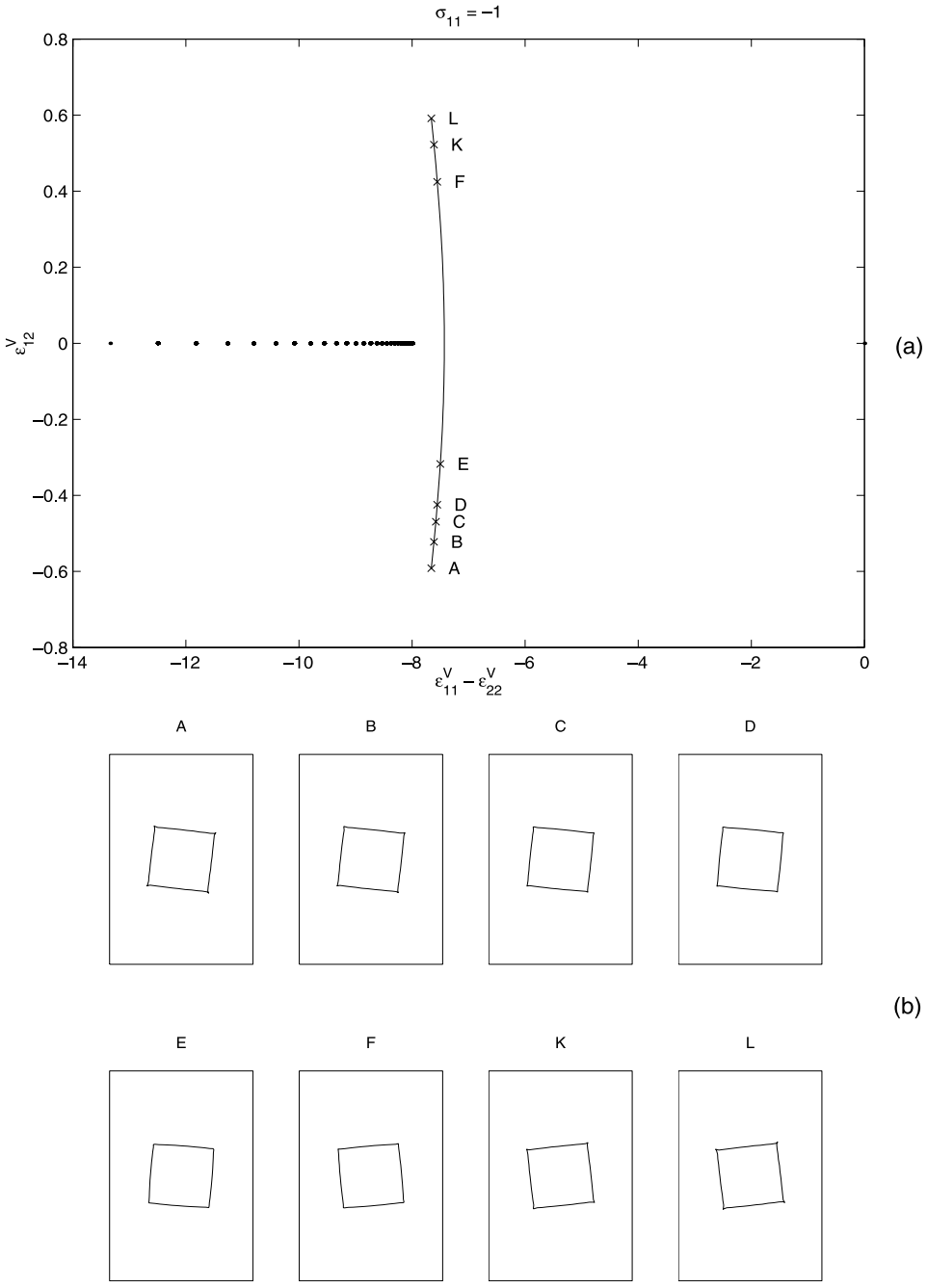


FIG. 10. (a) *The range of values of* $(\epsilon_{11}^V - \epsilon_{22}^V, \epsilon_{12}^V)$ *for a periodic array of well separated holes under pure shear loading with* $\sigma = -1$. *Elliptical holes, represented by the dots, necessarily have* $\epsilon_{12}^V = 0$. *The solid line is the envelope of numerical results for the optimal shaped holes.* (b) *The shapes of the optimal holes associated with the various points A–L on the solid line.*

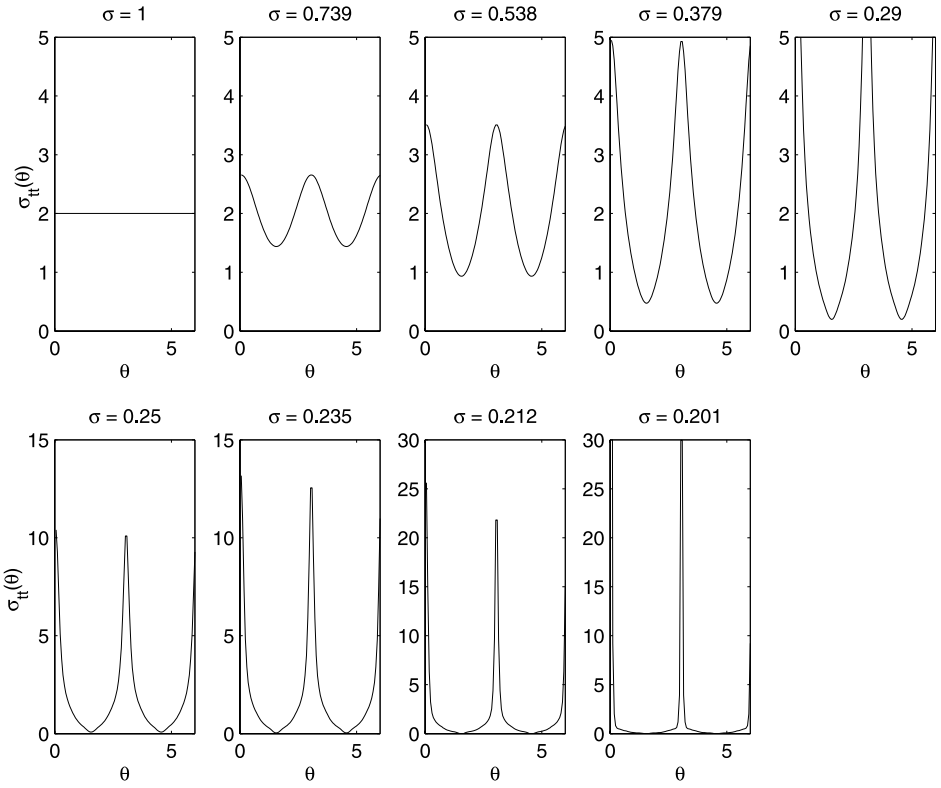


FIG. 11. *Plots of the tangential stress $\sigma_{tt}(\theta)$ along the hole boundary for the holes of Figure* 1, *with the loading σ as a function of the angle θ, where $\xi = e^{i\theta}$. Notice how the tangential stress concentrates at $\theta = 0$ and $\theta = \pi$ as σ approaches* 0.2. *The plots of the tangential stress $\sigma^0_{tt}(\theta)$ are similar but shifted horizontally by an angle of $\pi/2$.*

same hole. Indeed a hole which satisfies the optimality condition (3.4) is generally not going to satisfy the condition for any choice of $\boldsymbol{\sigma}^0_*$ when $\boldsymbol{\sigma}_*$ is rotated.

The calculations described above are based on numerical experiments where we have minimized over a finite set of conformal mapping coefficients c_n, treating them as controls. It is possible that we have not adequately explored the set of admissible shapes, and it is possible that the shapes found are only local minimizers and not global minimizers. These are difficult questions that one always faces when trying to approximate an infinite-dimensional minimization problem by a finite-dimensional one. Since relatively large perturbations to the boundary sometimes produce small changes in $\boldsymbol{\epsilon}_V$, we believe our sets of admissible values of $\boldsymbol{\epsilon}_V$ to be generally more reliable than our results for the shapes of the optimal holes.

5. Critical holes. Consider the graphs in Figure 2 of the tangential stress product $\sigma^0_{tt}\sigma_{tt}$ around the boundary of the hole. As σ approaches its critical value $\sigma \approx 0.2$, the product approaches zero. Further numerical investigations (see Figure 11) showed that both σ_{tt} and σ^0_{tt} approach zero on all smooth portions on the boundary, and that σ_{tt} blows up to infinity at two opposing corner points, while σ^0_{tt} blows up to infinity at the remaining two opposing corner points. Certainly the optimality criterion (3.4) will be satisfied if σ_{tt} (or alternatively σ^0_{tt}) is zero on all smooth portions of the boundary

Γ. Let us try to find holes having this property, which we will call critical holes. It was a surprise to us to discover that such holes exist, with the stress field $\boldsymbol{\sigma}(\mathbf{x})$ (or $\boldsymbol{\sigma}^0(\mathbf{x})$) being zero along the boundary except at the corner points. To our knowledge they have not been previously studied.

Without loss of generality let us suppose that the two corner points where σ_{tt} blows up correspond to $\xi = 1$ and $\xi = -1$. (We no longer assume that $\boldsymbol{\sigma}_*$ has the diagonal form (1.3).) The analytic function $\Phi(\xi)$, which determines σ_{tt} through (4.19), must have zero real part on the unit circle $|\xi| = 1$, except at the points $\xi = 1$ and $\xi = -1$, where it must have a delta function singularity. Also it must approach α in the limit $\zeta \to \infty$ because $\phi'(\xi)/\omega'(\xi)$ approaches 1 in this limit. These considerations imply that $\Phi(\xi)$ is given by the expression

$$\Phi(\xi) = \lim_{\epsilon \to 0} \left[\alpha + \frac{\alpha}{\xi + \varepsilon - 1} - \frac{\alpha}{\xi + 1 - \varepsilon} \right] = \frac{\alpha(\xi - 1)}{2(\xi + 1)} + \frac{\alpha(\xi + 1)}{2(\xi - 1)}. \tag{5.1}$$

In the ensuing calculations one should think of ϵ as being infinitesimal and strictly positive, so that the singularities of $\Phi(\xi)$ lie inside the unit circle, even though we will be looking at what happens in the limit $\epsilon \to 0$. (The only place one has to be careful in making the distinction between zero ϵ and infinitesimal ϵ is in evaluating the integral (5.4).)

To obtain $\omega(\xi)$ from $\Phi(\xi)$ we follow the approach of Cherkaev et al. (1998), who solve this type of problem with a different $\Phi(\xi)$. By differentiating (4.9) and using the relation $\overline{\Phi(t)} = -\Phi(t) + 2\,\mathrm{Re}\,\Phi(t)$, we see that

$$\omega'(\xi)\Phi(\xi) + \frac{1}{2\pi i} \int_{|t|=1} \frac{\omega(t)\Phi(t)}{(t - \xi)^2} dt - \frac{1}{\pi i} \int_{|t|=1} \frac{\omega(t)\,\mathrm{Re}\,\Phi(t)}{(t - \xi)^2} dt = \alpha + \overline{\gamma}\xi^{-2}. \tag{5.2}$$

The first integral can be evaluated using the Cauchy integral theorem,

$$\frac{1}{2\pi i} \int_{|t|=1} \frac{\omega(t)\Phi(t)}{(t - \xi)^2} dt = \alpha - \frac{d}{d\xi}[\omega(\xi)\Phi(\xi)] = \alpha - \omega'(\xi)\Phi(\xi) + \frac{4\alpha\xi\omega(\xi)}{(\xi^2 - 1)^2} \tag{5.3}$$

in which the leading constant α comes from the integral around a circle of very large radius. The second integral can be making the substitution (5.1), keeping ε finite, and taking the limit $\varepsilon \to 0$, giving

$$\frac{1}{\pi i} \int_{|t|=1} \frac{\omega(t)\,\mathrm{Re}\,\Phi(t)}{(t - \xi)^2} dt = \frac{\omega(1)\alpha}{(\xi - 1)^2} - \frac{\omega(-1)\alpha}{(\xi + 1)^2}. \tag{5.4}$$

Assuming $\omega(-1) = -\omega(1)$ and substituting (5.3) and (5.4) back into (5.2) results in a simple expression for $\omega(\xi)$ involving $\omega(1)$. The constant $\omega(1)$ can then be determined from the constraint that $\omega(\xi)/\xi$ approaches 1 as $\xi \to \infty$. In this way we see that necessarily

$$\omega(\xi) = \xi + (1 - 3k)\xi^{-1} + k\xi^{-3}, \quad \text{where } k = \frac{\overline{\gamma}}{4\alpha}. \tag{5.5}$$

This is an exact formula, not an approximation. The requirement that the mapping from the exterior of the unit disk onto its image be one-to-one will be satisfied if and only if k lies inside the intersection of the disk of radius $1/3$ with the region Q whose boundary is the curve traced out by

$$k = \frac{e^{i\eta}(1 - e^{i\eta})}{1 + 3e^{i\eta}} \tag{5.6}$$

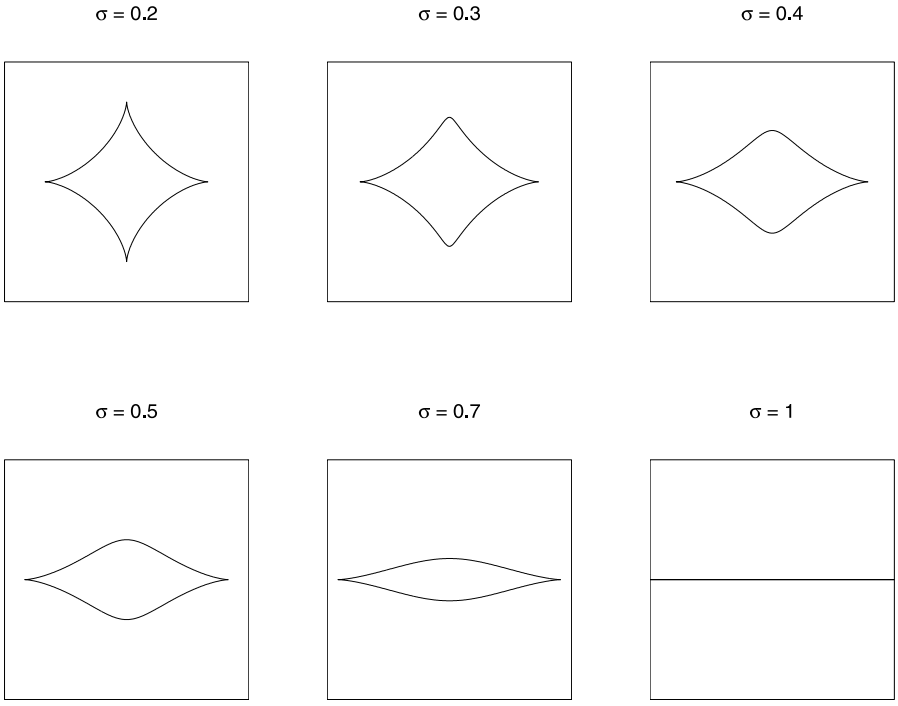


FIG. 12. *Critical holes associated with various values of* σ *between* 0.2 *and* 1. *The tangential stress* $\sigma_{tt}(\theta)$, *and hence all components of the stress field* $\mathfrak{E}(\mathbf{x})$, *are zero along the smooth portions of the boundary.*

as η varies between 0 and 2π. The constraint that $|k| \leq 1/3$ ensures that $\omega'(\xi)$ is nonzero when $|\xi| > 1$, while the constraint that $k \in Q$ ensures that the image of the circle $|\xi| = 1$ does not intersect itself and that $\omega(e^{i\theta})$ moves anticlockwise around the boundary as θ is increased from 0 to 2π. When k is real, these restrictions force k to lie between 0 and 1/3. Figure 12 shows examples of the hole shapes for $k = \overline{\gamma}/(4\alpha) = (1-\sigma)/[2(1+\sigma)]$ (corresponding to the loading (1.3)) as σ varies between 0.2 and 1. Notice that the hole shape for $\sigma = 0.2$ matches the hole shape found numerically in the next to last plot in Figure 1. Figure 13 shows examples of the hole shapes for complex values of k ranging over the boundary of admissible values. These unusual hole shapes might have been seen in our numerical simulations, had we explored the full range $\widetilde{D}(\boldsymbol{\sigma}_*)$ of values of $\boldsymbol{\epsilon}_V$, and not just the range $\widetilde{D}_1$ of values of the pair $(\epsilon_{11}^V, \epsilon_{22}^V)$.

By integrating $\phi'(\xi) = \omega'(\xi)\Phi(\xi)$, using (4.20) and the fact that $\overline{t} = 1/t$ when $|t| = 1$, one finds that the potentials ϕ and ψ are given by the expressions

$$\begin{aligned} \phi &= \alpha[\xi - (1+3k)\xi^{-1} - k\xi^{-3}] = \alpha z + Az^{-1} + \mathcal{O}(z^{-3}), \\ \psi &= 4\alpha\overline{k}\xi + \frac{4\alpha(\overline{k}-1)\xi}{(\xi^2-1)} = 4\alpha\overline{k}z + Bz^{-1} + \mathcal{O}(z^{-3}), \end{aligned} \tag{5.7}$$

in which the coefficients

$$A = -2\alpha, \quad B = 4\alpha(3|k|^2 - 1) \tag{5.8}$$

govern the leading corrections to the far field behavior in the z-plane. The average

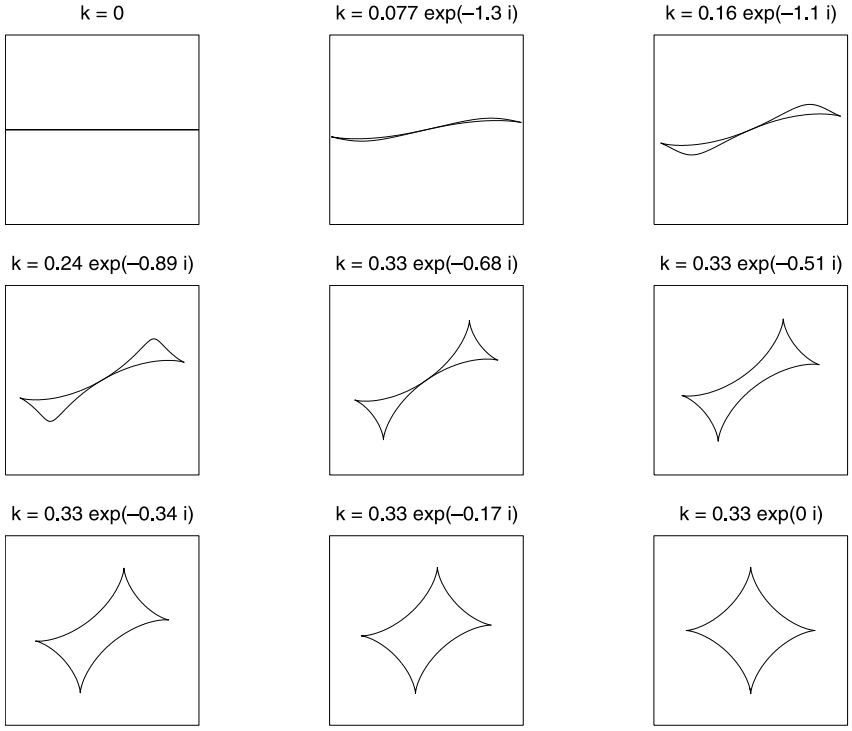


FIG. 13. *Critical holes associated with values of k at the boundary of its range of admissible values.*

strain in the void phase is determined by these coefficients and by the area V of the hole, given by (4.24). A straightforward computation (similar to the one given by Movchan and Serkov (1997) for determining $\mathbf{E}$) shows that

$$\boldsymbol{\epsilon}_V = \frac{\pi}{V}\begin{pmatrix} 2\,\mathrm{Re}(2A-B) & 4\,\mathrm{Im}(A) \\ 4\,\mathrm{Im}(A) & -2\,\mathrm{Re}(2A+B)\end{pmatrix}, \tag{5.9}$$

giving

$$\boldsymbol{\epsilon}_V = \frac{4\alpha}{3\,\mathrm{Re}(k)-6|k|^2}\begin{pmatrix} -3|k|^2 & 0 \\ 0 & 2-3|k|^2\end{pmatrix}. \tag{5.10}$$

For the loading (1.3), which corresponds to $\alpha = (1+\sigma)/4$ and $k = (1-\sigma)/[2(1+\sigma)]$, the formula (5.10) implies

$$\epsilon_{11}^V = -\frac{(1-\sigma^2)}{4\sigma}, \qquad \epsilon_{22}^V = \frac{(1+\sigma)(5+22\sigma+5\sigma^2)}{12\sigma(1-\sigma)}, \tag{5.11}$$

and the restrictions on k force σ to lie between 0.2 and 1. This average stress within the void phase is represented by the small square in Figures 4 and 5.

Suppose that the hole generated by (5.5) for a fixed value of k is subject to a different loading $\boldsymbol{\sigma}_*^0$ having loading coefficients

$$\alpha^0 = \frac{(\sigma_{11}^{*0}+\sigma_{22}^{*0})}{4}, \qquad \gamma^0 = \frac{(\sigma_{22}^{*0}-\sigma_{11}^{*0}+2i\sigma_{12}^{*0})}{2}. \tag{5.12}$$

The integral equation (4.9) (with α and γ being replaced by the new loading coefficients) can be solved for $\phi(\xi)$, giving

$$\phi(\xi) = \alpha^0\zeta - \frac{\overline{\gamma^0} + \alpha^0(1-k)}{\xi} + \frac{(4\alpha^0\overline{k} - \gamma^0)k - (4\alpha^0 k - \overline{\gamma^0})|k|^2}{(1-|k|^2)\xi} - \frac{\alpha^0}{\xi^3}, \tag{5.13}$$

which agrees with (5.7) when $\alpha^0 = \alpha$ and $\gamma^0 = 4\alpha\overline{k}$.

Now the optimality criterion (3.4) seems to say nothing about σ^0_{tt} when σ_{tt} is zero on all smooth portions of the boundary of Γ. However, here we will present a plausibility argument to suggest that σ^0_{tt} should be finite at $\xi = 1$ and $\xi = -1$. Assume otherwise, and suppose that there exist small perturbations of the loadings $\boldsymbol{\sigma}_*$ and $\boldsymbol{\sigma}^0_*$ that result in the optimal hole minimizing the bilinear form $\boldsymbol{\sigma}^0_* : \boldsymbol{\mathcal{E}}\boldsymbol{\sigma}_*$ having a completely smooth boundary. The functions σ_{tt} and σ^0_{tt} should remain approximately the same with sharp peaks at $\xi = 1$ and $\xi = -1$. But then the product $\sigma^0_{tt}\sigma_{tt}$ cannot be constant around the boundary, and we have a contradiction. On the other hand, assume that σ^0_{tt} was originally finite, taking the value s^0 at $\xi = 1$ and at $\xi = -1$. The product $\sigma^0_{tt}\sigma_{tt}$ will equal some small constant ε, and therefore $\sigma^0_{tt} = \varepsilon/\sigma_{tt}$ will have a sharp dip (with a minimum close to zero) after the perturbation. If σ_{tt} is of the order of ε away from its peak, then we can see that σ^0_{tt} could remain approximately equal to s^0 in the vicinity of $\xi = 1$ and $\xi = -1$, but away from the dip at these points. The dip can be regarded as being due to a small multiple of σ_{tt} being subtracted from the unperturbed potential σ^0_{tt} because of the slightly different loading. In the special case when $s^0 = 0$ the argument has to be modified, but everything works out if, for example, σ_{tt} is of the order of $\sqrt{\varepsilon}$ away from its peak. This is only a plausibility argument because it breaks down if all small perturbations of loading result in optimal holes having nonsmooth boundaries. However, the results we obtain by assuming that σ^0_{tt} is finite at $\xi = 1$ and $\xi = -1$ justify this assumption.

The potential $\Phi(\xi) = \phi'(\xi)/\omega'(\xi)$ will have a finite real part at $\xi = 1$ and at $\xi = -1$ if $\phi'(\xi)$ is zero at these points. This cancels the effect of $\omega'(\xi)$ being zero at these points. Thus the tangential stress component σ^0_{tt} will be finite at the cusp points when

$$\gamma^0 = 2\alpha^0(3|k|^2 - 1), \tag{5.14}$$

in which case $\boldsymbol{\sigma}^0_*$ takes the form

$$\boldsymbol{\sigma}^0_* = 2\alpha^0 \begin{pmatrix} 2 - 3|k|^2 & 0 \\ 0 & 3|k|^2 \end{pmatrix}. \tag{5.15}$$

Consequently we have

$$\boldsymbol{\sigma}^0_* : \boldsymbol{\mathcal{E}}\boldsymbol{\sigma}_* = \boldsymbol{\sigma}^0_* : \boldsymbol{\epsilon}_V = 0, \tag{5.16}$$

and if this represents the minimum value of the bilinear form $\boldsymbol{\sigma}^0_* : \boldsymbol{\mathcal{E}}\boldsymbol{\sigma}_*$, then we can place an island of plate material inside the inclusion (connected to the surrounding plate by a very thin bridge of plate material) without disturbing the minimum. Thus there would be a whole family of optimal holes minimizing the bilinear form, with values of $\boldsymbol{\epsilon}_V$ that are λ times the $\boldsymbol{\epsilon}_V$ given by (5.10), with $\lambda \geq 1$. Certainly the optimality criterion (3.4) is satisfied for such holes, since the product $\sigma^0_{tt}\sigma_{tt}$ is zero around the boundary of the hole including at the shore of the island (where both σ^0_{tt} and σ_{tt} are zero).

By substituting (5.14) back into (5.13), using (4.20) and the fact that $\bar{t} = 1/t$ when $|t| = 1$, one obtains the potentials ϕ and ψ associated with the loading $\boldsymbol{\sigma}_*^0$:

$$\begin{aligned}\phi &= \alpha^0\zeta + \alpha^0(1+3k)\xi^{-1} - \alpha^0\xi^{-3} = \alpha^0 z + A^0 z^{-1} + \mathcal{O}(z^{-3}),\\ \psi &= 2\alpha^0(3|k|^2-1)\xi - \frac{2\alpha^0[9|k|^2(k+1)-3k+1]\xi}{\xi^2+3k}\\ &= 2\alpha^0(3|k|^2-1)z + B^0 z^{-1} + \mathcal{O}(z^{-3}),\end{aligned} \tag{5.17}$$

in which

$$A^0 = 6k\alpha^0, \quad B^0 = -24|k|^2\alpha^0. \tag{5.18}$$

The associated average stress within the void phase by direct analogy with (5.9) is

$$\boldsymbol{\epsilon}_V^0 = \frac{1}{6\,\mathrm{Re}(k) - 12|k|^2}\begin{pmatrix} 2\,\mathrm{Re}(2A^0 - B^0) & 4\,\mathrm{Im}(A^0) \\ 4\,\mathrm{Im}(A^0) & -2\,\mathrm{Re}(2A^0 + B^0)\end{pmatrix}. \tag{5.19}$$

One can check that $\boldsymbol{\sigma}_* : \boldsymbol{\epsilon}_V^0 = 0$, which agrees with (5.16).

In the special case in which the loadings $\boldsymbol{\sigma}_*$ and $\boldsymbol{\sigma}_*^0$ take the forms (1.3) and (4.32), corresponding to $\alpha^0 = 1/(6|k|^2)$ and $k = (1-\sigma)/[2(1+\sigma)]$, we have

$$\sigma^0 = 2\alpha^0(2 - 3|k|^2) = \frac{5 + 22\sigma + 5\sigma^2}{3(1-\sigma)^2}, \tag{5.20}$$

and the constraint that σ lies between 0.2 and 1 forces σ^0 to lie between 5 and infinity. According to the formula (5.18) and (5.19) the matrix $\boldsymbol{\epsilon}_V^0$ has elements

$$\begin{aligned}\epsilon_{11}^{0V} &= \frac{1+\sigma^0}{\sigma} = \frac{(1+\sigma^0)(3\sigma^0-5)}{3\sigma^0 + 11 - 4\sqrt{6(\sigma^0+1)}},\\ \epsilon_{22}^{0V} &= -(1+\sigma^0), \quad \epsilon_{12}^{0V} = 0,\end{aligned} \tag{5.21}$$

where we have used (5.20) to express these elements entirely in terms of the constant σ^0 associated with the applied loading. Now let us rotate the hole by 90°, rotate $\boldsymbol{\sigma}_*^0$ and $\boldsymbol{\epsilon}_V^0$ by 90°, and divide both matrices by σ^0. Then, by relabelling $1/\sigma^0$ as σ, we see that the applied loading (1.3) with σ between 0 and 0.2 can produce an average stress $\boldsymbol{\epsilon}_V$ within the void phase having elements

$$\epsilon_{11}^V = -(1+\sigma), \quad \epsilon_{12}^V = 0, \quad \epsilon_{22}^V = \frac{(1+\sigma)(3-5\sigma)}{3 + 11\sigma - 4\sqrt{6\sigma(1+\sigma)}}. \tag{5.22}$$

This average stress within the void phase is represented by the small circle in Figure 6.

For negative values of σ, Figures 7, 8, and 9 suggest that the optimal holes are somewhat rectangular in shape. Accordingly, it makes sense to study critical holes for which the tangential stress (and hence the real part of $\Phi(\xi)$) vanishes except at four points on the boundary of the unit circle $|\xi| = 1$, which we take to be the points p, $-p$, $\overline{p} = 1/p$, and $-\overline{p} = -1/p$, where $p = e^{i\beta}$. Assuming that the applied loading is diagonal (i.e., γ is real), that σ_{tt} around the hole boundary has delta function singularities at the four points, and that the hole shape and potentials have inversion and reflection symmetry, which implies

$$\begin{aligned}&\omega(-\xi) = -\omega(\xi), \quad \phi(-\xi) = -\phi(\xi), \quad \Phi(-\xi) = \Phi(\xi),\\ &\overline{\omega(\xi)} = \omega(\overline{\xi}), \quad \overline{\phi(\xi)} = \phi(\overline{\xi}), \quad \overline{\Phi(\xi)} = \Phi(\overline{\xi}),\end{aligned} \tag{5.23}$$

the function $\Phi(\xi) = \phi'(\xi)/\omega'(\xi)$ is easily deduced to be

$$\begin{aligned}\Phi(\xi) &= \lim_{\varepsilon\to 0}\left[\alpha + \frac{\alpha p}{\xi+\varepsilon-p} - \frac{\alpha p}{\xi+p-\varepsilon} + \frac{\alpha\overline{p}}{\xi+\varepsilon-\overline{p}} - \frac{\alpha\overline{p}}{\xi+\overline{p}-\varepsilon}\right]\\ &= \frac{\alpha(\xi-p)}{4(\xi+p)} + \frac{\alpha(\xi+p)}{4(\xi-p)} + \frac{\alpha(\xi-\overline{p})}{4(\xi+\overline{p})} + \frac{\alpha(\xi+\overline{p})}{4(\xi-\overline{p})}.\end{aligned} \tag{5.24}$$

In the appendix we calculate the final form of $\omega(\xi)$, the area of the hole, and the strain in the void. It turns out that the loading can produce zero tangential stress along the smooth portions of the boundary only when $|\gamma/\alpha| \le C_1$, where $C_1 \approx 1.5$, i.e., when $1/C_2 > \sigma \ge C_2$, where $C_2 = (2-C_1)/(2+C_1) \approx 1/7$. Thus such holes are unlikely to be optimal when σ is negative. These calculations are presented in the appendix.

One reason why this analysis failed may be our assumption that σ_{tt} has simple delta function singularities at the four points, p, $-p$, $\overline{p}$, and $-\overline{p}$. When σ is negative, the tangential stress around the boundary of an optimal hole can take both positive and negative values. This is most easy to see when one minimizes the compliance energy $\boldsymbol{\sigma}_* : \boldsymbol{\mathcal{S}}_*\boldsymbol{\sigma}_*$. Cherkaev et al. (1998) have shown that the tangential stress changes sign at the four corner points but keeps constant magnitude around the boundary. Accordingly, the assumption that σ_{tt} is a positive valued measure around the boundary is probably too strong. Relaxing this constraint permits other forms of $\Phi(\xi)$, such as

$$\begin{aligned}\Phi(\xi) &= \frac{\alpha(\xi-p)}{4(\xi+p)} + \frac{\alpha(\xi+p)}{4(\xi-p)} + \frac{\alpha(\xi-\overline{p})}{4(\xi+\overline{p})} + \frac{\alpha(\xi+\overline{p})}{4(\xi-\overline{p})}\\ &\quad + \frac{i\beta(\xi-p)^2}{(\xi+p)^2} + \frac{i\beta(\xi+p)^2}{(\xi-p)^2} - \frac{i\beta(\xi-\overline{p})^2}{(\xi+\overline{p})^2} - \frac{i\beta(\xi+\overline{p})^2}{(\xi-\overline{p})^2},\end{aligned} \tag{5.25}$$

where β is an arbitrary real constant. This satisfies the symmetry constraints (5.23) and has zero real part except at the corner points. It would be interesting to see the hole shapes associated with this form of $\Phi(\xi)$. We did not do this, as the preliminary analysis was more involved than for the case $\beta = 0$ (treated in the appendix), which was already quite difficult.

Appendix. Here we calculate the final form of $\omega(\xi)$, the area of the hole, and the strain in the void when $\Phi(\xi)$ takes the form (5.24) and $p = e^{i\beta}$. These calculations are quite complicated and are best done with the aid of an algebraic manipulator, such as Maple. We follow the same procedure as used in section 5. We easily find that

$$\frac{1}{\pi i}\int_{|t|=1} \frac{\omega(t)\,\mathrm{Re}\,\Phi(t)}{(t-\xi)^2}\,dt = \frac{\alpha g}{2(\xi-p)^2} + \frac{\alpha g}{2(\xi+p)^2} + \frac{\alpha\overline{g}}{2(\xi-\overline{p})^2} + \frac{\alpha\overline{g}}{2(\xi+\overline{p})^2}, \tag{A.1}$$

where $g = p\omega(p)$. This leads to

(A.2)

$$\omega(\xi) = \frac{-g(i\xi^2+1/q)^2(i\xi^2+q)\xi^2 - \overline{g}(i\xi^2-q)^2(i\xi^2-1/q)\xi^2 + i(\gamma/\alpha)(i\xi^2-q)^2(i\xi^2+1/q)^2}{2\xi^3[(q-1/q)(\xi^4+1) - 4i\xi^2]},$$

in which $q = ip^2 = e^{i\tau}$ and $\tau = 2\beta + \pi/2$ have been introduced to simplify subsequent formulae. The denominator of $\omega(\xi)$ is zero when $\xi = 0$ and additionally when

$$\xi^2 = r_1 \equiv \frac{i(q+1)}{q-1} = \frac{1+\cos(\tau)}{\sin(\tau)} \quad \text{or} \quad \xi^2 = r_2 \equiv \frac{1}{r_1} = \frac{1-\cos(\tau)}{\sin(\tau)}. \tag{A.3}$$

Without loss of generality we can assume that $p = e^{i\beta}$ lies in the first quadrant with $\pi/2 \geq \beta \geq 0$, so that $3\pi/2 \geq \tau \geq \pi/2$. Then r_1 will lie inside the unit circle, while r_2 will lie outside it. In order for $\omega(\xi)$ to be analytic outside the unit disk, we require that the numerator of (A.2) be zero when $\xi^2 = r_2$. This imposes the constraint that

$$g(1-\overline{q})(q-\overline{q}+2) + \overline{g}(1-q)(\overline{q}-q+2) + \left(\frac{\gamma}{\alpha}\right)(q^2+\overline{q}^2+2) = 0, \tag{A.4}$$

in which $\overline{q} = 1/q$ is the complex conjugate of q. Additionally we have the constraint that

$$1 = \lim_{\xi\to\infty} \frac{\omega(\xi)}{\xi} = \frac{g + \overline{g} + \gamma/\alpha}{2i(\overline{q}-q)}. \tag{A.5}$$

These two real valued constraints can be used to solve for the complex constant g, giving

$$g = \frac{-2i(q-1)^2(q^2-2q-1) + (\gamma/\alpha)(3q^2-2q+1)}{(q-1)(q^2-4q+1)}, \tag{A.6}$$

which, when substituted back into the expression (A.2), gives

$$\omega(\xi) = \frac{2i\xi^2(q-1)^2(a_1\xi^4 + a_2\xi^2 + a_3) + i(\gamma/\alpha)q(b_1\xi^4 + b_2\xi^2 + b_3)}{2\xi^3 q(q-1)(q^2-4q+1)[i\xi^2(q-1)+q+1]} \tag{A.7}$$

with coefficients

$$\begin{aligned} &a_1 = q(q^2-4q+1), \quad a_2 = -i(q^4-1), \quad a_3 = q(q^2+4q+1), \\ &b_1 = 3q(3q^2-4q+3), \quad b_2 = i(q^2-1)(3q^2-8q+3), \quad b_3 = q(q^2-4q+1). \end{aligned} \tag{A.8}$$

Thus for a given loading, i.e., for given real values of α and γ, there is a one-parameter family of holes (parameterized by τ, with $q = e^{i\tau}$) such that the tangential stress is zero along the smooth portions of the boundary of each hole in the family. When $p = 1$, that is, $q = i$, the above expression for $\omega(\xi)$ reduces to (5.5), as it should. Also when $\gamma/\alpha = -4\sin(\tau)/3$, the expression for $\omega(\xi)$ reduces to

$$\omega(\xi) = \xi + 2\sin(\tau)\xi^{-1} - \frac{\xi^{-3}}{3}, \tag{A.9}$$

which corresponds to the inclusion generated from (5.5) with $k = -e^{2i\beta}/3$, rotated by an angle of $-\beta/2$. (One makes this substitution for k, multiplies the entire expression by $e^{-i\beta/2}$, and replaces ξ with $e^{i\beta/2}\xi$ to recover (A.9).) Another special case is for $q = -1$, and the expression for $\omega(\xi)$ reduces to

$$\omega(\xi) = \xi - \frac{\xi^{-3}}{3} - \left(\frac{\gamma}{\alpha}\right)\frac{(5\xi^{-1}+\xi^{-5})}{8}. \tag{A.10}$$

Of course there are restrictions on β which are necessary to ensure that the mapping $\omega(\xi)$ from the exterior of the unit disk onto its image is one-to-one. Differentiating (A.7) with respect to ξ gives

$$\omega'(\xi) = \frac{i(i\xi^2 q+1)(i\xi^2-q)(a\xi^4+b\xi^2+c)q^{3/2}}{\xi^4(q-1)(q^2-4q+1)[i\xi^2(q-1)+q+1]^2}, \tag{A.11}$$

where a, b, and c are the real valued coefficients

$$\begin{aligned}
a &= -i(q-1)^3(q^2-4q+1)q^{-5/2} = 2[\sin(5\tau/2) - 7\sin(3\tau/2) + 16\sin(\tau/2)],\\
b &= (q+1)(q-1)^2(q^2+4q+1)q^{-5/2} + 3i(\gamma/\alpha)(q-1)(3q^2-4q+3)]q^{-3/2}/2\\
&= 2[\cos(5\tau/2) + 3\cos(3\tau/2) - 4\cos(\tau/2)] - 3(\gamma/\alpha)[3\sin(3\tau/2) - 7\sin(\tau/2)],\\
c &= 3(\gamma/\alpha)(q+1)(q^2-4q+1)q^{-3/2}/2\\
&= 3(\gamma/\alpha)[\cos(3\tau/2) - 3\cos(\tau/2)].
\end{aligned} \tag{A.12}$$

The conformality constraint that $\omega'(\xi)$ be nonzero outside the unit disk is satisfied if and only if the quadratic $a\xi^4 + b\xi^2 + c$ has no root lying outside the unit disk, which holds if and only if

$$1 \geq \frac{c}{a} \geq \left|\frac{b}{a}\right| - 1. \tag{A.13}$$

Another restriction on the values that β can take arises from the constraints that $\omega(1)$ must lie on the positive real axis, and that $\omega(i)$ must lie on the positive imaginary axis, implying that

$$\begin{aligned}
0 &\leq 16[\sin(\tau/2)]^2[\sin(\tau/2) + \cos(\tau/2)]^3\\
&\quad +(\gamma/\alpha)[3\sin(3\tau/2) - 3\cos(3\tau/2) - 15\sin(\tau/2) + \cos(\tau/2)],\\
0 &\leq 16[\sin(\tau/2)]^2[\sin(\tau/2) - \cos(\tau/2)]^3\\
&\quad +(\gamma/\alpha)[-3\sin(3\tau/2) - 3\cos(3\tau/2) + 15\sin(\tau/2) + \cos(\tau/2)].
\end{aligned} \tag{A.14}$$

The constraints (A.13) and (A.14) confine the pair $(\gamma/\alpha, \tau)$ to lie within the region shown in Figure 14. One can see that this forces $|\gamma/\alpha|$ to be less than $C_1 \approx 1.5$. If $|\gamma/\alpha|$ is greater than this value, then the constraints cannot be satisfied for any choice of τ.

The potential ψ is obtained by substituting $\phi'(t) = \omega'(t)\Phi(t)$ into (4.20), using (5.24) and (A.7) and the fact that $\overline{\omega(t)} = \omega(\bar{t}) = \omega(1/t)$, and calculating the contour integral using the method of residues, giving

$$\psi = \frac{2i\xi\alpha(q-1)^2 s_1 + i\xi^3\gamma q s_2}{(q-1)(q^2-4q+1)(i\xi^2 q+1)(i\xi^2-q)}, \tag{A.15}$$

where

$$\begin{aligned}
s_1 &= -i\xi^2(q-1)(q^2+4q+1) + (q+1)(q^2-4q+1),\\
s_2 &= i\xi^2(q-1)(q^2-4q+1) + (q+1)(3q^2-4q+3).
\end{aligned} \tag{A.16}$$

The coefficients A and B appearing in the series expansions,

$$\phi = \alpha z + Az^{-1} + \mathcal{O}(z^{-3}), \quad \psi = \gamma z + Bz^{-1} + \mathcal{O}(z^{-3}), \tag{A.17}$$

of the potentials in powers of $1/z$ are found to be

(A.18)

$$\begin{aligned}
A &= -\lim_{\xi\to\infty} \xi^2[\omega(\xi)\Phi(\xi) - \alpha] - \alpha c_{-1} = i\alpha\left(q - \frac{1}{q}\right),\\
B &= \lim_{\xi\to\infty} \xi[\phi(\xi) - \gamma\xi] - \gamma c_{-1} = \frac{-4\alpha^2(q-1)^4(q^2+4q+1) - 3\gamma^2 q^2(3q^2-4q+3)}{2\alpha q(q-1)^2(q^2-4q+1)},
\end{aligned}$$

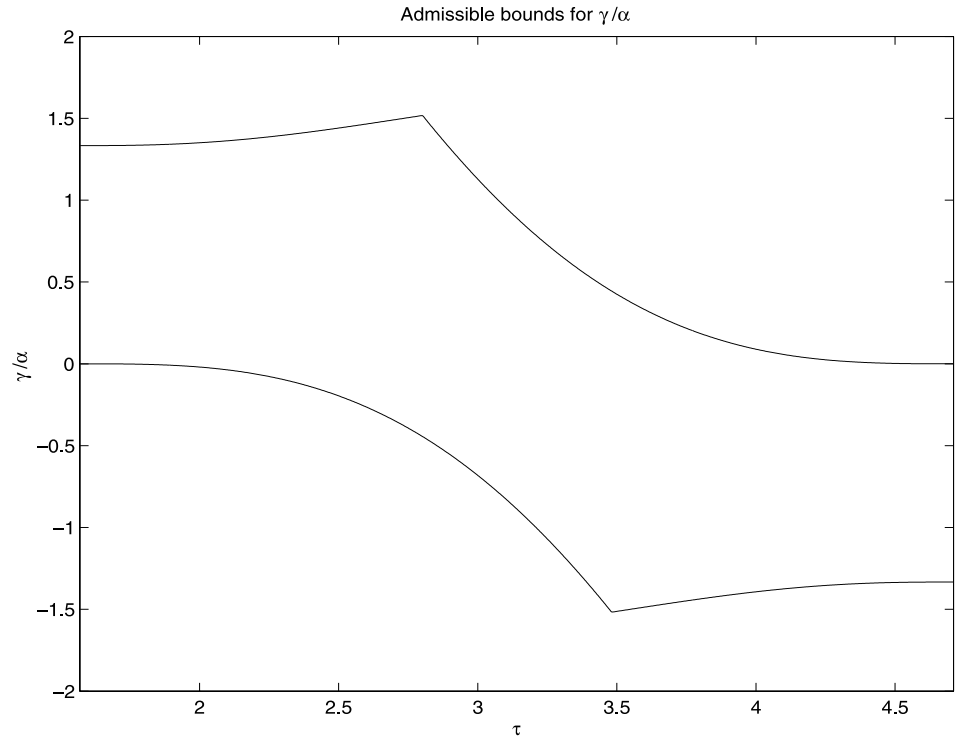


FIG. 14. *According to the constraints* (A.13) *and* (A.14), *the pair* $(\gamma/\alpha, \tau)$ *must lie within the region between these curves.*

where c_{-1} is the value of $\xi[\omega(\xi) - \xi]$ in the limit as $\xi \to \infty$. Substituting these expressions into (5.9) gives

$$\begin{aligned}
\epsilon_{11}^V &= \frac{4\alpha(1+i)(q-1)^3(q-i)^3 + 3(\gamma^2/\alpha)q^2(3q^2-4q+3)}{(V/\pi)q(q-1)^2(q^2-4q+1)} \\
&= \frac{8\alpha[1-\cos(\tau)-\sin(\tau)]^3 + 3(\gamma^2/\alpha)(3\cos(\tau)-2)}{(V/\pi)[5-6\cos(\tau)+\cos(2\tau)]}, \\
\epsilon_{22}^V &= \frac{4\alpha(1-i)(q-1)^3(q+i)^3 + 3(\gamma^2/\alpha)q^2(3q^2-4q+3)}{(V/\pi)q(q-1)^2(q^2-4q+1)} \\
&= \frac{8\alpha[1-\cos(\tau)+\sin(\tau)]^3 + 3(\gamma^2/\alpha)(3\cos(\tau)-2)}{(V/\pi)[5-6\cos(\tau)+\cos(2\tau)]},
\end{aligned} \tag{A.19}$$

in which V is the area of the inclusion. The area of the inclusion (assuming $\overline{\omega(\xi)} = \omega(\overline{\xi})$) is given by the contour integral

$$V = \frac{1}{2i}\int_{|\xi|=1} \overline{\omega(\xi)}\omega'(\xi)d\xi = \frac{1}{2i}\int_{|\xi|=1} \omega(1/\xi)\omega'(\xi)d\xi. \tag{A.20}$$

Substituting (A.7) and (A.11) into this formula and using the method of residues to evaluate the integral gives

$$\left(\frac{V}{\pi}\right) = v_1 + v_2\left(\frac{\gamma}{\alpha}\right) + v_3\left(\frac{\gamma}{\alpha}\right)^2, \tag{A.21}$$

where

$$
\begin{aligned}
v_1 &= \frac{-3(q^2+1)^3}{q(q^2-4q+1)^2} = \frac{-6[\cos(\tau)]^3}{[2-\cos(\tau)]^2},\\
v_2 &= \frac{-3i(q+1)(q^8-5q^7+17q^6-39q^5+48q^4-39q^3+17q^2-5q+1)}{q(q-1)^3(q^2-4q+1)^2}\\
&= \frac{3\cos(\tau/2)[\cos(4\tau)-5\cos(3\tau)+17\cos(2\tau)-39\cos(\tau)+24]}{8[\sin(\tau/2)]^3[2-\cos(\tau)]^2},\\
v_3 &= \frac{3(3q^8-3q^7-32q^6+87q^5-134q^4+87q^3-32q^2-3q+3)}{4(q-1)^4(q^2-4q+1)^2}\\
&= \frac{3[3\cos(4\tau)-3\cos(3\tau)-32\cos(2\tau)+87\cos(\tau)-67]}{128[\sin(\tau/2)]^4[2-\cos(\tau)]^2}. \qquad \text{(A.22)}
\end{aligned}
$$

Acknowledgments. We thank Valery Smyshlaev for useful and stimulating discussions, and the referee for helpful comments.

REFERENCES

G. Allaire and S. Aubry (1999), *On optimal microstructures for a plane shape optimization problem*, Structural Optim., 17, pp. 86–94.

G. Allaire and R. V. Kohn (1993a), *Optimal design for minimum weight and compliance in plane stress using extremal microstructures*, Eur. J. Mech. A Solids, 12, pp. 839–878.

G. Allaire and R. V. Kohn (1993b), *Explicit optimal bounds on the elastic energy of a two-phase composite in two space dimensions*, Quart. Appl. Math., 51, pp. 675–699.

M. Briane (1998), *Homogenization in some weakly connected domains*, Ricerche Mat., 47, pp. 51–94.

M. Briane and L. Mazliak (1998), *Homogenization of two randomly weakly connected materials*, Port. Math., 55, pp. 187–207.

A. V. Cherkaev (2000), *Variational Methods for Structural Optimization*, Appl. Math. Sci. 140, Springer-Verlag, Berlin.

A. V. Cherkaev, Y. Grabovsky, A. B. Movchan, and S. K. Serkov (1998), *The cavity of the optimal shape under the shear loading*, Int. J. Solids Structures, 35, pp. 4391–4410.

A. V. Cherkaev, K. A. Lurie, and G. W. Milton (1992), *Invariant properties of the stress in plane elasticity and equivalence classes of composites*, Proc. Roy. Soc. London A, 438, pp. 519–529.

A. V. Cherkaev and S. B. Vigdergauz (1986), *A hole in a plate, optimal for its biaxial extension-compression*, Appl. Math. Mech., 50, pp. 401–404.

A. R. Day, K. A. Snyder, E. J. Garboczi, and M. F. Thorpe (1992), *The elastic moduli of a sheet containing circular holes*, J. Mech. Phys. Solids, 40, pp. 1031–1051.

G. A. Francfort and F. Murat (1986), *Homogenization and optimal bounds in linear elasticity*, Arch. Ration. Mech. Anal., 94, pp. 161–177.

L. V. Gibiansky and A. V. Cherkaev (1984), *Design of composite plates of extremal rigidity*, Report 914, Ioffe Physicotechnical Institute, Leningrad; translated in *Topics in the Mathematical Modeling of Composite Materials*, A. Cherkaev and R. Kohn, eds., Progr. Nonlinear Differential Equations Appl. 31, Birkhäuser, Basel, Switzerland, pp. 95–137.

L. V. Gibiansky and A. V. Cherkaev (1987), *Microstructures of composites of extremal rigidity and exact bounds on the associated energy density*, Report 1115, Ioffe Physicotechnical Institute, Leningrad; translated in *Topics in the Mathematical Modeling of Composite Materials*, A. Cherkaev and R. Kohn, eds., Progr. Nonlinear Differential Equations Appl. 31, Birkhäuser, Basel, Switzerland, pp. 273–317.

L. V. Gibiansky, K. A. Lurie, and A. V. Cherkaev (1988), *Optimum focusing of a heat flux by a nonuniform heat-conducting medium (the "heat lens" problem)*, Zhurnal Tekhnicheskoi Fiziki, 58 (1988), pp. 67–74; translated in Sov. Phys. Tech. Phys., 33, pp. 38–42.

Y. Grabovsky and R. V. Kohn (1995), *Microstructures minimizing the energy of a two phase elastic composite in two space dimensions*. II: *The Vigdergauz microstructure*, J. Mech. Phys. Solids, 43, pp. 949–972.

G. K. Hu and G. J. Weng (2001), *A new derivative on the shift property of effective elastic compliances for planar and three-dimensional composites*, Proc. Roy. Soc. London A, 457, pp. 1675–1684.

M. Kachanov (1993), *Elastic solids with many cracks and related problems*, Adv. Appl. Mech., 30, pp. 259–445.

E. Ya. Khruslov (1979), *Asymptotic behavior of the solutions of the second boundary value problem in the case of the refinement of the boundary of the domain*, Matematicheskii Sbornik, 106 (1978), pp. 604–621; English translation in Math. USSR Sbornik, 35, pp. 266–282.

G. W. Milton (2002), *The Theory of Composites*, Cambridge University Press, Cambridge, UK.

G. W. Milton and A. V. Cherkaev (1995), *Which elasticity tensors are realizable?* ASME J. Engineering Materials Technol., 117, pp. 483–493.

G. W. Milton and S. K. Serkov (2000), *Bounding the current in nonlinear conducting composites*, J. Mech. Phys. Solids, 48, pp. 1295–1324.

A. B. Movchan and S. K. Serkov (1997), *The Pólya-Szegö matrices in asymptotic models of dilute composites*, European J. Appl. Math., 8, pp. 595–621.

F. Murat and L. Tartar (1985), *Calcul des variations et homogénísation*, in Les méthodes de l'homogénéisation: Théorie et applications en physique, Coll. de la Dir. des Études et Recherches de Électricité de France. Eyrolles, Paris, pp. 319–370; translated in *Topics in the Mathematical Modeling of Composite Materials*, A. Cherkaev and R. Kohn, eds., Progr. Nonlinear Differential Equations Appl. 31, Birkhäuser, Basel, Switzerland, pp. 139–173.

N. I. Muskhelishvili (1953), *Some Basic Problems in the Mathematical Theory of Elasticity*, Noordhoff, Groningen, The Netherlands.

W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling (1986), *Numerical Recipes. The Art of Scientific Computing*, Cambridge University Press, Cambridge, UK.

U. Ë. Raĭtum (1983), *Questions of the existence of a solution in problems of optimal control of leading coefficients of linear elliptic equations*, Differential Equations, 19, pp. 775–783.

S. K. Serkov (1998), *Asymptotic Analysis of Mathematical Models for Elastic Composite Media*, Ph.D. thesis, University of Bath, Bath, UK.

B. Shafiro and M. Kachanov (1999), *Solids with non-spherical cavities: Simplified representations of cavity compliance tensors and the overall anisotropy*, J. Mech. Phys. Solids, 47, pp. 877–898.

L. Tartar (1985), *Estimations fines des coefficients homogénéisés*, in Ennio De Giorgi's Colloquium, P. Krée, ed., Pitman Res. Notes Math. 125, Pitman Press, London, pp. 168–187.

L. Tartar (1995), *Remarks on the homogenization method in optimal design methods*, in Homogenization and Applications to Material Sciences, GAKUTO International Series, Math. Sci. Appl. 9, Gakkōtosho, Tokyo, pp. 393–412.

S. B. Vigdergauz (1986), *Effective elastic parameters of a plate with a regular system of equal-strength holes*, Inzhenernyi Zhurnal. Mekhanika Tverdogo Tela (MMT), 21, pp. 165–169.

S. B. Vigdergauz (1994), *Two-dimensional grained composites of extreme rigidity*, ASME J. Appl. Mech., 61, pp. 390–394.

S. B. Vigdergauz (1996), *Rhombic lattice of equi-stress inclusions in an elastic plate*, Quart. J. Mech. Appl. Math., 49, pp. 565–580.

S. B. Vigdergauz (1999), *Energy-minimizing inclusions in a planar elastic structure with macro-isotropy*, Structural Optim., 17, pp. 104-112.

Q.-S. Zheng and K. C. Hwang (1996), *Reduced dependence of defect compliance on matrix and inclusion properties in two-dimensional elasticity*, Proc. Roy. Soc. London A, 452, pp. 2493–2507.

Q.-S. Zheng and K. C. Hwang (1997), *Two-dimensional elastic compliances of materials with holes and microcracks*, Proc. Roy. Soc. London A, 453, pp. 353–364.

SIAM J. APPL. MATH.
Vol. 63, No. 3, pp. 1029–1055

DOUBLE HOPF BIFURCATIONS IN THE DIFFERENTIALLY HEATED ROTATING ANNULUS*

GREGORY M. LEWIS† AND WAYNE NAGATA‡

Abstract. We study a mathematical model of the differentially heated rotating fluid annulus experiment. In particular, we analyze the double Hopf bifurcations that occur along the transition between axisymmetric steady solutions and nonaxisymmetric rotating waves. The model uses the Navier–Stokes equations in the Boussinesq approximation. At the bifurcation points, center manifold reduction and normal form theory are used to deduce the local behavior of the full system of partial differential equations from a low-dimensional system of ordinary differential equations.

It is not possible to compute the relevant eigenvalues and eigenfunctions analytically. Therefore, the linear part of the equations is discretized, and the eigenvalues and eigenfunctions are approximated from the resulting matrix eigenvalue problem. However, the projection onto the center manifold and reduction to normal form can be done analytically. Thus, a combination of analytical and numerical methods is used to obtain numerical approximations of the normal form coefficients, from which the dynamics are deduced.

The results indicate that, close to the transition, there are regions in parameter space where there are multiple stable waves. Hysteresis of these waves is predicted. The validity of the results is shown by their consistency with experimental observations.

Key words. differentially heated rotating fluid experiment, axisymmetric to nonaxisymmetric transition, hysteresis of rotating waves, center manifold reduction, numerical approximation of normal form coefficients

AMS subject classifications. 37N10, 76U05, 37N99

PII. S0036139901386405

1. Introduction.

Laboratory experiments that isolate the effects of differential heating and rotation have long been regarded as useful tools for studying the behavior of large scale geophysical fluids, such as the atmosphere. The dynamic similarity of the various experiments to actual geophysical flows indicates that the form of the differential heating, the geometry of the system, the properties of the fluid, and the boundary conditions play a secondary role [13]. This is evidence that the character of large scale geophysical fluid flows is determined, to a large extent, by the differential heating and rotation. Consequently, the investigation of a mathematical model of a laboratory experiment itself can provide insight into the dynamical properties of large scale geophysical fluids. Furthermore, models of the experiments can be tractable, and the model and the method of analysis can be quantitatively validated via comparison with experimental observations. In contrast, a quantitative validation is not possible when studying direct, simplified models of large scale flows.

We study a model of a particular laboratory experiment in which the changes in the flow patterns in a differentially heated rotating annulus are observed as the imposed temperature gradient and rate of rotation are varied. We use an accurate

*Received by the editors March 14, 2001; accepted for publication (in revised form) June 4, 2002; published electronically February 25, 2003. This research was supported in part by the National Science and Engineering Research Council of Canada.

http://www.siam.org/journals/siap/63-3/38640.html

†Department of Mathematics, University of British Columbia, Vancouver, BC, V6T 1Z2 Canada. Present address: The Fields Institute, 222 College St., Toronto, ON, M5T 3J1 Canada (glewis@fields.utoronto.ca). The work of this author was supported in part by a fellowship from the Killam Trusts.

‡Department of Mathematics, University of British Columbia, Vancouver, BC, V6T 1Z2 Canada (nagata@math.ubc.ca).

mathematical model that is able to quantitatively reproduce some of the experimental observations [18], [19]. In the laboratory experiments, for small differential heating and rotation, a steady axisymmetric pattern is observed, i.e., the pattern is invariant under rotation. As the parameter values are increased, this relatively simple pattern becomes unstable and a wave motion appears. It is this transition from axisymmetric to nonaxisymmetric flow that is of interest here.

We study the transition by directly analyzing the partial differential equations (PDEs) that describe the fluid flow in the rotating annulus. In particular, we study the double Hopf bifurcations that are found at isolated points along the transition. These double Hopf bifurcation points occur when the linearization about the steady axisymmetric solution has two pairs of complex conjugate eigenvalues that simultaneously cross the imaginary axis as the parameters are varied. Center manifold reduction is used to find the dynamics of the PDEs close to the bifurcation point. This is a method of simplifying the equations in a way that takes into account all of the nonlinear interactions. The results are valid for parameter values close to the bifurcation point and when the bifurcating solutions are, in some sense, small. This method is sometimes referred to as weakly nonlinear analysis, because the nonlinear terms in the equations are assumed to be small but not negligible. Essentially, the technique is able to show the existence and stability of the bifurcating solutions and to give a first-order estimate of the solution itself, but it is not able to determine whether the solution persists for values of the parameters far from the bifurcation point.

This type of bifurcation analysis has been successful in other applications to fluid flow. One of the best known is the onset of motion in a layer of fluid heated from below, Rayleigh–Bénard convection (see, e.g., [24]). Another application of note is the Couette–Taylor problem (see [1] and the references contained therein), which is a fluid annulus experiment (without differential heating) where the inner and outer cylinders rotate at different rates, generating a shear flow in the fluid. A rich variety of behavior has been found by experiment, some of which can be explained with bifurcation theory. In addition, bifurcation analysis has made several predictions of flow patterns that were subsequently confirmed by experimental results. In the geophysical fluid dynamics literature, an asymptotic method, formally equivalent to center manifold reduction, was used to analyze "weakly nonlinear" wave-wave interactions (double Hopf bifurcations) in the two layer quasi-geostrophic potential vorticity equations in [20], [21], and [25] (see also [2] and [10]). The results indicated bistability and hysteresis of the wave solutions. For all of these models, it is possible to find the results analytically.

In the field of geophysical fluid dynamics, few models exist that can be studied purely analytically. Since the model we study does not fall into this category, we use an analytical-numerical hybrid analysis technique. Using center manifold reduction, it is possible to analytically reduce the time-dependent nonlinear PDEs to a series of steady linear PDE problems. These linear systems are then solved numerically, which results in numerical approximations for the coefficients of the normal form equations, from which the local dynamics can be deduced. Not only are the linear problems less difficult to numerically approximate, but also the validity of the approximations is more easily verified. Thus, although numerical approximations must be made, this method of analysis gives evidence that the predicted dynamics corresponds to those of the PDEs. Essentially equivalent methods are used in the Couette–Taylor problem [1] and in [7], where a double Hopf bifurcation was analyzed in a barotropic quasi-

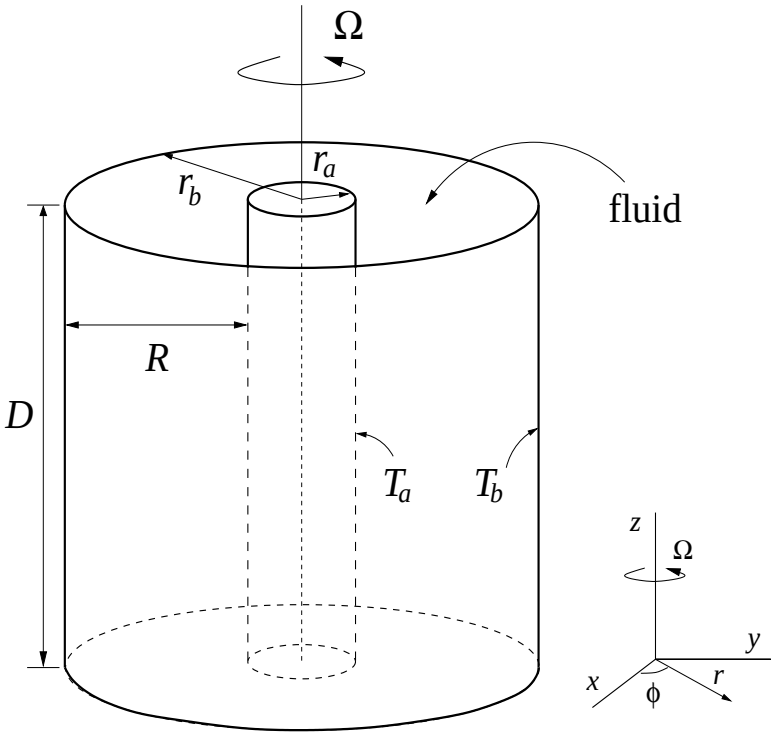


FIG. 1. *The differentially heated rotating annulus experiment, where the annulus is rotated at rate Ω and the inner wall is held at the fixed temperature T_a and the outer wall at temperature T_b, creating a differential heating $\Delta T = T_b - T_a$. Here r_a and r_b are the radii of the inner and outer cylinders, $R = r_b - r_a$, and D is the height of the annulus.*

geostrophic model. It should be noted that although similar methods were used in these problems, the numerics are substantially less intensive than those presented here. In fact, until recently, the numerics of this work would not have been possible on a personal computer.

In the next section, we describe the experiments in more detail. We discuss some general experimental results, and in so doing, introduce some of the flow features that our model reproduces. In section 3, the dynamical equations are written explicitly. The methods of analysis are discussed in the following two sections, where the analytical methods are presented in section 4 and the numerical methods are presented in section 5. In section 6, the results are described and discussed. A conclusion follows.

2. Experimental observations. Many different experiments have been performed in an attempt to develop an understanding of differentially heated rotating fluid systems (see, e.g., [13], [15], and [5]). The experiments often take the form of studying fluid flow in a rotating cylindrical annulus, where differential heating is obtained by maintaining the inner and outer walls of the annulus at different temperatures; see Figure 1. The experiments consist of finding the various stable flow patterns that occur at different values of the rotation rate and differential heating. The results are typically given in a diagram where the transitions between the different flow types are plotted in parameter space in terms of the Taylor number $\mathcal{T}$ and

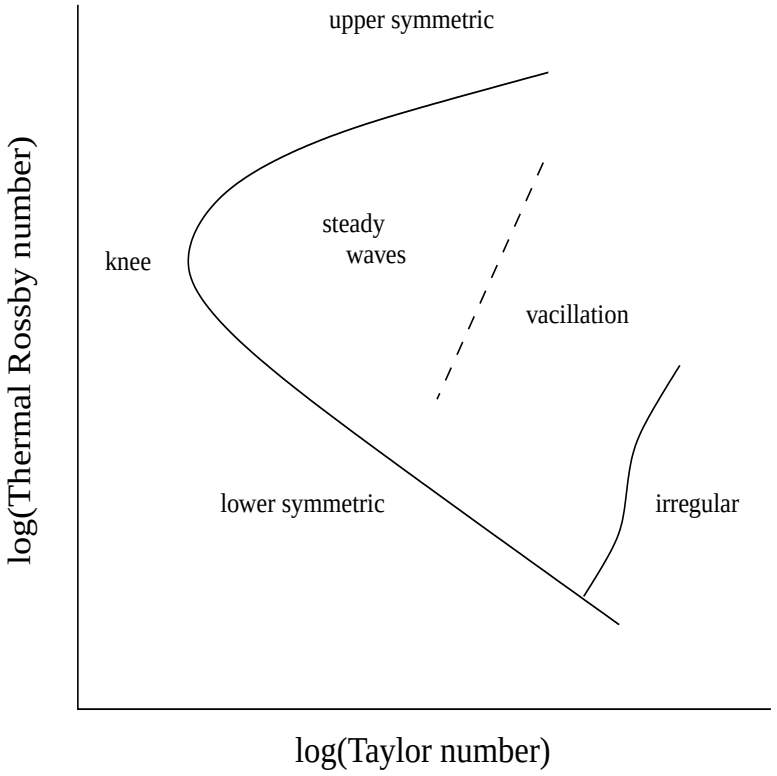


FIG. 2. *A schematic diagram depicting general experimental results; see, e.g.,* [13]. *To the left of all the curves is the axisymmetric regime, which is separated into three (dynamically similar) regions: lower symmetric, knee, and upper symmetric. To the right of the curve is the nonaxisymmetric regime, which is separated into three dynamically distinct regimes: steady waves, vacillation, and irregular flow.*

the thermal Rossby number $\mathcal{R}$ (see [4], [13]). The Taylor number

$$\mathcal{T} = \frac{4\Omega^2 R^4}{\nu^2}$$

is a dimensionless parameter measuring the relative importance of rotation to viscosity, where Ω is the rate of rotation, R is the gap width of the annulus, and ν is the kinematic viscosity of the fluid. The thermal Rossby number

$$\mathcal{R} = \frac{\alpha g D \Delta T}{\Omega^2 R^2}$$

is another dimensionless parameter measuring the relative importance of rotation to the differential heating, where ΔT is the difference in temperature between the inner and outer walls of the annulus, α is the coefficient of thermal expansion of the fluid, D is the depth, and g is the gravitational acceleration. If all other parameters are held fixed, there is a one-to-one relationship between these two dimensionless parameters and the two physical parameters that are varied during experiments: the differential heating ΔT and rate of rotation Ω.

Most of the experiments find four main flow regimes in different regions of parameter space (see Figure 2). (1) *Axisymmetric flow*: this flow is characterized by its

azimuthal invariance. (2) *Steady waves*: the flow in this region is nonaxisymmetric and resembles a rotating wave with constant amplitude and phase. Different wavelengths are seen in different subregions, with the possibility of observing stable waves of different wavelengths within the same subregion. The transitions between the subregions exhibit hysteresis. (3) *Vacillation*: in this region, the amplitude or structure of the observed wave varies apparently periodically in time. (4) *Irregular flow*: this region is characterized by its irregular nature in both space and time.

All of the observed flows have their counterparts in the Earth's atmosphere [6], [13]. The axisymmetric flow resembles the Hadley cell, which is observed in the atmosphere near the equator where the "local" rotation rate and differential heating are relatively small. In midlatitude regions of the Earth, the flow sometimes has wave characteristics that resemble the steady waves and vacillations seen in the experiments. Here, in both the atmosphere and experiments, the flow trajectories are curved, and vertical motion is inhibited.

Of particular interest to us are the transition from the axisymmetric to wave regime and the hysteresis of the waves which is observed in the steady wave regime. The hysteresis occurs between waves whose wave numbers differ by the integer one. By quantifying the double Hopf bifurcations that occur along the transition, we give evidence of the mechanism by which the hysteresis occurs.

3. Model equations. The dynamical equations of the fluid are taken to be the Navier–Stokes equations in the Boussinesq approximation. In particular, we consider the variations of all fluid properties to be negligible, and the equation of state of the fluid is assumed to be

$$\rho = \rho_0[1 - \alpha\,(T - T_0)], \tag{1}$$

where ρ is the density of the fluid, T is the temperature, α is the (constant) coefficient of thermal expansion, ρ_0 is the density at the reference temperature T_0, and $\alpha\,(T - T_0)$ is assumed to be small. A significant simplification due to the Boussinesq approximation is that the fluid can be considered incompressible. For the temperature evolution, we take the heat equation, with an advection term that couples the fluid velocity to the temperature. The boundaries are the inner wall of the cylindrical annulus with radius r_a, the outer wall with radius r_b, as well as a rigid flat bottom and top. At the boundaries, the no-slip condition is imposed on the fluid, and the temperature is T_a and T_b at the inner and outer walls, respectively, while the bottom and top are thermally insulating. The equations are written in circular cylindrical coordinates in a frame of reference corotating with the annulus at rate Ω. The radial, azimuthal, and vertical (or axial) coordinates are denoted r, φ, and z, respectively, with unit vectors $\mathbf{e}_r$, $\mathbf{e}_\varphi$, and $\mathbf{e}_z$ (see Figure 1).

We make a change of variables

$$r = Rr', \qquad z = Dz', \tag{2}$$

where $R = r_b - r_a$ is the gap width and D is the height of the annulus; write the fluid temperature as

$$T = T'(r', \varphi, z', t) + \Delta T\left(r' - \frac{r_a}{R}\right) + T_a, \tag{3}$$

where $\Delta T = T_b - T_a$ is the imposed temperature difference; and write the fluid pressure as

$$p = p'(r', \varphi, z', t) + \rho_0 g D(1 - z') + \frac{\rho_0 \Omega^2 R^2 (r')^2}{2}. \tag{4}$$

Then we drop the primes to obtain equations describing the evolution of the fluid velocity $\mathbf{u} = u(r,\varphi,z,t)\mathbf{e}_r + v(r,\varphi,z,t)\mathbf{e}_\varphi + w(r,\varphi,z,t)\mathbf{e}_z$, pressure deviation $p = p(r,\varphi,z,t)$, and temperature deviation $T = T(r,\varphi,z,t)$:

$$\frac{\partial \mathbf{u}}{\partial t} = \nu_s \nabla_s^2 \mathbf{u} - \frac{1}{R\rho_0}\nabla_s p - 2\Omega \mathbf{e}_z \times \mathbf{u} + \left(g\mathbf{e}_z - \Omega^2 R r \mathbf{e}_r\right)\alpha\left[T + \Delta T\left(r - \frac{r_a}{R}\right) + T_a - T_0\right] - \frac{1}{R}\left(\mathbf{u}\cdot\nabla_s\right)\mathbf{u}, \tag{5}$$

$$\frac{\partial T}{\partial t} = \kappa_s \nabla_s^2 T + \kappa_s \frac{\Delta T}{r} - \frac{\Delta T}{R}u - \frac{1}{R}\left(\mathbf{u}\cdot\nabla_s\right)T, \tag{6}$$

$$\nabla_s \cdot \mathbf{u} = \frac{\partial u}{\partial r} + \frac{u}{r} + \frac{\partial v}{\partial \varphi} + \frac{1}{\delta}\frac{\partial w}{\partial z} = 0, \tag{7}$$

where $\delta = D/R$, $\nu_s = \nu/R^2$, ν is the kinematic viscosity, $\kappa_s = \kappa/R^2$, κ is the coefficient of thermal diffusivity, g is the gravitational acceleration,

$$\nabla_s^2 = \frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}\frac{\partial^2}{\partial \varphi^2} + \frac{1}{\delta^2}\frac{\partial^2}{\partial z^2},$$

$$\nabla_s = \mathbf{e}_r\frac{\partial}{\partial r} + \mathbf{e}_\varphi\frac{1}{r}\frac{\partial}{\partial \varphi} + \mathbf{e}_z\frac{1}{\delta}\frac{\partial}{\partial z},$$

$$\begin{aligned}\left(\mathbf{u}_1\cdot\nabla_s\right)\mathbf{u}_2 = {} & \left(u_1\frac{\partial u_2}{\partial r} + \frac{v_1}{r}\frac{\partial u_2}{\partial \varphi} + \frac{1}{\delta}w_1\frac{\partial u_2}{\partial z} - \frac{v_1 v_2}{r}\right)\mathbf{e}_r \\ & + \left(u_1\frac{\partial v_2}{\partial r} + \frac{v_1}{r}\frac{\partial v_2}{\partial \varphi} + \frac{1}{\delta}w_1\frac{\partial v_2}{\partial z} + \frac{u_1 v_2}{r}\right)\mathbf{e}_\varphi \\ & + \left(u_1\frac{\partial w_2}{\partial r} + \frac{v_1}{r}\frac{\partial w_2}{\partial \varphi} + \frac{1}{\delta}w_1\frac{\partial w_2}{\partial z}\right)\mathbf{e}_z\end{aligned}$$

for velocity fields $\mathbf{u}_j = u_j(r,\varphi,z,t)\mathbf{e}_r + v_j(r,\varphi,z,t)\mathbf{e}_\varphi + w_j(r,\varphi,z,t)\mathbf{e}_z$, $j = 1, 2$, and

$$\left(\mathbf{u}\cdot\nabla_s\right)T = u\frac{\partial T}{\partial r} + \frac{v}{r}\frac{\partial T}{\partial \varphi} + \frac{1}{\delta}w\frac{\partial T}{\partial z}.$$

The domain is $r_a/R < r < r_b/R$, $0 \le \varphi < 2\pi$, $0 < z < 1$, and the boundary conditions are

$$\begin{aligned}\mathbf{u} &= 0 \quad \text{on} \quad r = \frac{r_a}{R},\ \frac{r_b}{R} \ \text{ and } \ z = 0, 1, \\ T &= 0 \quad \text{on} \quad r = \frac{r_a}{R},\ \frac{r_b}{R}, \\ \frac{\partial T}{\partial z} &= 0 \quad \text{on} \quad z = 0, 1,\end{aligned} \tag{8}$$

with 2π-periodicity in φ for $\mathbf{u}$, T, and p.

The solutions will not depend explicitly on the value of the reference temperature T_0. However, there is implicit dependence because the values of ν, κ, and ρ_0 are chosen to be those of the fluid at T_0. It is assumed that the difference between the temperature of the fluid and T_0 is everywhere small enough so that ν and κ can be considered as constants.

4. Analytical methods. We choose as the parameters of interest the rotation rate Ω and the temperature difference ΔT between the inner and outer annulus walls. These are the physical quantities (external variables) that are easily varied in an experiment. The other parameters describe the geometry of the annulus or properties of the fluid. Another choice is to use the dimensionless parameters, the Taylor number $\mathcal{T}$ and the thermal Rossby number $\mathcal{R}$ (see section 2), which have a one-to-one correspondence with Ω and ΔT. Our results are quoted in terms of these dimensionless parameters because experimental results are usually presented on a log-log plot of $\mathcal{T}$ versus $\mathcal{R}$. However, the analysis was carried out using the parameters Ω and ΔT, because nondimensionalization did not significantly simplify the equations (see [18]). The choice of parameters will not change the procedure or the results.

A summary of the main steps of the analysis is as follows:

1. Plot the neutral stability curves, by
   (a) calculating the steady axisymmetric solution at a particular location in parameter space,
   (b) solving the eigenvalue problem for this solution to find its linear stability,
   (c) repeating steps (a) and (b) at various locations in parameter space to find the parameter values at which the solution is neutrally stable.
2. Localize the point in parameter space where the double Hopf bifurcation occurs (find the intersections of the neutral stability curves; see below).
3. Calculate the eigenvalues and eigenfunctions at the bifurcation point.
4. Compute the appropriate normal form coefficients, which involves
   (a) calculating the adjoint eigenfunctions,
   (b) calculating the center manifold coefficients.

An analytical form for the steady axisymmetric solution is not known, and, therefore, numerical approximations must be made. This is also the case for the eigenvalues and eigenfunctions. In the analysis, this is dealt with by leaving the unknown functions unresolved when deriving the formulae for the normal form coefficients. That is, we write the normal form coefficients in terms of the unresolved functions. Then, for the numerical approximation of the normal form coefficients, the values of the unknown functions are needed only at specific locations (the grid points), and numerical approximations are used. We postpone discussion of the numerical methods until the next section, and for the remainder of this section we discuss the analytical methods. In particular, we discuss the equations necessary for the computation of the axisymmetric solution and the eigenfunctions, and we discuss briefly how center manifold reduction is used to derive the formulae for the normal form coefficients of interest. For a more detailed explanation of the center manifold reduction and normal form equations in the context of this model, see [18], and for a general context, see, e.g., [11].

4.1. The steady axisymmetric solution. The analysis begins with the computation of a steady axisymmetric solution. That is, we look for a solution of (5)–(7), with the boundary conditions (8), in the form

$$\mathbf{u} = \mathbf{u}^{(0)}(r,z) = u^{(0)}(r,z)\mathbf{e}_r + v^{(0)}(r,z)\mathbf{e}_\varphi + w^{(0)}(r,z)\mathbf{e}_z,$$

$$p = p^{(0)}(r,z), \qquad T = T^{(0)}(r,z),$$

independent of φ and t. The solution also depends on the parameters, but we do not indicate this dependence explicitly. We assume that such a solution exists, is unique and regular, and depends smoothly on the parameters, at least for the parameter

values of interest. We have not attempted to prove this, but we believe that, using standard techniques, it would be a feasible if lengthy digression to do so (see, for example [3]).

A stream function $\xi^{(0)}$ is introduced, so that the incompressibility condition (7) is automatically satisfied. The pressure terms can then be eliminated, and we obtain three equations in the three unknown functions $\xi^{(0)}$, $v^{(0)}$, and $T^{(0)}$. The resulting equations, computed using the Maple symbolic computation package, are sufficiently complicated that no insight is gained by explicitly writing them here. For more details, see [18]. In section 5, we describe how $\xi^{(0)}$, $v^{(0)}$, and $T^{(0)}$ are computed numerically.

4.2. The perturbation equations. Next, perturbation equations are required. It is on this system that the center manifold reduction is performed. We write

$$\mathbf{u} = \mathbf{u}^{(0)} + \hat{\mathbf{u}}, \quad p = p^{(0)} + \hat{p}, \quad T = T^{(0)} + \hat{T}, \tag{9}$$

where $(\mathbf{u}^{(0)}, p^{(0)}, T^{(0)})$ is the steady axisymmetric solution, substitute (9) into (5)–(7), and drop the hats, to obtain the perturbation equations

$$\begin{aligned}\frac{\partial \mathbf{u}}{\partial t} = {} & \nu_s \nabla_s^2 \mathbf{u} - \frac{1}{R\rho_0} \nabla_s p - 2\Omega \mathbf{e}_z \times \mathbf{u} + \left(g\mathbf{e}_z - \Omega^2 R r \mathbf{e}_r\right) \alpha T \\ & - \frac{1}{R}\left(\mathbf{u}^{(0)} \cdot \nabla_s\right)\mathbf{u} - \frac{1}{R}\left(\mathbf{u} \cdot \nabla_s\right)\mathbf{u}^{(0)} - \frac{1}{R}\left(\mathbf{u} \cdot \nabla_s\right)\mathbf{u},\end{aligned} \tag{10}$$

$$\frac{\partial T}{\partial t} = \kappa_s \nabla_s^2 T - \frac{\Delta T}{R} u - \frac{1}{R}\left(\mathbf{u}^{(0)} \cdot \nabla_s\right) T - \frac{1}{R}\left(\mathbf{u} \cdot \nabla_s\right) T^{(0)} - \frac{1}{R}\left(\mathbf{u} \cdot \nabla_s\right) T, \tag{11}$$

$$\nabla_s \cdot \mathbf{u} = 0, \tag{12}$$

with the boundary conditions (8). The trivial solution $\mathbf{u} = \mathbf{0}$, $p = 0$, $T = 0$ now satisfies these equations and corresponds to the steady axisymmetric solution of (5)–(7).

The perturbation equations (10)–(12) can be put into a suitable abstract form for which some important theoretical properties have been established. Following Henry [12, pp. 79–81], we can define a space $\mathcal{X}$ of vector functions $U = [\mathbf{u}, T]$ so that the incompressibility condition (12) and boundary conditions (8) are satisfied as part of the definition of the space. Then there is an abstract projection operator onto the space $\mathcal{X}$ that eliminates the pressure terms, and the system (10)–(12) together with boundary conditions (8) can be written as an abstract evolution equation in the space $\mathcal{X}$,

$$\dot{U} = \mathbf{L}\, U + \mathbf{N}(U), \tag{13}$$

where $\mathbf{L}\, U$ is the linear part of the equation (observe that $\mathbf{L}$ depends on the parameters, through the steady axisymmetric solution), and $\mathbf{N}(U)$ is the nonlinear part (it has the form $\mathbf{N}(U) = \mathbf{M}(U, U)$, where $\mathbf{M}$ is bilinear). If we assume that the steady axisymmetric solution of (5)–(8) exists, is unique, is regular, and depends smoothly on the parameters, then at least locally near $U = 0$ the initial-value problem for (13) in $\mathcal{X}$ has a unique solution $U(t)$, $t \geq 0$, that depends smoothly on initial conditions and parameters [12, Chapter 3]. Moreover, the principle of linearized stability holds, and the stability of the trivial solution $U = 0$ of (13) can be determined from the spectrum of the linearization $\mathbf{L}$ [12, Chapter 5].

4.3. The eigenvalue problem. The linearized stability of the steady axisymmetric solution is determined by the spectrum of the linearization of (10)–(12) about the trivial solution. Since $\mathbf{L}$ is the sum of a self-adjoint operator and a bounded linear operator, it is sectorial. The spatial domain is bounded, so the spectrum consists entirely of isolated eigenvalues of finite multiplicity. The eigenvalue problem is formally obtained by assuming that the unknown functions may be written as $\mathbf{u} = \mathbf{u}(r, \varphi, z, t) = e^{\lambda t}\tilde{\mathbf{u}}_m(r, z)e^{im\varphi}$, with m an integer, and likewise for T and p, and then linearizing (10)–(12). A linear eigenvalue problem for the eigenvalues λ and the eigenfunctions $[\tilde{\mathbf{u}}_m(r, z), \tilde{T}_m(r, z)]e^{im\varphi}$ is obtained for each azimuthal wave number m. By the principle of linearized stability, if all eigenvalues λ have negative real parts, then the steady axisymmetric solution is asymptotically stable, while if any eigenvalue λ has a positive real part, then the steady axisymmetric solution is unstable. We are especially interested in locating critical parameter values, where a finite number of eigenvalues have zero real parts and the rest have negative real parts. The solution is then neutrally stable, and we expect a bifurcation of solutions of the nonlinear equations as parameters are varied near the critical values. The azimuthal wave numbers m of the eigenfunctions corresponding to the eigenvalues that have zero real part at the critical parameter values are defined as the critical wave numbers.

If $m \neq 0$, it is possible to eliminate the pressure and azimuthal velocity terms. The resulting three equations in the three remaining unknowns $\tilde{u}_m(r, z)$, $\tilde{w}_m(r, z)$, and $\tilde{T}_m(r, z)$ may be written in the form of a generalized linear eigenvalue problem

$$\lambda \mathbf{A}_m \tilde{U}_m = \mathbf{L}_m \tilde{U}_m, \tag{14}$$

where

$$\tilde{U}_m = \begin{pmatrix} \tilde{u}_m \\ \tilde{w}_m \\ \tilde{T}_m \end{pmatrix}$$

and $\mathbf{A}_m$ and $\mathbf{L}_m$ are 3×3 matrices of linear operators. If $m = 0$, a stream function method can be used in exactly the same manner as in the calculation of the axisymmetric solution. Again the equations were calculated using Maple and are too lengthy to write here.

Finally, the adjoint eigenvalue problem is necessary to calculate the adjoint eigenfunctions. The adjoint operators are calculated using the inner product, which for two vector functions $U_1 = [\mathbf{u}_1, T_1]$ and $U_2 = [\mathbf{u}_2, T_2]$ is taken to be

$$\langle U_1, U_2 \rangle = \int_0^1 \int_0^{2\pi} \int_{\frac{r_a}{R}}^{\frac{r_b}{R}} \left(\mathbf{u}_1 \cdot \overline{\mathbf{u}}_2 + T_1 \overline{T}_2 \right) r \; dr \; d\varphi \; dz, \tag{15}$$

where the overbar denotes complex conjugation. The adjoint eigenfunctions have the form $[\tilde{\mathbf{u}}^*_m(r, z), \tilde{T}^*_m(r, z)]e^{im\varphi}$.

4.4. Normal form coefficients. The numerical results, presented in section 6, predict that there are critical parameter values at which the linear eigenvalue problem has two complex conjugate pairs of eigenvalues with zero real parts, while the other eigenvalues have negative real parts. Therefore, suppose that the critical parameter values occur at $\Omega = \Omega_0$ and $\Delta T = \Delta T_0$, so that for Ω near Ω_0 and ΔT near ΔT_0 the eigenvalue problem has eigenvalues

$$\lambda_1 = \mu_1 + i\omega_1, \quad \overline{\lambda}_1, \quad \lambda_2 = \mu_2 + i\omega_2, \quad \overline{\lambda}_2, \tag{16}$$

and when $\Omega = \Omega_0$ and $\Delta T = \Delta T_0$, we have $\mu_1 = \mu_2 = 0$. Also, assume that all the other eigenvalues have negative real parts, with the real parts uniformly bounded below zero.

The eigenfunctions corresponding to the above eigenvalues are

$$\Phi_1, \quad \overline{\Phi}_1, \quad \Phi_2, \quad \overline{\Phi}_2,$$

where they have the form

$$\Phi_j = [\tilde{\mathbf{u}}_{m_j}(r,z), \tilde{T}_{m_j}(r,z)]e^{im_j\varphi},$$

with m_j ($j = 1, 2$, $m_1 \neq m_2$) being the azimuthal wave number corresponding to Φ_j. The center eigenspace E^c is the span of the eigenfunctions corresponding to the eigenvalues with zero real parts when $\Omega = \Omega_0$ and $\Delta T = \Delta T_0$,

$$E^c = \text{span}\{\Phi_1, \overline{\Phi}_1, \Phi_2, \overline{\Phi}_2\}.$$

The stable eigenspace E^s is the span of all the other eigenfunctions, which are the eigenfunctions that correspond to eigenvalues with negative real parts. The adjoint eigenfunctions corresponding to the Φ_j are denoted by Φ_j^*, where the Φ_j^* are found from the adjoint eigenvalue problem. The eigenfunctions and their adjoints are normalized so that their inner products satisfy

$$\langle \Phi_j, \Phi_j^* \rangle = 1 \tag{17}$$

for $j = 1, 2$. Due to a rescaling (see below), the results do not depend on the way in which the second normalization constant is determined.

The projection of U onto the center eigenspace E^c is given by

$$PU = \langle U, \Phi_1^* \rangle \Phi_1 + \langle U, \overline{\Phi}_1^* \rangle \overline{\Phi}_1 + \langle U, \Phi_2^* \rangle \Phi_2 + \langle U, \overline{\Phi}_2^* \rangle \overline{\Phi}_2. \tag{18}$$

Using this projection, we may then decompose U as follows:

$$U = z_1\Phi_1 + \overline{z}_1\overline{\Phi}_1 + z_2\Phi_2 + \overline{z}_2\overline{\Phi}_2 + \Psi, \tag{19}$$

where $PU = z_1\Phi_1 + \overline{z}_1\overline{\Phi}_1 + z_2\Phi_2 + \overline{z}_2\overline{\Phi}_2 \in E^c$, $(I - P)U = \Psi \in E^s$, and I is the identity operator. This implies that the complex amplitudes z_1 and z_2 are given by the inner products $z_1 = z_1(t) = \langle U, \Phi_1^* \rangle$ and $z_2 = z_2(t) = \langle U, \Phi_2^* \rangle$.

Taking the projection of (13), we get

$$\begin{aligned} \dot{z}_1 &= \lambda_1 z_1 + \langle \mathbf{N}(U), \Phi_1^* \rangle, \\ \dot{z}_2 &= \lambda_2 z_2 + \langle \mathbf{N}(U), \Phi_2^* \rangle, \end{aligned} \tag{20}$$

where U is given by (19). The complex conjugate equations contain redundant information, and so are omitted. For (20), we use center manifold theory to write U solely in terms of the center eigenspace variables z_1 and z_2, and in so doing we decouple the system.

Given that the assumptions stated above for (13) and the eigenvalues for the linearization are valid, then, for $(\Omega, \Delta T)$ in a neighborhood of $(\Omega_0, \Delta T_0)$, the center manifold theorem [12, p. 168] implies that there exists a differentiable center manifold for (13):

$$W_{loc}^c = \left\{ U = z_1\Phi_1 + \overline{z}_1\overline{\Phi}_1 + z_2\Phi_2 + \overline{z}_2\overline{\Phi}_2 + H(z_1\Phi_1, \overline{z}_1\overline{\Phi}_1, z_2\Phi_2, \overline{z}_2\overline{\Phi}_2) \right\}, \tag{21}$$

where $H : E^c \to E^s$ is defined for $\|z_1\Phi_1 + \overline{z}_1\overline{\Phi}_1 + z_2\Phi_2 + \overline{z}_2\overline{\Phi}_2\|$ small and $\|\cdot\|$ is the norm that corresponds to the inner product (15). The local center manifold W^c_{loc} is locally invariant, is tangent to the center eigenspace E^c at $U = 0$ when $\Omega = \Omega_0$ and $\Delta T = \Delta T_0$, and is locally exponentially attracting.

Therefore, on the center manifold, we can write

$$\Psi = H(z_1, \overline{z}_1, z_2, \overline{z}_2) = O(|z_1, \overline{z}_1, z_2, \overline{z}_2|^2) \tag{22}$$

and then expand the center manifold function H in a Taylor series as

$$\begin{aligned} H(z_1, \overline{z}_1, z_2, \overline{z}_2) = {} & H_{2000}z_1^2 + H_{1100}z_1\overline{z}_1 + H_{0020}z_2^2 + H_{0011}z_2\overline{z}_2 \\ & + H_{1010}z_1z_2 + H_{1001}z_1\overline{z}_2 + c.c. + O(3), \end{aligned} \tag{23}$$

where H_{ijkl} are the Taylor series coefficients of H, $O(n) = O(|z_1, \overline{z}_1, z_2, \overline{z}_2|^n)$, and *c.c.* denotes the complex conjugates of the previous terms that are written explicitly. We also write

$$\begin{aligned} N(z_1, \overline{z}_1, z_2, \overline{z}_2) = {} & N_{2000}z_1^2 + N_{1100}z_1\overline{z}_1 + N_{0020}z_2^2 + N_{0011}z_2\overline{z}_2 \\ & + N_{1010}z_1z_2 + N_{1001}z_1\overline{z}_2 + c.c. + O(3), \end{aligned} \tag{24}$$

where $N(z_1, \overline{z}_1, z_2, \overline{z}_2)$ is the nonlinear term of (13) written in terms of $z_1, \overline{z}_1, z_2$, and $\overline{z}_2$, using the decomposition of U given in (19), and with Ψ written using (22) and (23). With the nonlinear part written as (24), the system is decoupled, and (20) reduces to a four-dimensional ODE that describes the dynamics on the center manifold. Because the center manifold is locally exponentially attracting, the behavior of the original PDEs, close to the bifurcation point, can be deduced from the reduced system.

The normal form for the nonresonant case is

$$\begin{aligned} \dot{z}_1 &= \lambda_1 z_1 + G_{11}z_1^2\overline{z}_1 + G_{12}z_1z_2\overline{z}_2 + O(4), \\ \dot{z}_2 &= \lambda_2 z_2 + G_{21}z_1\overline{z}_1z_2 + G_{22}z_2^2\overline{z}_2 + O(4), \end{aligned} \tag{25}$$

where $\lambda_j{=}\lambda_j(\Omega, \Delta T)$, and the normal form coefficients G_{kl} are given by

$$\begin{aligned} G_{11} &= \langle N_{2100}, \Phi_1^* \rangle, \\ G_{12} &= \langle N_{1011}, \Phi_1^* \rangle, \\ G_{21} &= \langle N_{1110}, \Phi_2^* \rangle, \\ G_{22} &= \langle N_{0021}, \Phi_2^* \rangle. \end{aligned} \tag{26}$$

The normal form (25) is obtained from (20) by using a series of near-identity coordinate transformations (see, e.g., [26]). This normal form requires the nonresonance condition that the imaginary parts of the eigenvalues, ω_1 and ω_2, satisfy $n_1\omega_1 + n_2\omega_2 \neq 0$ for all integers n_1 and n_2 with $|n_1| + |n_2| \leq 4$ at the critical parameter values $\Omega = \Omega_0$ and $\Delta T = \Delta T_0$.

In general, the formulae (26) for the normal form coefficients also depend on the coefficients of the terms that are quadratic in z_1 and z_2 (e.g., N_{2000}). However, in our case these terms vanish in the projection (20) because, due to their φ-dependence, they are orthogonal to the adjoint eigenfunctions. In the same manner it can be shown that to find the normal form coefficient $G_{11} = \langle N_{2100}, \Phi_1^* \rangle$, only $\tilde{N}_{2100}^{(m_1)}$ is needed, where $\tilde{N}_{ijkl}^{(m_1)}$ is defined as the coefficient of $e^{im_1\varphi}$ in the expansion $N_{ijkl}(r, \varphi, z) = \sum_m \tilde{N}_{ijkl}^{(m)}(r, z)\, e^{im\varphi}$. That is, all terms with a factor $e^{im\varphi}$, $m \neq m_1$, vanish in the

inner product because they are orthogonal to Φ_1^*. Furthermore, due to the form of the nonlinear part, only the eigenfunctions Φ_1, $\overline{\Phi}_1$ and the particular coefficients of the center manifold function, $\tilde{H}_{1100}^{(0)}$ and $\tilde{H}_{2000}^{(2m_1)}$, appear in the formula for $\tilde{N}_{2100}^{(m_1)}$, where the $\tilde{H}_{ijkl}^{(m)}$ are defined in a manner similar to the $\tilde{N}_{ijkl}^{(m)}$, i.e., $H_{ijkl}(r,\varphi,z) = \sum_m \tilde{H}_{ijkl}^{(m)}(r,z)\, e^{im\varphi}$.

Thus, in addition to the eigenfunctions, the normal form coefficient

- G_{11} can be written as a function of only the coefficients $\tilde{H}_{1100}^{(0)}$ and $\tilde{H}_{2000}^{(2m_1)}$.

Similarly, in addition to the eigenfunctions, the normal form coefficients

- G_{12} can be written as a function of only the coefficients, $\tilde{H}_{0011}^{(0)}$, $\tilde{H}_{1001}^{(m_1-m_2)}$, and $\tilde{H}_{1010}^{(m_1+m_2)}$;
- G_{21} can be written as a function of only the coefficients $\tilde{H}_{1100}^{(0)}$, $\tilde{H}_{0110}^{(m_2-m_1)}$, and $\tilde{H}_{1010}^{(m_1+m_2)}$;
- G_{22} can be written as a function of only the coefficients $\tilde{H}_{0011}^{(0)}$ and $\tilde{H}_{0020}^{(2m_2)}$.

The equations satisfied by the $H_{ijkl}(r,\varphi,z)$ are derived using the local invariance of the center manifold (see [11]). From these equations, it follows that the relevant $\tilde{H}_{ijkl}^{(m)}(r,z)$ satisfy

$$
\begin{aligned}
\left[2\lambda_1\mathbf{I} - \tilde{\mathbf{L}}^{(2m_1)}\right]\tilde{H}_{2000}^{(2m_1)} &= \tilde{N}_{2000}^{(2m_1)},\\
\tilde{\mathbf{L}}^{(0)}\tilde{H}_{1100}^{(0)} &= -\tilde{N}_{1100}^{(0)},\\
\left[2\lambda_2\mathbf{I} - \tilde{\mathbf{L}}^{(2m_2)}\right]\tilde{H}_{0020}^{(2m_2)} &= \tilde{N}_{0020}^{(2m_2)},\\
\tilde{\mathbf{L}}^{(0)}\tilde{H}_{0011}^{(0)} &= -\tilde{N}_{0011}^{(0)},\\
\left[(\lambda_1+\lambda_2)\,\mathbf{I} - \tilde{\mathbf{L}}^{(m_1+m_2)}\right]\tilde{H}_{1010}^{(m_1+m_2)} &= \tilde{N}_{1010}^{(m_1+m_2)},\\
\left[\left(\lambda_1+\overline{\lambda}_2\right)\mathbf{I} - \tilde{\mathbf{L}}^{(m_1-m_2)}\right]\tilde{H}_{1001}^{(m_1-m_2)} &= \tilde{N}_{1001}^{(m_1-m_2)},
\end{aligned}
\tag{27}
$$

where the $\tilde{\mathbf{L}}^{(m)}$ are defined by $\mathbf{L}[\tilde{U}(r,z)e^{im\varphi}] = e^{im\varphi}[\tilde{\mathbf{L}}^{(m)}\tilde{U}(r,z)]$ and $\mathbf{I}$ is the identity operator. For $m \neq 0$, the same solution method that is used for the eigenvalue problem can be used here (i.e., elimination of the pressure term and one velocity component). For $m = 0$, the stream function method (as for the axisymmetric solution) can be used.

We write $z_1 = \rho_1 e^{i\theta_1}/\sqrt{|G_{11}^r|}$ and $z_2 = \rho_2 e^{i\theta_2}/\sqrt{|G_{22}^r|}$, where G_{ij}^r is the real part of the normal form coefficients G_{ij}, and substitute these expressions into (25). In these scaled polar coordinates, the truncated normal form equations are

$$
\begin{aligned}
\dot{\rho}_1 &= \rho_1\left(\mu_1 + a\rho_1^2 + b\rho_2^2\right),\\
\dot{\rho}_2 &= \rho_2\left(\mu_2 + c\rho_1^2 + d\rho_2^2\right),\\
\dot{\theta}_1 &= \omega_1,\\
\dot{\theta}_2 &= \omega_2,
\end{aligned}
\tag{28}
$$

where

$$
\begin{aligned}
a &= \frac{G_{11}^r}{|G_{11}^r|} = \pm 1,\\
b &= \frac{G_{12}^r}{|G_{22}^r|},\\
c &= \frac{G_{21}^r}{|G_{11}^r|},
\end{aligned}
\tag{29}
$$

$$d = \frac{G_{22}^r}{|G_{22}^r|} = \pm 1,$$

and $\lambda_j = \mu_j + i\omega_j$. The $O(|\rho_1, \rho_2|^4)$ terms are ignored in the $\dot{\rho}_j$ equations, and the $O(|\rho_1, \rho_2|^2)$ terms are ignored in the $\dot{\theta}_j$ equations. Ignoring these terms does not affect the local dynamics, except for fine details of the dynamics on invariant tori.

In summary, given m_1 and m_2, the coefficients of the scaled normal form equations a, b, c, d can be written in terms of the following functions, which are all functions only of the two spatial variables r and z:

- the eigenfunctions and adjoint eigenfunctions

$$\tilde{\Phi}_{m_1}, \quad \tilde{\Phi}^*_{m_1}, \quad \tilde{\Phi}_{m_2}, \quad \tilde{\Phi}^*_{m_2},$$

  where $\Phi_j(r, \varphi, z) = \tilde{\Phi}_{m_j}(r, z)e^{im_j\varphi}$, $j = 1, 2$;
- certain Taylor series coefficients of the center manifold function

$$\tilde{H}^{(0)}_{1100}, \quad \tilde{H}^{(2m_1)}_{2000}, \quad \tilde{H}^{(m_1-m_2)}_{1001}, \quad \tilde{H}^{(m_1+m_2)}_{1010}, \quad \tilde{H}^{(0)}_{0011}, \quad \text{and} \quad \tilde{H}^{(2m_2)}_{0020},$$

where $H_{ijkl}(r, \varphi, z) = \sum_m \tilde{H}^{(m)}_{ijkl}(r, z)e^{im\varphi}$. The eigenfunctions are found from the eigenvalue problem (14), and the coefficients of the center manifold function are found from (27).

5. Numerical methods. In order to find values of the normal form coefficients, the axisymmetric solution, the eigenfunctions, and the Taylor series coefficients of the center manifold function must be known. Because analytic solutions of these are not known, they are approximated numerically. Upon discretization, the axisymmetric solution is approximated from a system of nonlinear algebraic equations, while the partial differential eigenvalue problems become matrix eigenvalue problems and the partial differential boundary value problems for finding the coefficients of the center manifold function become systems of linear equations. In all cases, the discretization leads to large sparse systems, and thus we seek appropriate solution techniques.

In this section, we discuss some of the details of the numerical approximations, including the discretization and solution techniques. Also included is a brief discussion of convergence for the approximation.

5.1. The mesh: Nonuniform spacing. We employ centered finite differencing to discretize the spatial derivatives. The values of the unknown functions in the interior of the domain are approximated at $N \times N$ grid points, labeled by $(r, z) = (r_k, z_l)$, with $1 \le k \le N$ and $1 \le l \le N$, where N, k, l are positive integers. We define $r_0 = r_a/R$, $r_{N+1} = r_b/R$, $z_0 = 0$, and $z_{N+1} = 1$. This leads to a discretized solution vector of size $3N^2 + 2N$ (because $T(r_k, z_0)$ and $T(r_k, z_{N+1})$ are also unknown for $1 \le k \le N$).

With the no-slip boundary conditions and the small parameter ν multiplying a second derivative term, boundary layers form in the fluid flow. For this reason, a scaling method is used to choose the locations of the grid points. This consists of making a change of coordinates and calculating the solutions on a uniform grid in the new coordinates. The transformation is chosen such that its inverse takes a uniform grid to a grid with many points near the boundary. The transformation that takes the new coordinates (x, y) to the original coordinates (r, z) is given by

$$r = \frac{\tan^{-1}(\eta x)}{2\tan^{-1}\left(\frac{\eta}{2}\right)} + \frac{1}{2} + \frac{r_a}{R}, \qquad z = \frac{\tan^{-1}(\eta y)}{2\tan^{-1}\left(\frac{\eta}{2}\right)} + \frac{1}{2}, \tag{30}$$

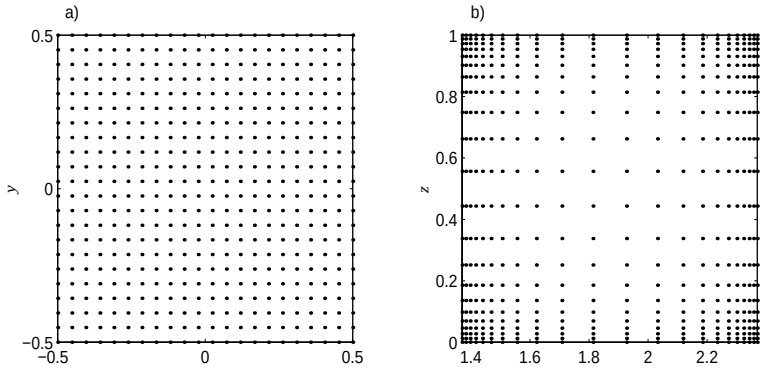


FIG. 3. *The transformation of the grid points.* (a) *A uniform grid (equally spaced grid points) of* $N = 20$, (b) *the grid obtained by applying the change of coordinates* (30) *with* $\eta = 6$.

where η is a scaling factor that determines the magnitude of compression near the boundary; see Figure 3. The domain $r_a \leq r \leq r_b/R$, $0 \leq z \leq 1$ is mapped to $-1/2 \leq x \leq 1/2$, $-1/2 \leq y \leq 1/2$, and the solutions are approximated on a uniform grid in the (x, y) coordinates.

The boundary layers observed in the eigenfunctions are not as severe as those in the axisymmetric solutions. In fact, significant errors are introduced in the eigenvalues and eigenfunctions if the points in the interior are too sparse. This occurs even if the axisymmetric solutions appear to be well resolved. This suggests that different scaling factors should be used for the axisymmetric and eigenvalue problems. However, errors introduced in interpolation seem to negate the potential benefit of using different scaling factors. In the calculations presented, the scaling factor $\eta = 6$ is used. This is the smallest value that leads to qualitatively good results for the axisymmetric problem when $N = 20$; for smaller values of η, the boundary layer is not resolved well enough. Also, for larger values of η (for $N = 20$), there is an insufficient number of interior points to adequately describe the eigenfunctions. However, for larger values of N, the results are consistent and not as sensitive to the choice of η.

5.2. Solution techniques. For the computation of the axisymmetric solution we use Newton's method. This method can be combined with a predictor-corrector continuation technique to find the axisymmetric solution for a wide range of parameter values. If $\Omega = 0$ and $\Delta T = 0$, then the trivial solution satisfies the axisymmetric equations. Thus for Ω and ΔT small, the trivial solution is a reasonable prediction of the solution, and Newton's method is used for the correction. For small increments in the parameter values, the previous solution is a reasonable prediction. To make larger increments in the parameter values, a secant line approximation can be used for the prediction.

Each point on a neutral stability curve is found using an iterative secant method, where the real part of the eigenvalue with largest real part is considered as a function of the parameters. The iterative procedure for the localization of the double Hopf points uses the fact that the points occur at intersections of two neutral stability curves. In both procedures, iteration continues until the magnitudes of the real parts of the relevant eigenvalues are less than a specified tolerance (10^{-8} for the results presented below).

The discretized transformed equations and the entries of the coefficient matrices

are computed symbolically using Maple. The generalized matrix eigenvalue problem, which results from the discretization of (14), is solved in Matlab using the implicitly restarted Arnoldi method [17], which is a memory-efficient iterative method for finding a specified number of eigenvalues with the largest magnitudes. A generalized Cayley transformation is made so that the Arnoldi iteration finds the eigenvalues with largest real parts [8]. The parameters of the transformation can also be chosen to improve convergence properties. In particular, the generalized Cayley transformation

$$\mathbf{C}(\mathbf{L},\mathbf{A}) = (\mathbf{L} - \alpha_1 \mathbf{A})^{-1} (\mathbf{L} - \alpha_2 \mathbf{A}) \tag{31}$$

maps eigenvalues λ of the generalized matrix eigenvalue problem $\lambda \mathbf{A} v = \mathbf{L} v$ to eigenvalues σ of the transformed matrix $\mathbf{C}(\mathbf{L},\mathbf{A})$, such that the eigenvalues λ with $\mathrm{Real}(\lambda) > (\alpha_1 + \alpha_2)/2$ are mapped to the eigenvalues σ with $|\sigma| > 1$, where α_1 and α_2 are the real parameters of the Cayley transformation. The matrix $\mathbf{C}(\mathbf{L},\mathbf{A})$ does not have to be formed explicitly, because the Arnoldi iteration only requires matrix-vector products involving $\mathbf{C}(\mathbf{L},\mathbf{A})$; see [17]. Thus, the sparseness properties of $\mathbf{L}$ and $\mathbf{A}$ can be exploited, and computer memory requirements can be reduced.

5.3. Convergence. For the centered differencing that was used, the local truncation error is $O(h^2)$ (i.e., approximately a constant times h^2, as $h \to 0$), where h is the mesh size. Given this and a few standard assumptions, the accuracy of the approximations for the boundary value problems will be $O(h^2)$. In addition, if the approximate solution and the differencing scheme for the derivative are both $O(h^2)$, then the approximations of derivatives of the solutions are also $O(h^2)$. However, for the present application, although the approximation of the partial differential eigenvalue problem by the matrix eigenvalue problem can be assumed to converge, the order of this convergence is unknown. Considering this, it is reasonable to assume that the approximations of the normal form coefficients converge, even though we could not say to what order.

An additional comment should be made concerning the eigenfunction approximation. It is obvious that a finite-dimensional approximation is not able to approximate all the solutions of the infinite-dimensional continuous eigenvalue problem. We expect that it is the highly oscillatory high wave number eigenfunctions that the matrix problem is unable to resolve. Because the critical eigenfunctions of interest have relatively low wave numbers and are not highly oscillatory, we expect that these functions are resolved and that the errors in the differencing are relatively small.

Our results, which are presented in the next section, indicate that the approximation of the normal form coefficients seems to be convergent. However, the mesh size h could not be taken small enough to obtain an accurate estimate of the order of convergence.

6. Results. The results of our study are presented in this section. The parameter values specifying the geometry of the annulus and fluid properties are listed in Table 1. These values correspond to the experiments performed by Fein [4]. Our results are compared with those obtained in that study.

6.1. The axisymmetric solution. An example of the axisymmetric solution is plotted in Figure 4. Qualitatively, the form of the solution is the same for all values of the parameters. The figure shows that the fluid velocity in the interior of the fluid is predominantly in the azimuthal direction. The radial velocity is almost zero everywhere except near the upper and lower boundaries, where it is negative and positive, respectively. The vertical velocity is largest at the inner and outer walls,

TABLE 1

The annulus geometry and fluid properties used in the analysis, after [4]. *See section* 3 *for definitions of symbols.*

r_a	3.48	cm
r_b	6.02	cm
R	2.54	cm
D	5	cm
ν	$1.01e^{-2}$	cm^2/sec
κ	$1.41e^{-3}$	cm^2/sec
α	$2.06e^{-4}$	$1/^\circ$ C
ρ_0	0.998	gm cm^3
T_0	20.0	$^\circ$ C
g	980	gm/cm^3

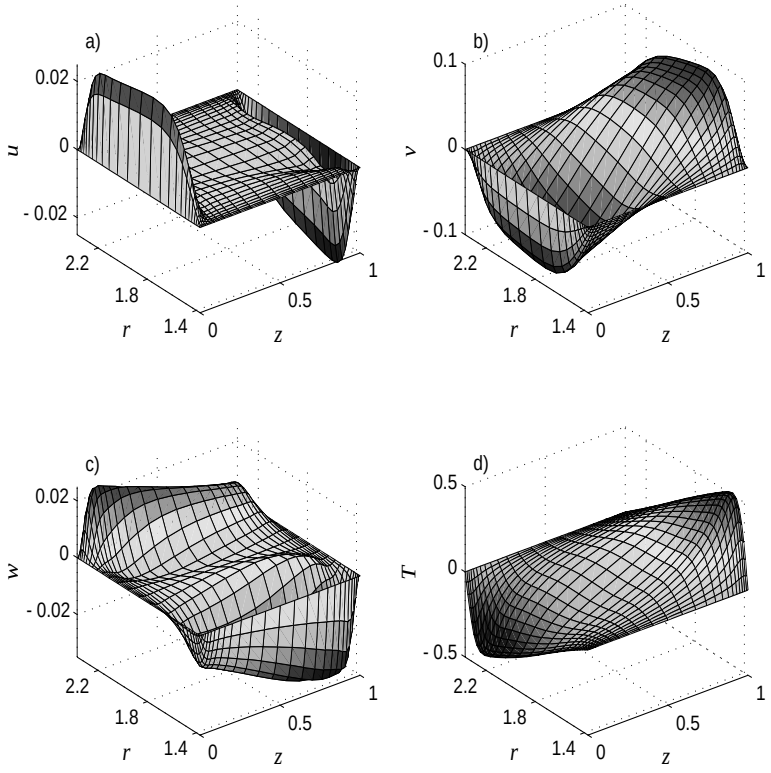


FIG. 4. *The axisymmetric solution:* (a) u, *the fluid velocity in the radial direction,* (b) v, *the fluid velocity in the azimuthal direction,* (c) w, *the fluid velocity in the vertical direction, and* (d) T, *the deviation of the temperature of the fluid from* $\Delta T\,(r - r_a/R) + T_a$. *This solution is for the* $N = 25$ *case and is observed at the* $(m_1, m_2) = (6, 7)$ *double Hopf point, where* $\Omega = 0.5927$ *and* $\Delta T = 0.6950$.

where there is rising at the warmer outer wall and sinking at the cooler inner wall. The interior azimuthal velocity exhibits an almost linear shear in the vertical, with a positive velocity in the upper half of the annulus and negative velocity in the lower half. The resulting circulation is a convection cell that is tilted from the radial plane such

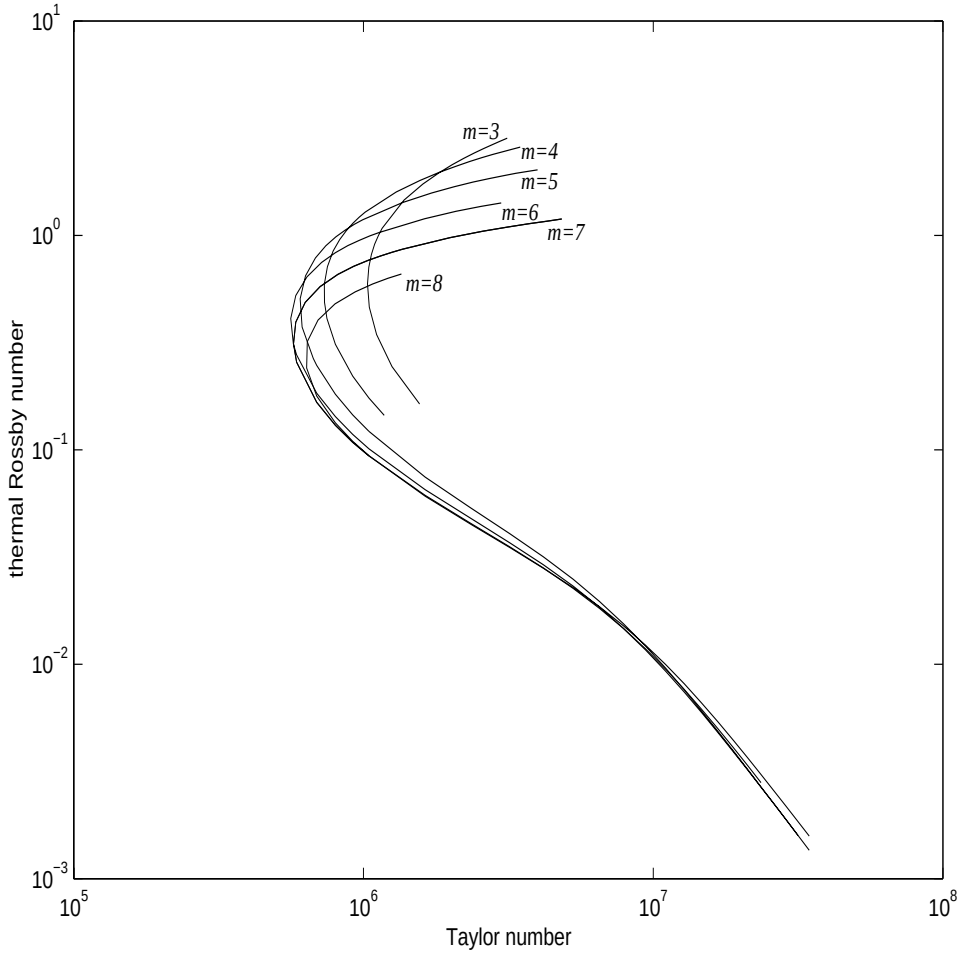


FIG. 5. *Neutral stability curves are plotted for the wave numbers $m = 3$ to $m = 8$. The curves are calculated by finding the parameter values where for each m the eigenvalues of* (14) *all have negative real part except one with zero real part. The curves are plotted on a log-log graph of thermal Rossby number versus Taylor number.*

that, at the upper and lower boundaries, the inward and outward motion is deflected to the right. Although quantitative information of the experimental axisymmetric flow was not available, the computed flow profile qualitatively reproduces all the features of the experimental flow.

6.2. Neutral stability and transition curves. The neutral stability curves are presented in Figure 5. There is a separate curve for each azimuthal wave number. The curves consist of points in the parameter space where, for the given wave number, there is one pair of complex conjugate eigenvalues with zero real part while all other eigenvalues associated to that wave number have negative real part. Wave numbers from $m = 2$ to 10 were calculated, and it was found that $m = 3$ to 8 were the only critical wave numbers. Therefore, only these values are shown in Figure 5. It is not possible to calculate the neutral stability curves of all wave numbers, but it can be argued that the higher wave numbers will not be critical in the parameter range of interest. We refer the interested reader to [18] and here justify investigation of only a

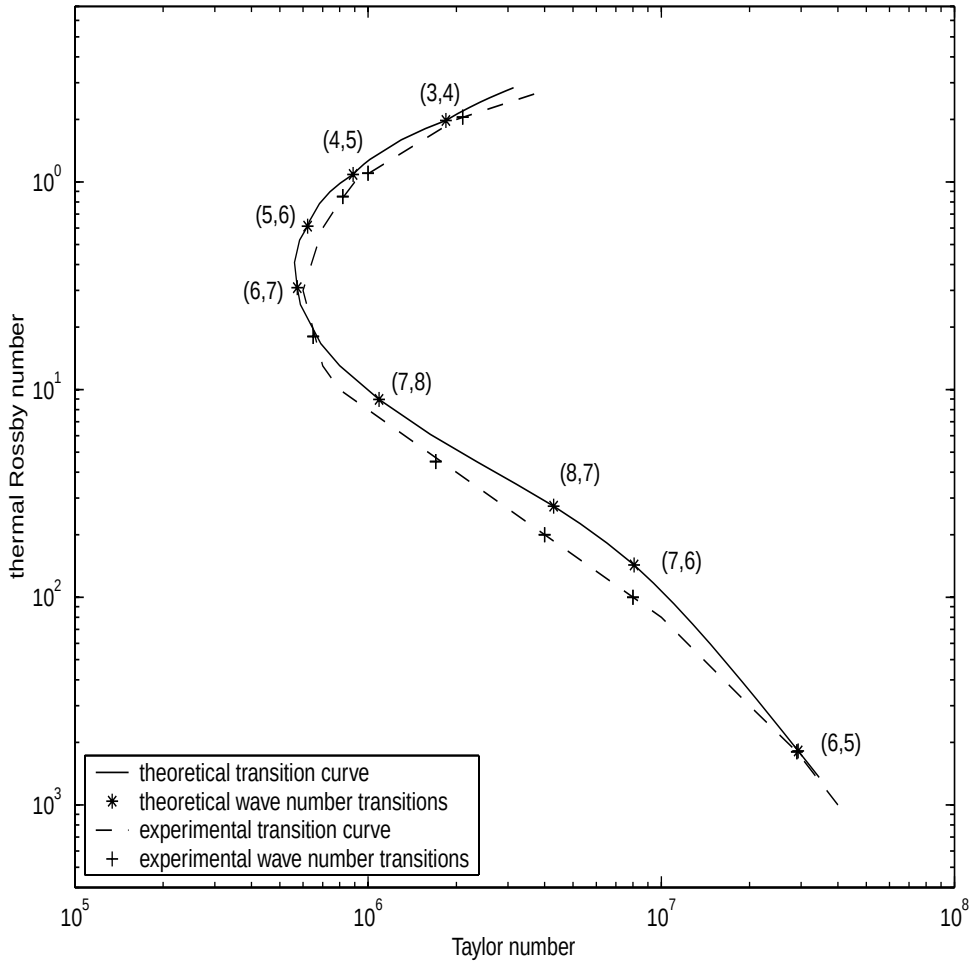


FIG. 6. *Transition curves for theory and experiment delineating the axisymmetric from the nonaxisymmetric regimes. The critical wave number transitions (double Hopf bifurcation points), labeled as* (m_1, m_2)*, are also plotted along the curve.*

finite number of wave numbers by comparison with the experimental results. A 25×25 grid was used for the calculations of all curves shown.

In Figure 6, the curve that separates the axisymmetric regime from the nonaxisymmetric regime is plotted. To the left of this curve, the axisymmetric solution is linearly stable (to perturbations of all wave numbers), while to the right, it is unstable. Along this curve it can be seen that there are transitions of the critical wave number. These transitions occur at intersections of the neutral stability curves and correspond to the double Hopf bifurcation points; i.e., at these points there are two complex conjugate pairs of eigenvalues with zero real parts. Also plotted in Figure 6 is the experimentally observed transition curve taken from Fein [4], with critical wave number transitions. This is the curve along which a transition from the axisymmetric to steady wave flow was observed. All curves are plotted on a log-log graph of the Taylor number $\mathcal{T}$ versus the thermal Rossby number $\mathcal{R}$ (see section 2).

Linear analysis reproduces many of the experimental observations. By inspection

TABLE 2

Numerical results for double Hopf bifurcation points; $a = -1$ and $d = -1$ for all N. Here, m_1 and m_2 are the critical wave numbers, N is the number of grid points on one side, $(\Omega_0, \Delta T_0)$ is the location in parameter space where the bifurcation occurs, ω_1 and ω_2 are the imaginary parts of the eigenvalues at $(\Omega_0, \Delta T_0)$, and a,b,c, and d are the normal form coefficients.

N	m_1, m_2	Ω_0	ΔT_0	$-\omega_1$	$-\omega_2$	b	c
30	3, 4	1.025	12.76	$2.320 \cdot 10^{-2}$	$3.305 \cdot 10^{-2}$	-1.0051	-2.362
40	3, 4	1.030	12.93	$2.514 \cdot 10^{-2}$	$3.632 \cdot 10^{-2}$	-0.9720	-2.651
50	3, 4	1.034	13.04	$2.560 \cdot 10^{-2}$	$3.712 \cdot 10^{-2}$	-0.9813	-2.723
30	4, 5	0.7313	3.772	$1.450 \cdot 10^{-2}$	$1.936 \cdot 10^{-2}$	-1.332	-2.273
40	4, 5	0.7271	3.795	$1.508 \cdot 10^{-2}$	$2.018 \cdot 10^{-2}$	-1.327	-2.367
50	4, 5	0.7276	3.840	$1.533 \cdot 10^{-2}$	$2.053 \cdot 10^{-2}$	-1.324	-2.414
20	5, 6	0.6354	1.543	$7.946 \cdot 10^{-3}$	$1.039 \cdot 10^{-2}$	-1.360	-2.134
30	5, 6	0.6102	1.490	$8.462 \cdot 10^{-3}$	$1.091 \cdot 10^{-2}$	-1.470	-2.187
40	5, 6	0.6048	1.502	$8.721 \cdot 10^{-3}$	$1.124 \cdot 10^{-2}$	-1.483	-2.237
50	5, 6	0.6036	1.517	$8.867 \cdot 10^{-3}$	$1.141 \cdot 10^{-2}$	-1.488	-2.265
20	6, 7	0.6117	0.6972	$3.711 \cdot 10^{-3}$	$4.960 \cdot 10^{-3}$	-1.473	-2.253
30	6, 7	0.5838	0.6944	$4.046 \cdot 10^{-3}$	$5.398 \cdot 10^{-3}$	-1.532	-2.256
40	6, 7	0.5757	0.7008	$4.217 \cdot 10^{-3}$	$5.611 \cdot 10^{-3}$	-1.556	-2.269
50	6, 7	0.5727	0.7071	$4.317 \cdot 10^{-3}$	$5.735 \cdot 10^{-3}$	-1.568	-2.277
20	7, 8	0.8699	0.3959	$8.582 \cdot 10^{-4}$	$1.161 \cdot 10^{-3}$	-1.628	-2.433
30	7, 8	0.7925	0.3758	$9.294 \cdot 10^{-4}$	$1.283 \cdot 10^{-3}$	-1.616	-2.408
40	7, 8	0.7652	0.3713	$9.750 \cdot 10^{-4}$	$1.356 \cdot 10^{-3}$	-1.625	-2.404
50	7, 8	0.7505	0.3704	$10.09 \cdot 10^{-4}$	$1.410 \cdot 10^{-3}$	-1.631	-2.399
20	8, 7	1.603	0.4692	$4.010 \cdot 10^{-4}$	$3.493 \cdot 10^{-4}$	-2.309	-1.748
30	8, 7	1.635	0.4581	$3.748 \cdot 10^{-4}$	$3.284 \cdot 10^{-4}$	-2.274	-1.723
40	8, 7	1.655	0.4559	$3.602 \cdot 10^{-4}$	$3.156 \cdot 10^{-4}$	-2.270	-1.722
50	8, 7	1.670	0.4556	$3.501 \cdot 10^{-4}$	$3.064 \cdot 10^{-4}$	-2.268	-1.722
20	7, 6	2.226	0.4625	$1.553 \cdot 10^{-4}$	$1.361 \cdot 10^{-4}$	-2.311	-1.734
30	7, 6	2.231	0.4457	$1.441 \cdot 10^{-4}$	$1.229 \cdot 10^{-4}$	-2.309	-1.719
40	7, 6	2.250	0.4391	$1.323 \cdot 10^{-4}$	$1.109 \cdot 10^{-4}$	-2.310	-1.718
50	7, 6	2.269	0.4356	$1.232 \cdot 10^{-4}$	$1.018 \cdot 10^{-4}$	-2.310	-1.717
20	6, 5	3.843	0.2559	$2.064 \cdot 10^{-5}$	$2.083 \cdot 10^{-5}$	-2.376	-1.746
30	6, 5	4.696	0.1718	$2.278 \cdot 10^{-5}$	$2.198 \cdot 10^{-5}$	-2.350	-1.733
40	6, 5	5.886	0.1148	$2.257 \cdot 10^{-5}$	$2.165 \cdot 10^{-5}$	-2.336	-1.729
50	6, 5	7.449	0.0764	$2.315 \cdot 10^{-5}$	$2.220 \cdot 10^{-5}$	-2.330	-1.730

of Figure 6, it can be seen that there is a good correspondence between the numerical and experimental transition curves. It has also been shown, via comparison with experimentally measured wave speeds at the transition, that the imaginary parts of the eigenvalues are also in agreement. See [19] for further discussion.

6.3. Double Hopf normal form coefficients: Hysteresis. The numerical results are presented in Table 2. Included are the locations of the double Hopf bifurcation points and the values of the normal form coefficients, as well as the values of

the imaginary parts of the critical eigenvalues at the bifurcation point. The double Hopf points are labeled in terms of the associated critical wave numbers m_1 and m_2. For all double Hopf points, the critical wave numbers m_1 and m_2 differ by the integer one, and the normal form coefficients satisfy $a = -1$, $b < 0$, $c < 0$, and $d = -1$, as well as the condition $A = ad - bc < 0$.

The dynamics are found from an investigation of the fixed points of the equations obtained by ignoring the θ_j in the normal form equations (28). To lowest order, the $\dot{\theta}_j$ equations add a constant rotation for each corresponding dimension. See [9] for a complete analysis of the normal form equations (28). Here, there are fixed points when

(i) $\rho_1 = \rho_2 = 0$,
(ii) $\rho_2 = 0$ and $\rho_1 = \rho_p = \sqrt{\mu_1/-a}$,
(iii) $\rho_1 = 0$ and $\rho_2 = \rho_q = \sqrt{\mu_2/-d}$,
(iv) $\rho_1 = \rho_1^{(T)} = \sqrt{(-d\mu_1 + b\mu_2)/A}$ and $\rho_2 = \rho_2^{(T)} = \sqrt{(c\mu_1 - a\mu_2)/A}$,

where $A = ad - bc$, and with the condition that the quantities inside the square root signs must be positive.

Fixed point (i) is a fixed point of the normal form equations for all values of the parameters. By inspection of the normal form equations (28), it is fairly easy to see that, regardless of the values of the coefficients, for small ρ_1 and ρ_2, $\dot{\rho}_1$ and $\dot{\rho}_2$ will have the same sign as μ_1 and μ_2, respectively. This means that solution (i) will be stable if both μ_1 and μ_2 are negative, and unstable if either one is greater than zero. In the fluid annulus, this solution corresponds to the steady axisymmetric flow. Fixed points (ii) and (iii) correspond to periodic solutions of the normal form equations and exist when $\mu_1 > 0$ and $\mu_2 > 0$, respectively (because we have $a = -1$ and $d = -1$). In the fluid, these solutions correspond to nonaxisymmetric steadily rotating waves. The fixed point (iv) corresponds to a 2-torus for the normal form equations and exists when $(-d\mu_1 + b\mu_2)/A > 0$ *and* $(c\mu_1 - a\mu_2)/A > 0$. Because we have $A < 0$, $a = -1$, $b < 0$, $c < 0$, and $d = -1$, the 2-torus exists in the wedge in (μ_1, μ_2) parameter space given by $d\mu_1/b < \mu_2 < c\mu_1/a$, $\mu_1 > 0$, $\mu_2 > 0$. These solutions correspond to modulated wavy flow in the fluid.

A linear stability analysis of fixed points (ii), (iii), and (iv) gives the local behavior near the bifurcation points; see Figure 7 for the bifurcation diagram. The results indicate that both of the bifurcating waves, corresponding to the fixed points (ii) and (iii), are stable in the wedge where the 2-torus exists. Thus, the boundaries of the wedge $\mu_2 = d\mu_1/b$ and $\mu_2 = c\mu_1/a$ give the boundaries of the region of bistability of the wave solutions. Furthermore, because only one of the bifurcating waves loses stability on each of the boundaries of the wedge, there is hysteresis. The results also indicate that the 2-torus is always unstable. In Figure 9, the approximate boundaries of the region of bistability are drawn. The bifurcation diagram in Figure 8 shows the hysteresis that occurs when a one-parameter path through the parameter space crosses the region of bistability. The parameter s could be either Ω or ΔT or a function of both, depending on the particular circumstances. An example of such a path is indicated on Figure 7.

Quantitative verification of the predicted hysteresis is not currently possible due to the lack of experimental data for the specific annulus studied here. Furthermore, although the extent of hysteresis has been mapped for other transitions in other regions of parameter space (see, e.g., [5], [14], [16], and [4]), there is relatively little data concerning the hysteresis that occurs in the transitions between steady waves near the axisymmetric regime. Also, many experimental results in the steady wave

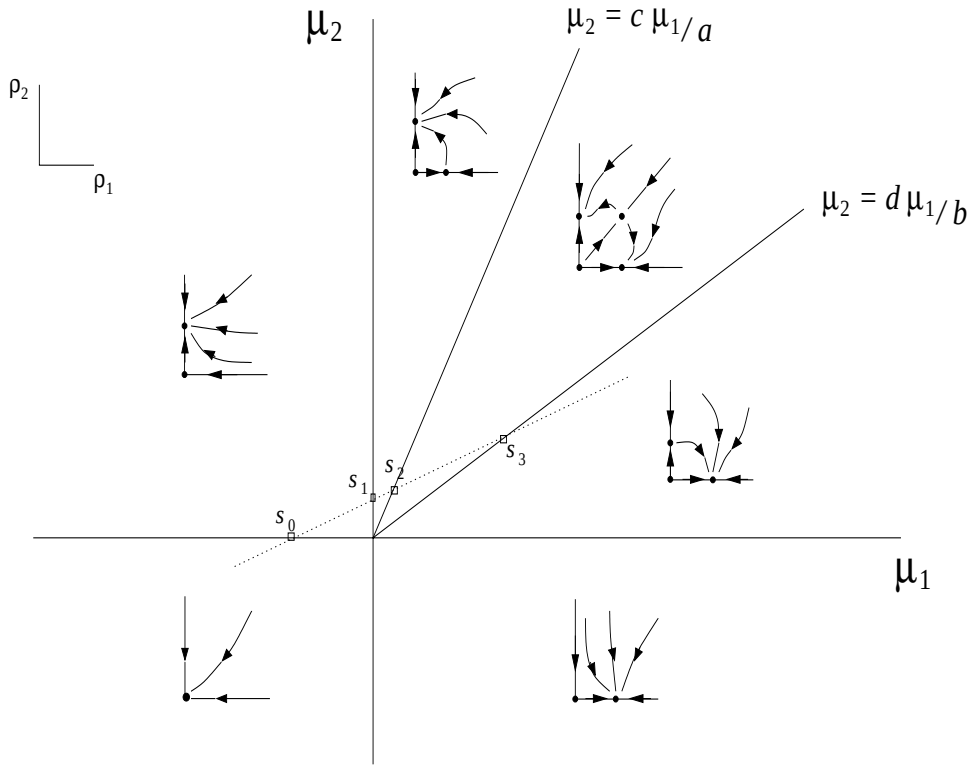


FIG. 7. *The two-dimensional bifurcation diagram. The diagram is displayed using the real parts of the critical eigenvalues μ_1, μ_2 as the bifurcation parameters. The regions of different character are separated by solid lines. In each region, the corresponding phase portrait is drawn, where the phase portraits are presented in ρ_1, ρ_2 coordinates. The θ_1 and θ_2 equations in* (28) *add a constant rotation to each coordinate. The dotted line indicates a possible one-parameter path which will lead to hysteresis. The bifurcation points along this path are indicated by s_0, s_1, s_2, and s_3 (see Figure* 8*).*

regime are quoted in terms of the wave number that is most likely to occur. Our analysis cannot predict this.

It can be seen in Table 2 that the numerical differences between the normal form coefficients at the different levels of discretization decrease with increasing discretization level. This is an indication of the convergence of the numerical approximations. However, it appears that N is not large enough for us to make an estimation of the order of convergence. Yet, the differences in the normal form coefficients at different N are quite small, which is evidence that these results are at least qualitatively accurate. To say this with more certainty, the analysis must be performed using higher levels of discretization. This was not possible with the available resources. Because the results accurately reproduce the experimental results, we conclude that the approximations are satisfactory.

The results for the $(m_1, m_2) = (3, 4)$ and $(m_1, m_2) = (4, 5)$ double Hopf points are not complete (see Table 2). For these wave number pairs with $N = 20$, the eigenfunctions are not well resolved, and the eigenvalues are inaccurate. Also, the evidence of convergence of the normal form coefficients (see Table 2) is weaker for the double Hopf points at higher differential heating. It seems that the increase in numerical difficulty is not caused by the difficulty of resolving the boundary layer in

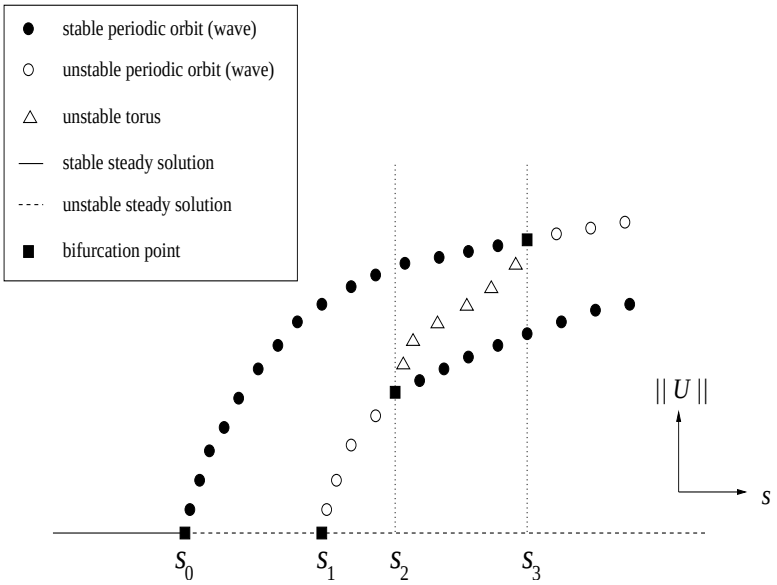


FIG. 8. *The one-dimensional bifurcation diagram depicting the bifurcation observed along the path indicated with the dotted line in Figure* 7. *The bifurcation points are labeled as* s_0, s_1, s_2, *and* s_3. $\|U\|$ *is a measure of the size of the solution, and* s *is the bifurcation parameter.*

the axisymmetric solution, but rather by the difficulty of resolving the eigenfunctions in the interior of the domain.

6.4. The eigenfunctions: Bifurcating wave form. An example of an eigenfunction is plotted in Figure 10. This is the eigenfunction with wave number $m = 6$ that is observed at the $(m_1, m_2) = (6, 7)$ double Hopf point (see Table 2). From (28), the periodic orbit corresponding to the wave with wave number m_1, to lowest order in μ_1, is given by

$$\begin{aligned} \rho_1 &= \sqrt{\frac{-\mu_1}{a}} + O(\mu_1), \\ \theta_1 &= \omega_1 t + O(\mu_1) \end{aligned} \tag{32}$$

or, in terms of z_1,

$$z_1 = \sqrt{\frac{-\mu_1}{G_{11}^r}} e^{i\omega_1 t} + O(\mu_1), \tag{33}$$

which describes a near-circular periodic orbit. The periodic orbit corresponding to the wave with wave number m_2 is given by a similar expression. In terms of the variables of the perturbation equations (10)–(12), to lowest order, the periodic solution corresponding to (32) is given by $U = [\mathbf{u}, T] = z_1\Phi_1 + \overline{z}_1\overline{\Phi}_1 = \mathrm{Re}\,(z_1\Phi_1)$. That is,

$$\begin{aligned} U &= \mathrm{Re}\left[\sqrt{\frac{-\mu_1}{G_{11}^r}} e^{i\omega_1 t}\tilde{\Phi}_{m_1} e^{im_1\varphi}\right] + O(\mu_1) \\ &= \sqrt{\frac{-\mu_1}{G_{11}^r}}\left[\tilde{\Phi}_{m_1}^r \cos(m_1\varphi + \omega_1 t) - \tilde{\Phi}_{m_1}^i \sin(m_1\varphi + \omega_1 t)\right] + O(\mu_1), \end{aligned} \tag{34}$$

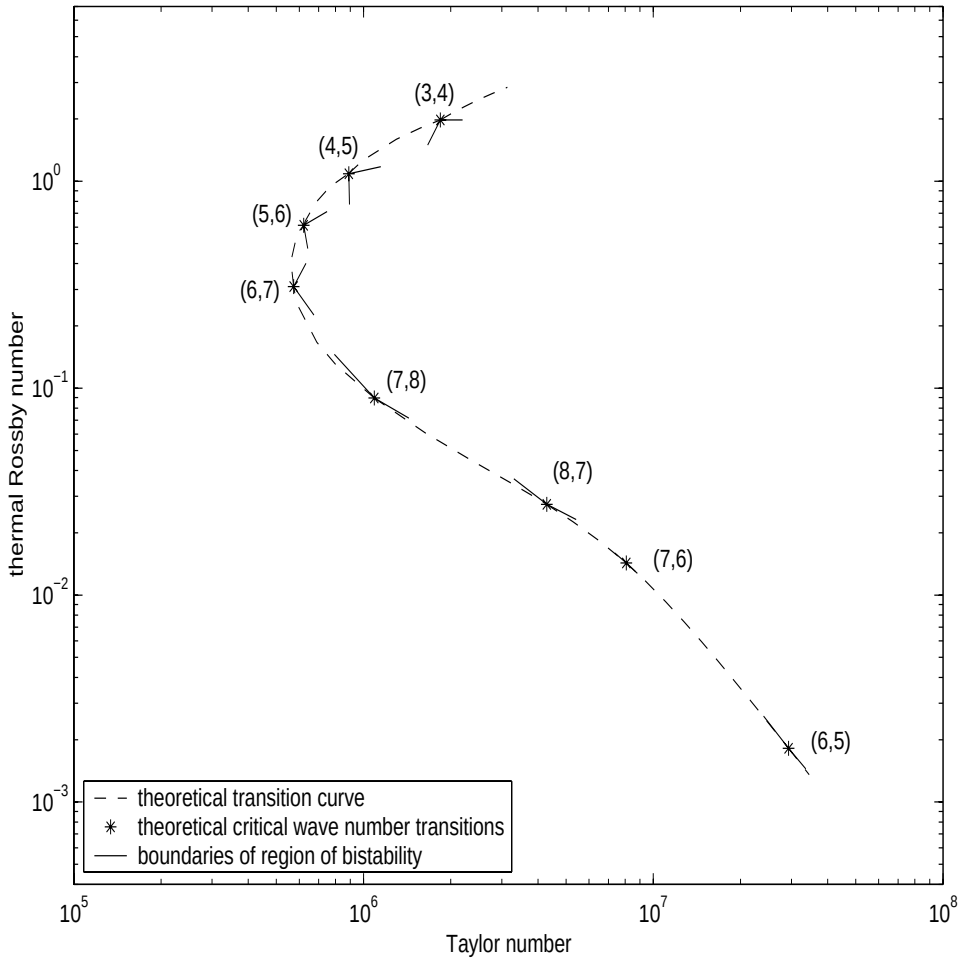


FIG. 9. *Theoretical transition curve between the axisymmetric and the nonaxisymmetric regimes including the boundaries of the region of bistability. The boundaries are the solid lines attached to the double Hopf points. For each double Hopf point, the area between the boundaries is the region where there is bistability of wave solutions.*

i.e., a rotating wave, where $\Phi_1 = \tilde{\Phi}_{m_1} e^{im_1\varphi}$ and where $\tilde{\Phi}^r_{m_1}$ and $\tilde{\Phi}^i_{m_1}$ denote the real and imaginary parts of $\tilde{\Phi}_{m_1}$, respectively. In terms of the variables of the original equations (5)–(7), the solution U corresponds to deviations from the axisymmetric solution $[\mathbf{u}^{(0)}, T^{(0)}]$ of the same equations. Also, if t is fixed, then at different φ, the periodic solution is a different linear combination of $\tilde{\Phi}^r_{m_1}$ and $\tilde{\Phi}^i_{m_1}$, and thus, the form of the eigenfunction gives the form of the bifurcating wave to the lowest order of approximation. That is, the approximation is valid for parameter values that are close to the axisymmetric-to-wave transition curve.

The form of the bifurcating wave is consistent with previous results. Measurements, from experiments with the same annulus geometry as is used for our results, indicate that the temperature has a maximum at midradius middepth [4]. Furthermore, the coarse features of the wave form are consistent with the detailed experimental and numerical results of [15], as well as the numerical results of [27], even

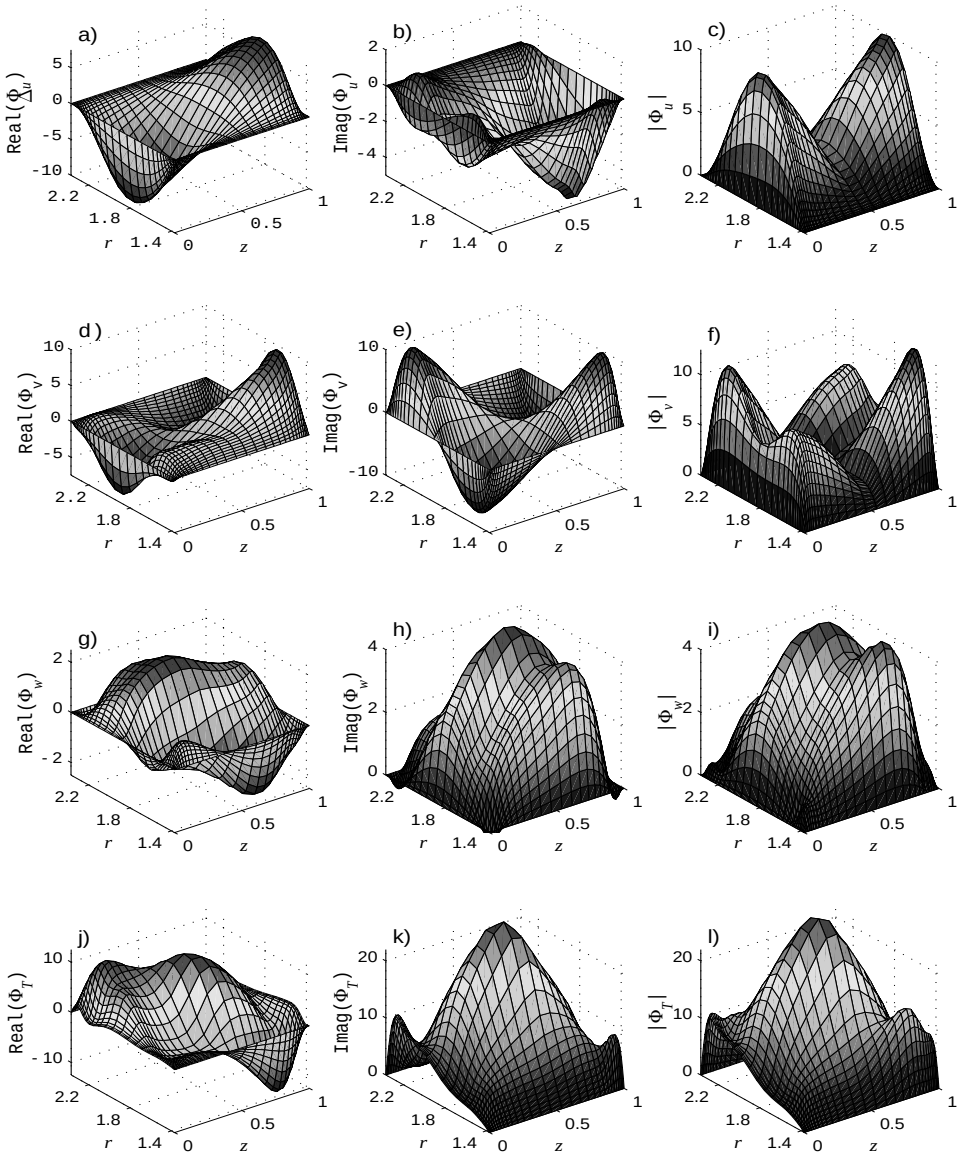


FIG. 10. *An example of the radial and vertical dependence of an eigenfunction with* $m = 6$ *and* $N = 30$ *at* $\Omega = 0.5838$ *and* $\Delta T = 0.6944$: (a) *real part,* (b) *imaginary part, and* (c) *amplitude of the radial component of the eigenfunction;* (d) *real part,* (e) *imaginary part, and* (f) *amplitude of the azimuthal component of the eigenfunction;* (g) *real part,* (h) *imaginary part, and* (i) *amplitude of the vertical component of the eigenfunction;* (j) *real part,* (k) *imaginary part, and* (l) *amplitude of the temperature component of the eigenfunction. That is, the actual components of the eigenfunctions are the plotted functions multiplied by* $e^{im\varphi}$.

though different annulus geometries, waves with different dominant wave numbers, and waves far from the axisymmetric-to-wave transition curve are studied. This includes (see Figures 5 and 6 of [15]) the radial dependence of the Fourier amplitude of the dominant wave number of the radial velocity at various heights, and the radial dependence of the Fourier amplitude of the dominant wave number of the azimuthal velocity at middepth. In our case, the square of the Fourier amplitude is given by

$(\tilde{\Phi}^r_{m_j})^2 + (\tilde{\Phi}^i_{m_j})^2$. See also Figure 8 in [27] for the vertical dependence of the deviations from the azimuthally averaged flow of both temperature and velocity for the (numerical) wave forms in an annulus without a rigid lid. However, compared to our bifurcating waves (34), the waves of these experimental and numerical studies show a relative decrease in the amplitude at midradius of the azimuthal average of the azimuthal velocity (i.e., the wave number zero Fourier component of the azimuthal velocity). In our case, the azimuthal average to first order is given by the axisymmetric solution $[\mathbf{u}^{(0)}, p^{(0)}, T^{(0)}]$ (see Figure 4). The waves studied in [15] and [27] are observed in regions of parameter space far from the axisymmetric-to-wave transitions, where the higher-order terms in (34), which we have ignored, may be important. Although these higher-order terms do not produce a significant qualitative change in the deviations from the azimuthally averaged flow, they do seem to produce a small, but noticeable, qualitative difference on the the azimuthal averaged flow itself. In order to study this effect, the bifurcating waves (34) would have to be calculated for parameter values far from the transition curve.

7. Conclusion. In this paper, we study the transitions from axisymmetric steady solutions to nonaxisymmetric waves in a Navier–Stokes model of the differentially heated rotating annulus experiment. An analytical-numerical center manifold reduction is used to analyze the double Hopf bifurcation points that occur at this transition. The results, which are obtained by numerically approximating the coefficients of the normal form equations, show that there are stable waves that bifurcate from the axisymmetric solution via a Hopf bifurcation, and that hysteresis (and bistability) of the bifurcating waves occurs near critical wave number transitions. Associated with the hysteresis is the existence of an unstable torus. Approximate boundaries to the region of bistability are drawn. The results are consistent with laboratory experiments, which supports not only the validity of the model, but also the validity of the analysis. Although the convergence of the numerical approximations cannot be proven, the evidence of convergence and the correspondence with experimental results supports the claim that the behavior that is predicted by our results occurs in the full PDE model.

The behavior seen in the model of the experiment is qualitatively the same as that seen in the models of the analytical studies discussed in the first section. This is very interesting because the models of these analytical studies are simplified models of atmospheric circulation, and so they are of a very different scale from that of the experiment. That is, the method of analysis is able to highlight the dynamical similarity of two geophysical fluid models of vastly different scales. The similarity is evidence for the usefulness of studying the models of both scales and for the statement that both types of models incorporate the fundamental properties of differentially heated rotating systems.

The study presented here is a beginning, and there are many possible directions future work could take. The models of atmospheric circulation of the analytical studies mentioned above are quite simplified. The success of the numerical computations of the present study gives confidence that analysis of this type could be applied to more realistic atmospheric models, such as that presented in [3]. Also, in the analysis of the annulus experiment, there is the possibility of resonant behavior close to an experimentally observed "triple-point," which is a point in parameter space that is shared by three regimes (the axisymmetric, the wave, and the irregular regimes; see Figure 2). The $(m_1, m_2) = (6, 5)$ double Hopf point, which occurs in a similar region in parameter space as does the experimentally observed triple-point, is close to being

resonant; i.e., the imaginary parts of the two complex conjugate pairs of eigenvalues with largest real part are nearly equal. Thus, a strongly resonant double Hopf bifurcation might be found by varying a third parameter, and in this case the dynamics found close to the resonant bifurcation may explain the existence of the triple-point.

Another interesting direction would be to attempt to follow the bifurcating solutions as the parameters move away from the bifurcation point. Two interesting flows that are observed in the annulus (both experimentally and numerically) are amplitude vacillation and wave dispersion [14], [22]. It has been hypothesized that the mechanism responsible for both of these flows is an interaction of two waves via a stable torus, where amplitude vacillation results from an interaction of two waves of the same dominant azimuthal wave number, while wave dispersion results from an interaction of waves with different dominant azimuthal wave numbers [23], [21]; see also [5] for experimental evidence.

The unstable torus, which we have shown to exist in the steady wave regime, is such an interaction of two waves with different wave numbers. Thus, it is possible that if the unstable torus could be followed further into the steady wave regime, a bifurcation to a stable torus (and wave dispersion) might be discovered. Alternatively, if the stable periodic orbits corresponding to the wave solutions could be followed further into the wave regime, a bifurcation to a stable torus might occur, which might result in the discovery of either amplitude vacillation or wave dispersion. At the moment, such a study seems computationally prohibitive. However, if the curvature of the annulus is neglected, a symmetry of the resulting system leads to a bifurcation to a steady solution as opposed to a periodic orbit. In this case, the computation may be possible.

The comparison of theoretical and experimental results that took place in the investigation of the Taylor–Couette flow led to many more discoveries about the system than otherwise would have occurred. The work presented here begins such a comparison for the differentially heated rotating annulus flow. Some of our results are confirmed by comparison with experiments, and some predictions, concerning the boundaries of the region of bistability, have yet to be verified. For future work, we expect that the use of such techniques will lead to further discovery of new dynamics, both theoretical and experimental, which in turn will lead to a better general understanding of differentially heated rotating fluid systems.

REFERENCES

[1] P. Chossat and G. Iooss, *The Couette-Taylor Problem*, Appl. Math. Sci. 102, Springer-Verlag, New York, 1994.

[2] P. Drazin, *Nonlinear baroclinic instability of a continuous zonal flow of viscous fluid*, J. Fluid Mech., 55 (1972), pp. 577–587.

[3] J. Dutton and P. Kloeden, *The existence of Hadley convective regimes of atmospheric motion*, J. Austral. Math. Soc. Ser. B, 24 (1983), pp. 318–338.

[4] J. Fein, *An experimental study of the effects of the upper boundary condition on the thermal convection in a rotating cylindrical annulus of water*, Geophys. Fluid Dynamics, 5 (1973), pp. 213–248.

[5] W.-G. Früh and P. Read, *Wave interactions and the transition to chaos of baroclinic waves in a thermally driven rotating annulus*, Philos. Trans. Royal Soc. London A, 355 (1997), pp. 101–153.

[6] M. Ghil and P. Childress, *Topics in Geophysical Fluid Dynamics*, Appl. Math. Sci. 60, Springer-Verlag, New York, 1987.

[7] J. Gierling and R. Eckhard, *Double Hopf bifurcation and interaction of Rossby-waves induced by the zonal flow*, Contrib. Atmospheric Phys., 71 (1998), pp. 427–444.

[8] W. J. F. Govaerts, *Numerical Methods for Bifurcations of Dynamical Equilibria*, SIAM, Philadelphia, 2000.

[9] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields*, Appl. Math. Sci. 42, Springer-Verlag, New York, 1983.
[10] J. Hart, *Wavenumber selection in nonlinear baroclinic instability*, J. Atmospheric Sci., 38 (1984), pp. 400–408.
[11] B. Hassard, N. Kazarinoff, and Y.-H. Wan, *Theory and Applications of Hopf Bifurcation*, London Math. Soc. Lecture Note Ser. 41, Cambridge University Press, London, Cambridge, 1981.
[12] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Math. 840, Springer-Verlag, Berlin, 1981.
[13] R. Hide and J. Mason, *Sloping convection in a rotating fluid*, Adv. Geophys., 24 (1975), pp. 47–100.
[14] P. Hignett, *Characteristics of amplitude vacillation in a differentially heated rotating fluid annulus*, Geophys. Astrophys. Fluid Dynam., 31 (1985), pp. 247–281.
[15] P. Hignett, A. White, R. Carter, W. Jackson, and R. Small, *A comparison of laboratory measurements and numerical simulations of baroclinic wave flows in a rotating cylindrical annulus*, Quart. J. Roy. Meteorolog. Soc., 111 (1985), pp. 131–154.
[16] E. Koschmieder and H. White, *Convection in a rotating, laterally heated annulus: The wave number transitions*, Geophys. Astrophys. Fluid Dynam., 18 (1981), pp. 279–299.
[17] R. B. Lehoucq, D. C. Sorensen, and C. Yang, *ARPACK Users' Guide: Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods*, Software Environ. Tools 6, SIAM, Philadelphia, 1998.
[18] G. Lewis, *Double Hopf Bifurcations in Two Geophysical Fluid Dynamics Models*, Ph.D. thesis, University of British Columbia, Vancouver, 2000.
[19] G. Lewis and W. Nagata, *Linear stability analysis for the differentially heated rotating annulus*, submitted to Geophys. Astrophys. Fluid Dynam.
[20] J. Mansbridge, *Wavenumber transition in baroclinically unstable flows*, J. Atmospheric Sci., 41 (1984), pp. 925–930.
[21] I. Moroz and P. Holmes, *Double Hopf bifurcation and quasi-periodic flow in a model for baroclinic instability*, J. Atmospheric Sci., 41 (1984), pp. 3147–3160.
[22] R. Pfeffer and W. Fowlis, *Wave dispersion in a rotating, differentially heated cylindrical annulus of fluid*, J. Atmospheric Sci., 25 (1968), pp. 361–371.
[23] D. Rand, *Dynamics and symmetry. Predictions for modulated waves in rotating fluids*, Arch. Ration. Mech. Anal., 79 (1982), pp. 1–37.
[24] D. Tritton, *Physical Fluid Dynamics*, Oxford University Press, Oxford, UK, 1988.
[25] W. Weimer and H. Haken, *Chaotic behavior and subcritical formation of flow patterns of baroclinic waves for finite dissipation*, J. Atmospheric Sci., 46 (1989), pp. 1207–1218.
[26] S. Wiggins, *Introduction to Applied Nonlinear Dynamical Systems and Chaos*, Texts Appl. Math. 2, Springer-Verlag, New York, 1990.
[27] G. Williams, *Baroclinic annulus waves*, J. Fluid Mech., 49 (1971), pp. 417–449.

SIAM J. APPL. MATH.
Vol. 63, No. 3, pp. 1056–1082

DEPINNING TRANSITIONS IN DISCRETE REACTION-DIFFUSION EQUATIONS*

A. CARPIO† AND L. L. BONILLA‡

Abstract. We consider spatially discrete bistable reaction-diffusion equations that admit wave front solutions. Depending on the parameters involved, such wave fronts appear to be pinned or to glide at a certain speed. We study the transition of traveling waves to steady solutions near threshold and give conditions for front pinning (propagation failure). The critical parameter values are characterized at the depinning transition, and an approximation for the front speed just beyond threshold is given.

Key words. discrete reaction-diffusion equations, traveling wave fronts, propagation failure, wave front depinning

AMS subject classifications. 34E15, 92C30

PII. S003613990239006X

1. Introduction.

Spatially discrete systems describe physical reality in many different fields: atoms adsorbed on a periodic substrate [13], motion of dislocations in crystals [32], propagation of cracks in a brittle material [35], microscopic theories of friction between solid bodies [18], propagation of nerve impulses along myelinated fibers [23, 24], pulse propagation through cardiac cells [24], calcium release waves in living cells [6], sliding of charge density waves [19], superconductor Josephson array junctions [39], or weakly coupled semiconductor superlattices [3, 9]. No one really knows why, but spatially discrete systems of equations often have smooth solutions of the form $u_n(t) = u(n - ct)$, which are monotone functions approaching two different constants as $(n - ct) \to \pm\infty$. Existence of such *wave front* solutions has been proved for particular discrete systems having dissipative dynamics [40]. In the case of discrete systems with conservative dynamics, a wave front solution was explicitly constructed by Flach, Zolotaryuk, and Kladko [16]. However, a general proof of wave front existence for discrete conservative systems with bistable sources is lacking.

A distinctive feature of spatially discrete reaction-diffusion systems (not shared by continuous ones) is the phenomenon of wave front pinning: for values of a control parameter in a certain interval, wave fronts joining two different constant states fail to propagate [24]. When the control parameter surpasses a threshold, the wave front depins and starts moving [23, 19, 32, 9]. The existence of such thresholds is thought to be an intrinsically discrete fact, which is lost in continuum approximations. The characterization of propagation failure and front depinning in discrete systems is thus an important problem, which is not yet well understood despite the numerous inroads made in the literature [23, 6, 19, 32, 25, 26, 28, 30, 27, 36, 37, 38].

*Received by the editors April 3, 2002; accepted for publication (in revised form) August 20, 2002; published electronically February 25, 2003. This research was supported by the Spanish MCyT grant BFM2002-04127-C02, by the Third Regional Research Program of the Autonomous Region of Madrid (Strategic Groups Action), and by the European Union under grant RTN2-2001-00349.
http://www.siam.org/journals/siap/63-3/39006.html

†Departamento de Matemática Aplicada, Universidad Complutense de Madrid, 28040 Madrid, Spain (carpio@mat.ucm.es).

‡Departamento de Matemáticas, Escuela Politécnica Superior, Universidad Carlos III de Madrid, Avenida de la Universidad 30, 28911 Leganés, Spain, and Unidad Asociada al Instituto de Ciencia de Materiales de Madrid (CSIC), 28049 Cantoblanco, Spain (bonilla@ing.uc3m.es).

In this paper, we study front depinning for infinite one-dimensional nonlinear spatially discrete reaction-diffusion (RD) systems. When confronted with a spatially discrete RD system, a possible strategy is to approximate it by a continuous RD system. For generic nonlinearities, the width of the pinning interval is exponentially small as the continuum limit is approached. Pinning in the continuum limit has been analyzed by many authors using exponential asymptotics, also known as asymptotics beyond all orders. As far as we can tell, usage of these techniques for discrete equations goes back to two classic papers by Indenbom [22] (for the FK potential) and by Cahn [7] (for the double-well potential). In both cases, an exponential formula for the critical field was derived by means of the Poisson sum rule. In the context of dislocation motion, exponential formulas for the depinning shear stress of the Peierls–Nabarro (PN) model were found earlier by Peierls [33] and Nabarro [31]. Descriptions of wave front pinning near the continuum limit can also be found in more recent work [20, 25, 27].

Analyzing the continuum limit of a discrete system by means of exponential asymptotics is a costly strategy for describing pinning for two reasons. It is not numerically accurate as we move away from the continuum limit, and it ceases to be useful if convective terms [11] or disorder [12] alter the structure of the discrete system (quite common in applications). Thus other authors have tried to describe the opposite strongly discrete limit. For discrete RD equations, Erneux and Nicolis [14] studied a finite discrete RD equation with a cubic nonlinearity, a Dirichlet boundary condition at one end, and a Neumann boundary condition at the other end. They considered a particular limit in which two of the three zeroes of the cubic nonlinearity coalesced as diffusivity went to zero. Erneux and Nicolis's calculation is essentially a particular case of our active point approximation that involves only one active point and makes an additional assumption on the nonlinearity (not needed in our calculations). They found that the wave front velocity scales as the square root of $(d-d_c)$ (d is the diffusivity and d_c its critical value at which wave fronts are pinned). Essentially the same results can be found in the appendix of [27]. Kladko, Mitkov, and Bishop [28] introduced an approximation called the single active site theory. In this approximation, the wave front is described by two linear tails (solution of the RD equation linearized about each of the two constants joined by the front) *patched* at one point. This approximation is used to estimate the critical field for wave front depinning.

By a combination of numerical and asymptotic calculations, we arrive at the following description [10, 11]. The nature of the depinning transition depends on the nonlinearity of the model and is best understood as propagation failure of the traveling front. Usually, but not always, the wave front profiles become less smooth as a parameter F (external field) decreases. They become *discontinuous* at a critical value F_c. Below F_c, the front is pinned at discrete positions corresponding to a stable steady state. As a consequence of the maximum principle for spatially discretized parabolic equations, stationary and moving wave fronts cannot simultaneously exist for the same value of F (see [8]). This is *not* the case for chains with conservative dynamics, which are spatially discretized hyperbolic equations without a maximum principle. For chains with conservative Hamiltonian dynamics, an inverse method due to Flach, Zolotaryuk, and Kladko [16] explicitly shows that stationary and moving wave fronts may coexist for the same value of the parameters.

We consider chains of diffusively coupled overdamped oscillators in a potential V, subject to a constant external force F:

$$\frac{du_n}{dt} = u_{n+1} - 2u_n + u_{n-1} + F - A\,g(u_n). \tag{1.1}$$

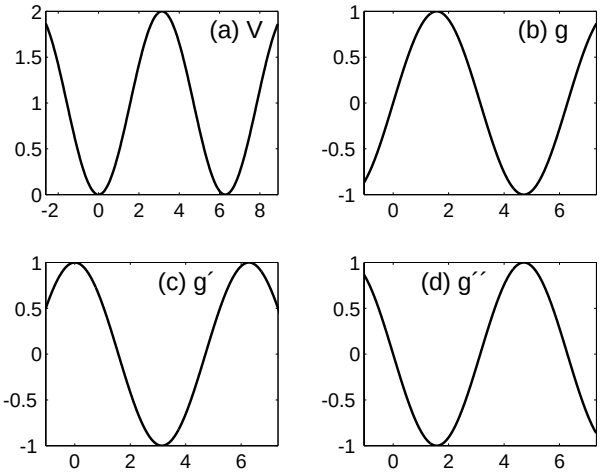


FIG. 1.1. *FK model:* (a) *potential* $1-cos(x)$, (b) *source term* $g(u) = \sin(u)$, (c) $g'(u) = \cos(u)$, (d) $g''(u) = -\sin(u)$.

Here $g(u) = V'(u)$ is at least C^1, and it presents a "cubic" nonlinearity (see Figure 1.1), such that $A\,g(u) - F$ has three zeroes, $U_1(F/A) < U_2(F/A) < U_3(F/A)$ in a certain force interval ($g'(U_i(F/A)) > 0$ for $i = 1, 3$, $g'(U_2(F/A)) < 0$). Provided that $g(u)$ is odd with respect to $U_2(0)$, there is a symmetric interval $|F| \le F_c$ where the discrete wave fronts joining the stable zeroes $U_1(F/A)$ and $U_3(F/A)$ are pinned [23, 8]. For $|F| > F_c$, there are smooth traveling wave fronts, $u_n(t) = u(n - ct)$, with $u(-\infty) = U_1$ and $u(\infty) = U_3$, as proved in [40, 8]. The velocity $c(A, F)$ depends on A and F, and it satisfies $cF < 0$ and $c \to 0$ as $|F| \to F_c$ (see [8]). Examples are the overdamped Frenkel–Kontorova (FK) model ($g = \sin u$; see Figure 1.1) [17] and the quartic double well potential ($V = (u^2 - 1)^2/4$). Less symmetric nonlinearities yield a nonsymmetric pinning interval, and our analysis applies to them with trivial modifications. Note that coexistence of fronts traveling in opposite directions can occur in the case of conservative systems, but not for (1.1) due to the maximum principle (which is the basis of comparison techniques) [8].

For the overdamped FK model given by (1.1) with $g = \sin u$, Figure 1.2 shows wave front profiles near the critical field. Individual points undergo abrupt jumps at particular times, which gives the misleading impression that the motion of the discrete fronts proceeds by successive jumps. Actually, the points remain very close to their stationary values at $F = F_c$, say $u_n(A, F_c)$, during a very long time interval of order $|F - F_c|^{-\frac{1}{2}}$. Then, at a specific time, *all* the points $u_n(t)$ jump to a vicinity of $u_{n+1}(A, F_c)$. The method of matched asymptotic expansions can be used to describe this two-stage motion of the points $u_n(t)$. Then the wave front profile can be reconstructed by using the definition $u_n(t) = u(n - ct)$. The slow stage of front motion is described by the normal form of a saddle-node bifurcation, and it yields an approximation to the wave front velocity, which scales with the field as $|F-F_c|^{\frac{1}{2}}$. This scaling has been mentioned by other authors: it was found numerically in [1], and by means of exponential asymptotics in the limit A small in [27]. It is also conjectured in [26] on the correct basis that the depinning transition consists of a saddle-node bifurcation (a similar claim was stated in [30] for continuous reaction diffusion equations with localized sources). However, the derivation of the *local* saddle-node normal form and

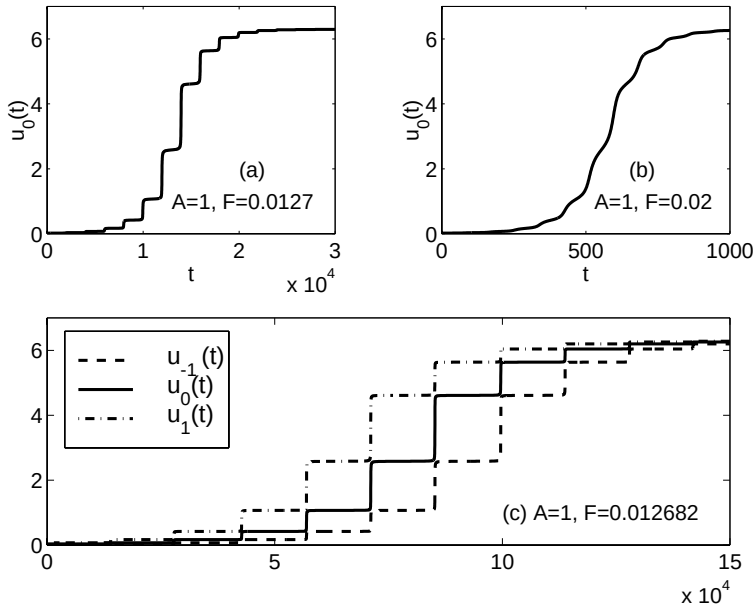


FIG. 1.2. *Wave front profiles for the overdamped FK model when* $A = 1$ *near* F_c.

the correct description of the *global* saddle-node bifurcation involving matching with a fast stage during which the front jumps abruptly one lattice period were apparently omitted by the authors of [26], who used energy arguments. Our picture of the wave front depinning transition has essentially been corroborated in the continuum limit (as an appropriate dimensionless lattice length goes to zero) by King and Chapman, who used asymptotics beyond all orders [27]. An independent confirmation follows from Fáth's calculations for a spatially discrete reaction-diffusion equation with a piecewise linear source term [15] (except that the velocity should scale differently with $|F - F_c|$ in this case).

For exceptional nonlinearities, the wave front does not lose continuity as the field decreases. In this case, there is a continuous transition between wave fronts moving to the left for $F > 0$ and moving to the right for $F < 0$; as for continuous systems, front pinning occurs at only a single field value $F = 0$ (see [27, 16, 36, 37, 38]). Wave front velocity then scales linearly with the field. We discuss the characterization of the critical field (including analytical formulas in the strongly discrete limit), describe depinning anomalies (discrete systems having zero critical field [36, 37, 38, 16]), and give a precise characterization of stationary and moving fronts near depinning (including front velocity) by singular perturbation methods. Our approximations show excellent agreement with numerical simulations.

The rest of the paper is organized as follows. In section 2, we characterize wave front depinning. We also explain that pinning of wave fronts normally occurs at force values belonging to an interval with nonzero length. However, there are nonlinearities for which pinning occurs only at $F = F_c = 0$. In section 3, we present a theory of wave front depinning for the strongly discrete case (A large). This theory enables us to predict the critical field and the speed and shape of the wave fronts near threshold. The main ideas of our theory are very simple. First, a wave front profile $u_n(t) =$

$u(n-ct)$ can be reconstructed if we follow the motion of one point during a sufficiently long time interval. Secondly, the analysis of (1.1) is complicated by the presence of the discrete diffusion term $u_{n+1}-2u_n+u_{n-1}$. Previous authors have tried to approximate this term by its continuum limit (corresponding to $A \to 0$), which leads to using exponential asymptotics [27]. (See also [25] on using exponential asymptotics for the Hamiltonian version of our model.) However, we are only interested in constructing solutions of (1.1) joining constant values. For sufficiently large A (say, $A = 0.1$ for the FK model), u_i is approximately either $U_1(F/A)$ or $U_3(F/A)$ except for a finite number of points (the *active* points). Then we can approximate the infinite system (1.1) by a closed system of ordinary differential equations (only one equation for $A \geq 10$ in the FK model). The depinning transition is a global bifurcation of this system, as explained in section 3. Some auxiliary technical results are collected in the appendix.

2. Front pinning as propagation failure. To describe monotone stationary solutions of (1.1) joining $U_1(F/A)$ and $U_3(F/A)$ for $|F| \leq F_c$, it is better to start by considering traveling wave fronts for $|F| > F_c$. It has been proved (and corroborated by numerical calculations) that traveling wave fronts and stationary profiles cannot coexist at the same value of F (see [9]). Furthermore, numerical computations of wave fronts near the critical fields F_c for the FK and other usual potentials show staircase-like wave front profiles, which sharpen as F approaches F_c. At $F = F_c$, a series of gaps open up, and one is left with a discontinuous stationary profile $s(x)$ solving

$$\begin{aligned} s(x+1) - 2s(x) + s(x-1) &= Ag(s(x)) - F_c, \quad x \in \mathbb{R}, \\ s(-\infty) = U_1\left(\frac{F_c}{A}\right), &\quad s(\infty) = U_3\left(\frac{F_c}{A}\right). \end{aligned} \tag{2.1}$$

The profile $s(x)$ is increasing and piecewise constant. The sequence of constant values attained by $s(x)$ defines a steady solution u_n of (1.1) with $F = F_c$. A stationary solution can thus be understood as a wave front that fails to propagate and is *pinned* at discrete values. Figure 1.2 illustrates the pinning transition for the FK model with $A = 1$. As F decreases from 0.02 to 0.0127, a series of steps are formed. Figure 1.2(c) depicts the paths described by three consecutive points. All profiles look identical and are obtained by shifting any one of them some multiple of a certain constant length. This implies that the length of all steps in the profile is the same and that all the points $u_n(t)$ in (1.1) proceed to climb the next step in the staircase at the same time. This behavior indicates that the wave front is a traveling wave, $u_n(t) = u(n - ct)$. Proofs of this fact for some sources can be found in [40].

2.1. Limiting front profile at the critical field. Let us start by showing that the limit of the traveling waves as $F \to F_c$ is singular if $F_c > 0$. This fact can be guessed from the differential-difference equations satisfied by the wave profiles. The traveling waves for $|F| > F_c$ have the form $u_n(t) = u(n - ct)$, where the profile $u(z)$ solves (see [8])

$$\begin{aligned} -cu_z &= u(z+1) - 2u(z) + u(z-1) - Ag(u(z)) + F, \quad z \in \mathbb{R}, \\ u(-\infty) = U_1\left(\frac{F}{A}\right), &\quad u(\infty) = U_3\left(\frac{F}{A}\right). \end{aligned} \tag{2.2}$$

The solution u is as smooth as allowed by $g(u)$. (u is C^{k+1} if $g(u)$ is C^k, with $k \geq 1$.) Then multiplying (2.2) by u_z and integrating it, we get

$$-c\int_{-\infty}^{\infty} u_z^2 dz = F\left[U_3\left(\frac{F}{A}\right) - U_1\left(\frac{F}{A}\right)\right]. \tag{2.3}$$

A first obvious conclusion is that the sign of c is opposite to the sign of F. Let F_c be positive. As $F \to F_c$, $c \to 0$ and $F\,[U_3(F/A) - U_1(F/A)] \to F_c\,[U_3(F_c/A) - U_1(F_c/A)] \neq 0$. Therefore the integrals $\int u_z^2 dz \to \infty$ as $F \to F_c$. Thus the limiting profile must be discontinuous if $F_c > 0$.

If $F_c = 0$, the relation (2.3) can be used to show that (2.1) has a smooth solution. In fact, provided that $c \sim -K\,F$ $(K > 0)$ as $F \to 0$, we can use (2.3) to uniformly bound the derivatives of the solutions u in (2.2) for $F \neq 0$. Then we obtain a smooth solution of (2.1) in the limit as $F \to 0$. We will come back to this question later on in subsection 2.3. Note that the stationary equation $s(x+1)-2s(x)+s(x-1) = Ag(s)-F$ has no continuous solutions joining $U_1(F/A)$ to $U_3(F/A)$ unless $F = 0$. To see this [8], we multiply the equation by s_x (in the sense of distributions if necessary) and integrate to get $F = 0$.

2.2. Characterization of the critical field. Some results are available in the continuum limit $A \to 0$. For $g = \sin u$, it is well known that F_c vanishes exponentially fast as A goes to zero. An exponential formula for F_c was first found by Indenbom [22] using the Poisson sum rule (following the calculations of the PN energy barrier for the PN model by Peierls [33] and Nabarro [31]) and numerically checked by Hobart [21] in the context of the Peierls stress and energy for dislocations. For the discrete bistable RD equation, Cahn [7] derived an exponential dependence of F_c by a similar technique. Related ideas can be found in Kladko, Mitkov, and Bishop [28]. These arguments can be used for other potentials and suggest that $F_c \sim C\,e^{-\eta/\sqrt{A}}$ as $A \to 0+$ (with positive C and η independent of A) holds for a large class of nonlinearities. Using exponential asymptotics, King and Chapman [27] have obtained precise formulas for the critical field and the wave front velocity of a discrete RD equation. Particularized to the FK potential, their formulas for the critical field and for the wave front velocity after depinning are $F_c \sim \Lambda\, e^{-\pi^2/[2\sinh^{-1}(\sqrt{A}/2)]}$, $\Lambda \approx 356.1$, and $c \sim D\,\sqrt{(F^2 - F_c^2)/A}$, respectively. This latter result agrees with the scaling law $c \sim |F - F_c|^{\frac{1}{2}}$, found in a large class of discrete RD equations [9, 10, 11, 26] and in continuous equations with localized sources [30]. However, exponential asymptotics [27] does not work for A large. We shall therefore follow a different approach. We shall begin by considering stationary increasing discrete front profiles and study under which conditions they start moving. Since stationary fronts are pinned wave fronts, we can call the transition from stationary to moving fronts the *depinning transition*.

Two facts distinguish the depinning transition: (i) the smallest eigenvalue of (1.1) linearized about a stable stationary profile becomes zero (see below), and (ii) stationary and moving wave fronts cannot coexist for the same values of the field. First, the following comparison principle [23] for (1.1) can be used to show that stationary and traveling wave fronts cannot coexist for the same value of F (see [8]).

Comparison principle. Assume that we have two configurations $w_n(t)$ and $l_n(t)$. If initially $w_n(0) \geq l_n(0)$ for all n, and at any later time $t > 0$

$$\frac{dw_n}{dt} \geq w_{n+1} - 2w_n + w_{n-1} - Ag(w_n) + F, \tag{2.4}$$

$$\frac{dl_n}{dt} \leq l_{n+1} - 2l_n + l_{n-1} - Ag(l_n) + F \tag{2.5}$$

for all $n \in \mathbb{Z}$, then necessarily $w_n(t) \geq l_n(t)$ for all n and t. Here w_n satisfying (2.4) is said to be a supersolution, and l_n satisfying (2.5) is said to be a subsolution.

Front pinning can be proved using stationary sub- and supersolutions, which can be constructed, provided that the stationary solution is linearly stable. The

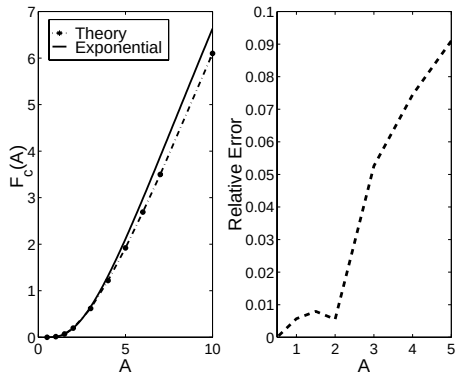


FIG. 2.1. (a) *Critical field, $F_c(A)$, for $A \in (\frac{1}{2}, 10)$. We have compared the result of approximately solving $\lambda_1(A, F) = 0$ for F as a function of A (see the appendix) to the asymptotic result $F_c \sim 356.1\, e^{-\pi^2/[2\sinh^{-1}(\sqrt{A}/2)]}$ of* [27]. (b) *Relative error of the exponential asymptotics approximation.*

smallest eigenvalue of the linearization of (1.1) about a stationary profile $u_n(A, F)$, $u_n(t) = u_n(A, F) + v_n e^{-\lambda t}$, is given by

$$\lambda_1(A, F) = \min \frac{\sum[(v_{n+1} - v_n)^2 + Ag'(u_n(A, F))v_n^2]}{\sum v_n^2}, \tag{2.6}$$

over a set of functions v_n, which decay exponentially as $n \to \pm\infty$. We show in the appendix that the minimum is attained at a positive eigenfunction.

The critical field can be uniquely characterized by $\lambda_1(A, F_c) = 0$ and $\lambda_1(A, F) > 0$ for $|F| < F_c$. The details are given in the appendix. Notice that $\lambda_1(A, F) > 0$ implies that (2.1) does not have smooth solutions $s(x)$; otherwise, $v_n = s'(n)$ is an eigenfunction corresponding to $\lambda_1 = 0$ as it happens in the continuum limit. The previous characterization is the basis of a procedure for calculating $F_c(A)$. In section 3, we shall show that wave fronts near the depinning transition are described by a reduced system of equations for a finite number of points $u_n(t)$ which "jump" from about a discrete value corresponding to the stationary solution, $u_n(A, F_c)$, to the next one, $u_{n+1}(A, F_c)$, during front motion. The smallest eigenvalue for the linearization of the reduced system of equations about a stationary solution approximates λ_1 well. The critical field obtained by this procedure has been depicted in Figure 2.1 for the FK potential and compared to King and Chapman's asymptotic result (obtained by keeping two terms in their formulas). Notice that the asymptotic result loses accuracy as A increases.

Equation (2.6) shows that the critical field is positive for large A and typical nonlinearities. In fact, consider the FK potential. For $F = 0$ there are two one-parameter families of stationary solutions which are symmetric with respect to U_2 (see Figure 2.2), one taking on the value U_2 (unstable dislocation), and the other one having $u_n \neq U_2$ (stable dislocation) [21, 8]. The centers of two stable (or two unstable) dislocations differ in an integer number of lattice periods. Except for a possible rigid shift, the stable dislocation, $u_n(A, 0)$, is a dynamically stable stationary solution towards which step-like initial conditions evolve. Figures 2.3(a) and (b) show two initial conditions that evolve (exponentially fast) towards the stable dislocation. Half the initial points $u_n(0)$ have been selected to be below U_2, and the other half are above this value. In Figure 2.3(a), $u_n(0) - u_n(A, 0) = \epsilon_n$, where ϵ_n are real random

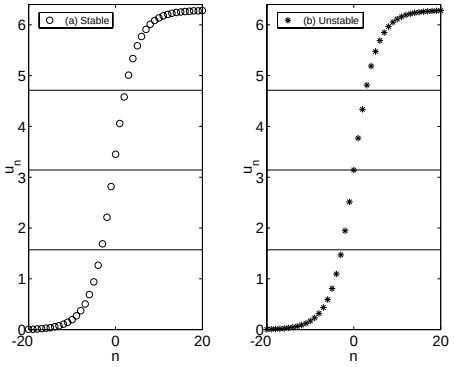


FIG. 2.2. *Stable and unstable dislocations for the FK model when* $F = 0$ *and* $A = 0.1$.

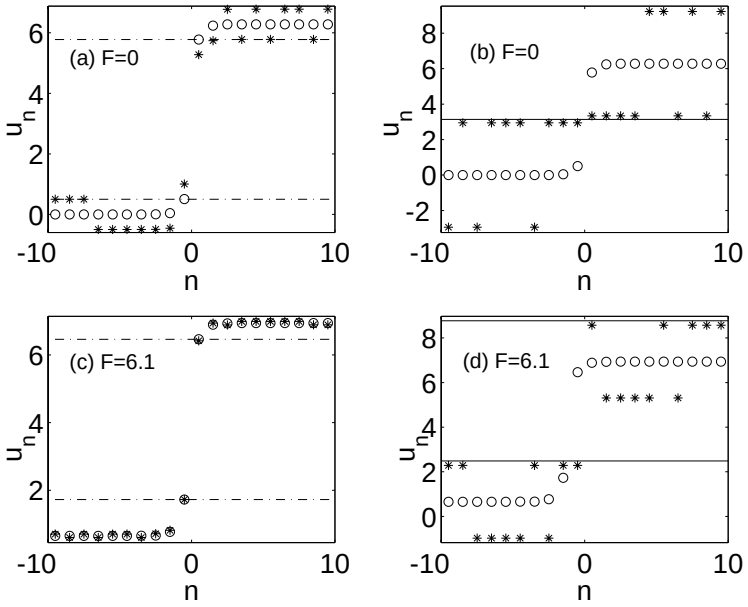


FIG. 2.3. *Initial condition* $u_n(0)$ *(asterisks) and its large time limit, the stable dislocation (circles), for the FK model with* $A = 10$*;* $F = 0$ *for* (a) *and* (b)*, and* $F = 6.1 < F_c$ *for* (c) *and* (d)*. The initial points are selected as indicated in the text.*

numbers with $|\epsilon_n| < 0.5$. In Figure 2.3(b), $u_n(0) - U_{1,3} = \delta_n B$, $0 < B = U_2 - U_1 - 0.2$, and δ_n randomly takes on the values 1 or -1. By using comparison methods, it is possible to prove that a small disturbance of the stable dislocation evolves towards it. The same results hold for the stable stationary solution $u_n(A, F)$ for $0 < |F| < F_c$. As $|F|$ increases, a disturbance of the stable stationary solution typically evolves towards the same stationary solution displaced an integer number of lattice periods unless the disturbance is sufficiently small. See Figure 2.3(d) for an example of this phenomenon for F slightly smaller than F_c. Carefully selecting the initial condition avoids this, as in Figure 2.3(c).

For large A, the stable dislocation has $g'(u_n) > 0$ for all n, and (2.6) gives

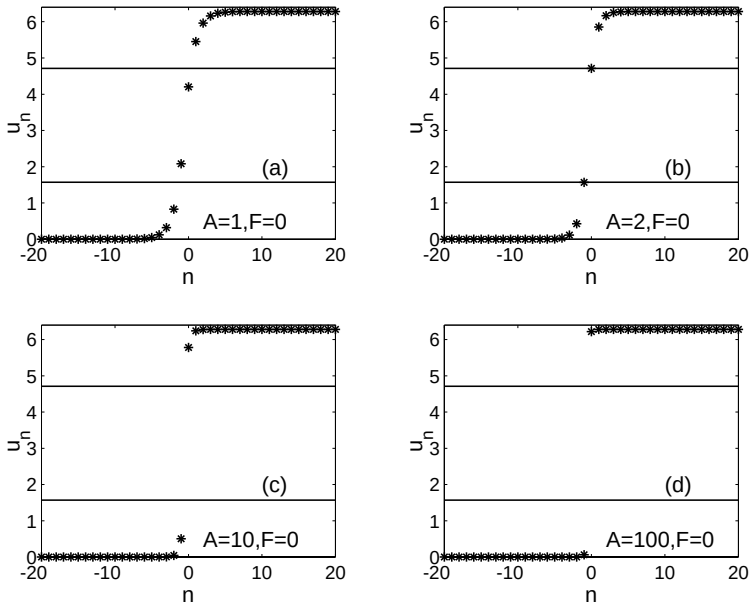


FIG. 2.4. *Stationary solutions for the FK model when* $A = 1, 2, 10, 100$.

$\lambda_1(A, 0) > 0$. Since $\lambda_1(A, F_c) = 0$, this implies that the critical field is nonzero. (Different proofs are given in [23, 9].) As $A > 0$ decreases, several u_n may enter the region of negative slope $g'(u)$: the number of points with $g'(u_n) < 0$ increases as A decreases; see Figures 2.2 and 2.4. It should then be possible to have $\lambda_1(A, 0) = 0$, i.e., $F_c = 0$, for a discrete system! Examples of this *pinning anomaly* will be given next.

2.3. Pinning failure. Despite widespread belief, it is not true that the critical field is positive for all discrete systems. This point was already raised by Hobart [21], who proposed the following numerical criterion to check whether for a given source g the critical field for (1.1) is zero.

Let us assume for the sake of simplicity that g is odd about 0. Then $U_2(0) = 0$ and $U_1(0) = -U_3(0)$. For any $x \in (U_1(0), U_3(0))$, we can compute numerically a unique value $y(x)$ such that the sequence u_n defined by $u_0 = x$, $u_1 = y(x)$, and $u_n = 2u_{n-1} - u_{n-2} + g(u_{n-1}), n > 1$, tends to $U_3(0)$ as $n \to \infty$. Hobart conjectured that $F_c = 0$ for a given nonlinearity g, provided that the function $y(x)$ satisfies

$$y^{-1}(x) = -y(-x), \quad y(x) - y(-x) = 2x + g(x) \tag{2.7}$$

for $x \in (U_1(0), U_3(0))$. It is fairly easy to construct examples of nonlinearities $g(x)$ for which (2.7) holds. It suffices to choose some smooth odd increasing function $u(x)$ such that $u(x) \to \pm a$ as $x \to \pm\infty$ for some $a > 0$. We define $g(u(x)) = u(x+1) - 2u(x) + u(x-1)$ so that $g(z) = u(u^{-1}(z) + 1) - 2z + u(u^{-1}(z) - 1)$ and $y(z) = u(u^{-1}(z) + 1)$. Choosing $u(x) = \tanh(x)$ (see [34, 5, 16]), we get an explicit formula for g: $g(z) = -2\gamma z(1 - z^2)/(1 - \gamma z^2)$ with $\gamma = \tanh^2(1)$. Notice that one or two points of the stationary solutions, $u_n = \tanh(n + p)$ (p is any constant), enter the region where $g' < 0$; see Figure 2.5(a).

By following this procedure, we find examples of bistable source terms for which

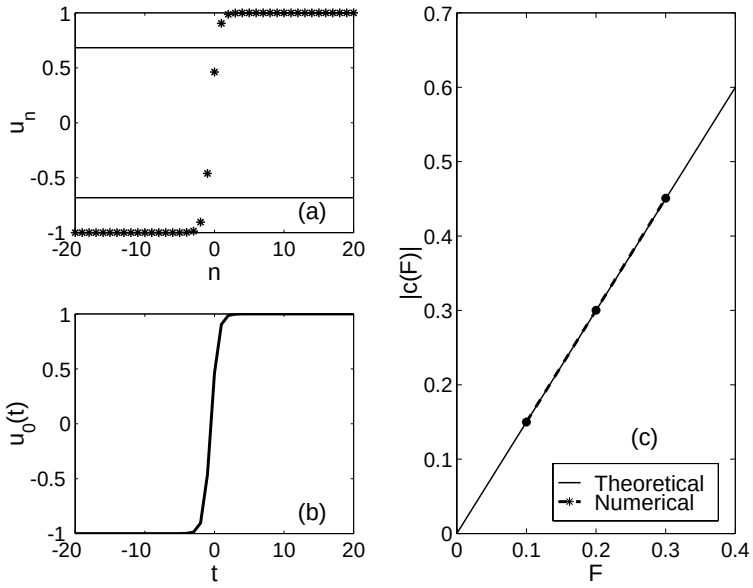


FIG. 2.5. (a) *Stationary solution* $u_n = \tanh(n)$, (b) *wave front for* F *small,* (c) *numerically calculated versus predicted speed for* $g(u) = -2\gamma u(1-u^2)/(1-\gamma u^2)$, *with* $\gamma = \tanh^2(1)$.

(1.1) has a uniparametric family of continuous stationary solutions, $u_n = u(n+p)$, $0 \le p < a$, satisfying $u_{n+1} - 2u_n + u_{n-1} = g(u_n)$ and $u_{-\infty} = -a$, $u_\infty = a$. In this case, (1.1) does not have stationary solutions joining $U_1(F/A)$ and $U_3(F/A)$ unless $F = 0$ (see [8]). The existence of continuous steady solutions for $F = 0$ implies that there is a continuous transition from wave fronts traveling to the left ($c < 0$) for $F > 0$ to wave fronts traveling to the right ($c > 0$) for $F < 0$. Only at $F = 0$ are wave fronts stationary (pinned). This pinning anomaly is stated more precisely as follows.

THEOREM 2.1. *Let* $g \in C^2$ *be as in the Introduction with* $g(0) = 0$, *and let* $\mathcal{L}(F)$ *be the operator*

$$\mathcal{L}(F)\, v_n = A\, g'(u_n) v_n + 2v_n - v_{n+1} - v_{n-1}, \tag{2.8}$$

corresponding to the evolution equation (1.1) *linearized about the stationary solution* $u_n = u_n(A, F)$ *at field* F. *Let us assume that for* $F = 0$ *there exists a differentiable increasing stationary solution* $u(x)$ *such that* $u(x) \to \pm U_3(0)$ *as* $x \to \pm\infty$. *Then,*

1. *zero is the smallest eigenvalue of the operator* $\mathcal{L}_0 = \mathcal{L}(0)$, *corresponding to the evolution equation* (1.1) *linearized about the stationary solution* $u_n(A, 0) = u(n)$;

2. $F_c(A) = 0$ *for* (1.1);

3. *traveling wave fronts exist for all* $F \neq 0$. *Furthermore, their speed increases linearly with the force for small* F. *We have*

$$c \sim -F\, \frac{U_3(0) - U_1(0)}{\int_{-\infty}^{\infty} \left(\frac{du}{dx}\right)^2 dx} \tag{2.9}$$

as $F \to 0$.

Moreover, statement 3 *implies the existence of steady differentiable solutions* $u(x)$ *of* (1.1) *such that* $u(x) \to \pm U_3(0)$ *as* $x \to \pm\infty$ *for* $F = 0$.

It is not our goal here to give a rigorous proof of this result, but to sketch the main ideas. First of all, note that the derivative $v_n = u_x(n) > 0$ is a *positive* eigenfunction of the elliptic operator $\mathcal{L}_0$ corresponding to the eigenvalue $\lambda = 0$ and decaying exponentially at infinity. Statement 1 immediately follows. This fact can be used to construct propagating sub- and supersolutions for (1.1) which forbid pinning for any $F \neq 0$. Thus, $F_c = 0$, which is statement 2. For $F = \epsilon > 0$ sufficiently small, the propagating subsolutions are $l_n(t) = l(n+\epsilon c_0 t)$, with $c_0 > 0$ and $l(x) = u(x)+\epsilon u_x(x)$. For $F = -\epsilon$, the propagating supersolutions are $w_n(t) = w(n-\epsilon c_0 t)$, with $c_0 > 0$ and $w(x) = u(x) - \epsilon u_x(x)$. In both cases, we have to choose $c_0 < 1/\max(u_x)$. A subsolution traveling to the left "pushes" the fronts to the left. Similarly, the supersolutions traveling to the right "push" the fronts to the right.

Let us now obtain statement 3. If $|F| > 0$, we have traveling wave front solutions $u_n(t) = \mathcal{U}(n - ct)$ of (1.1), whose profile $\mathcal{U}(z)$ satisfies the differential-difference equation

$$-c\frac{d\mathcal{U}}{dz}(z) = \mathcal{U}(z+1) - 2\mathcal{U}(z) + \mathcal{U}(z-1) - g(\mathcal{U}(z)) + F, \tag{2.10}$$

and $\mathcal{U}(\pm\infty) = \pm U_3(0)$; see [8]. Let $F = F_0\epsilon$ with $0 < \epsilon \ll 1$. The traveling wave solution can be written as $\mathcal{U}(n-ct) = u(n-ct) + \epsilon w(n-ct) + o(\epsilon)$, where $u(x)$ is the smooth stationary profile. Let $z = n - ct$ and $c = c_0\epsilon + o(\epsilon)$. Then w obeys

$$w(z+1) - 2w(z) + w(z-1) - Ag'(u(z))w(z) = -c_0\frac{du}{dz}(z) - F_0,$$

$$w(-\infty) = w(\infty) = \frac{1}{g'(U_1)} = \frac{1}{g'(U_3)}.$$

By the Fredholm alternative, this linear nonhomogeneous equation has a solution if the left-hand side $-c_0 du/dz - F_0$ is orthogonal to the eigenfunction du/dz, which yields (2.9).

In section 3, we show that the wave front speed c scales as $|F-F_c|^{\frac{1}{2}}$ if $F_c > 0$. Our linear scaling (2.9) of the velocity in statement 3 therefore implies that $F_c = 0$. The linear scaling (2.9) with $F_c = 0$ implies the existence of smooth stationary solutions at $F = 0$ as discussed in the first subsection.

Remark 1. We conjecture that the three statements in Theorem 2.1 are equivalent. To prove this, it would be enough to show that $F_c = 0$ implies the linear scaling of the speed of the waves (statement 3). Then, existence of differentiable stationary solutions follows. This implies statement 1 ($\lambda_1(A, 0) = 0$), which implies statement 2 ($F_c = 0$), as we showed above.

Remark 2. When stationary wave front solutions have smooth profiles, pinning failure occurs for discrete RD equations and for discrete equations with conservative dynamics. In the latter case, translation-invariant smooth profiles have the same energy, and therefore the PN energy barrier (defined as the smallest energy barrier that must be overcome for a kink or wave front to move [4]) vanishes. Pinning of a wave front usually results if the energy difference between the stable and the unstable front solutions (see Figure 2.2) is not zero. This energy difference provides an estimation of the PN energy barrier. Discussions of the PN potential and the PN barrier can be found in section 2.3 of [4] and in section III.B of [16]. The mathematical meaning and usefulness of the PN barrier for an infinite system with conservative dynamics are worth studying.

Remark 3. Speight and Ward [36] and Speight [37, 38] have developed a technique to discretize some continuum conservative models in such a way that kink-like initial

profiles may propagate without getting trapped. Their idea is to seek a discrete version of the potential energy which admits minimals satisfying a first order difference equation called the Bogomol'nyi equation so that there is no PN barrier. On the other hand, the difference operators in Speight and Ward [36] discretized equations of motion have a structure different from discrete diffusion and are hard to justify physically.

Remark 4. In discrete RD equations, moving and pinned fronts cannot coexist for the same value of the applied field. For chains with conservative Hamiltonian dynamics, the situation is less clear. In fact, it is possible to have two stationary front solutions with a positive energy difference between them (which would imply a nonzero critical field and therefore wave front pinning according to general belief), and yet a moving wave front may coexist with the stationary fronts for the same parameter values. An explicit example of this situation has been constructed by Flach, Zolotaryuk, and Kladko using an inverse method [16].

3. Asymptotic theory of wave front depinning. In this section we introduce a systematic procedure for deriving analytic expressions for the critical field $F_c > 0$ as a function of A, and for the front profiles and their velocity as functions of $F - F_c$ and A. Our methods work best in the strongly discrete case for large A. Our ideas are quite general and may be applied successfully to more complex discrete models [11]. We shall assume that $g \in C^2$ throughout this section.

3.1. Theory with a single active point. We choose A large enough for the stable dislocation in Figure 2.2 not to enter the region where $g' < 0$; see Figure 2.4. When $F > 0$, this solution is no longer symmetric with respect to U_2. If F is not too large, all $u_n(A, F)$ avoid the region of negative slope $g'(u) < 0$. For larger F and generic potentials (FK, double-well, ...), we have observed numerically that $g' < 0$ for a single point, labelled $u_0(A, F)$. This property persists until F_c is reached; see Figure 3.1.

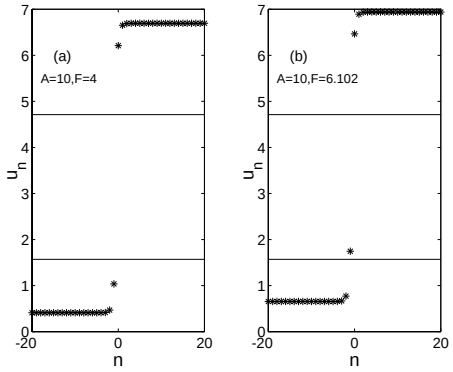


FIG. 3.1. *Stationary solutions for the FK model with* $A = 10$*:* (a) *No points are found in the region* $g' < 0$ *for sufficiently small* F*;* (b) *one point enters the region* $g' < 0$ *for sufficiently large* $F < F_c$*.*

First, consider the symmetric stationary profile with $u_n \neq U_2$ for $F = 0$. The front profile consists of two tails with points very close to U_1 and U_3, plus two symmetric points u_0, u_1 in the gap region between U_1 and U_3. As $F > 0$ increases, this profile changes slightly: the two tails are still very close to $U_1(F/A)$ and $U_3(F/A)$. As for the two middle points, u_1 gets closer and closer to U_3, whereas u_0 moves away from

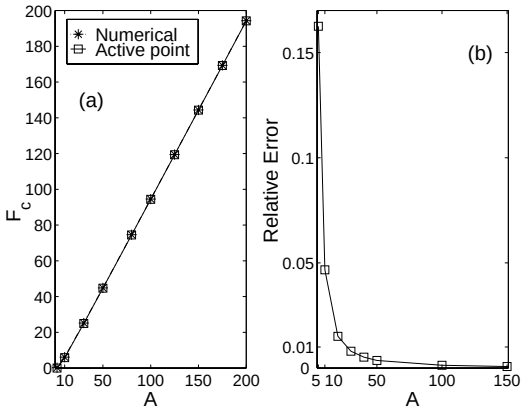


FIG. 3.2. *Approximation of* (1.1) *by the equation with one active point for the FK potential and* $A > 2$*:* (a) *critical force versus* A, (b) *error in the approximation of* $F_c(A)$.

U_1. This structure is preserved by the traveling fronts above the critical field: there is only one active point most of the time, which we can adopt as our u_0. Then the wave front profile (2.2) can be calculated as $u(-ct) = u_0(t)$. In (1.1), we can approximate $u_{-1} \sim U_1$, $u_1 \sim U_3$, thereby obtaining

$$\frac{du_0}{dt} \approx U_1\left(\frac{F}{A}\right) + U_3\left(\frac{F}{A}\right) - 2u_0 - A\,g(u_0) + F. \tag{3.1}$$

This equation has three stationary solutions for $F < F_c$, two stable and one unstable, and only one stable stationary solution for $F > F_c$. Let us consider $F < F_c$. Only two out of the three solutions of (3.1) approximate stationary fronts for the exact system: those having smaller values of u_0. The one having smallest u_0 approximates the stable stationary front; the other one approximates the unstable stationary front. Recall that the unstable front had a value $u_0 = [U_1(0) + U_3(0)]/2$ at the middle of the gap for $F = 0$. As $F > 0$ increases, u_0 decreases towards $U_1(F/A)$. Thus one active point will also approximate the profile of the unstable stationary front. The stationary solution of (3.1) having the largest value of u_0 (slightly below $U_3(F/A)$) is not consistent with the assumptions we made to derive (3.1), and therefore it does not approximate a physically existing stationary front. If $F > F_c$, the only stationary solution of (3.1) is the unphysical one. The critical field F_c is such that the expansion of the right-hand side of (3.1) about the two coalescing stationary solutions has zero linear term, $2 + Ag'(u_0) = 0$, and

$$2u_0 + A\,g(u_0) \sim U_1\left(\frac{F_c}{A}\right) + U_3\left(\frac{F_c}{A}\right) + F_c. \tag{3.2}$$

These equations for F_c and $u_0(A, F_c)$ have been solved for the FK potential, for which $u_0 = \cos^{-1}(-2/A)$ and $U_1 + U_3 = 2\sin^{-1}(F_c/A) + 2\pi$. The results are depicted in Figure 3.2 and show excellent agreement with those of direct numerical simulations for $A > 10$. Our approximation performs less well for smaller A, and it breaks down at $A = 2$ with the wrong prediction $F_c = 0$. Notice that $F_c(A)/A \sim 1$ as A increases. In practice, only steady solutions are observed for very large A.

Let us now construct the profile of the traveling wave fronts after depinning for

F slightly above F_c. Then $u_0(t) = u_0(A, F_c) + v_0(t)$ obeys the following equations:

$$\frac{dv_0}{dt} = \alpha\,(F - F_c) + \beta\, v_0^2, \tag{3.3}$$

$$\alpha = 1 + \frac{1}{A\,g'(U_1(F_c/A))} + \frac{1}{A\,g'(U_3(F_c/A))}\,, \tag{3.4}$$

$$\beta = -\frac{A}{2}\,g''(u_0), \tag{3.5}$$

where we have used $2 + Ag'(u_0) = 0$ and (3.2) and ignored terms of order $(F - F_c)\,v_0$ and higher. These terms are negligible after rescaling $v_0 = (F - F_c)^{\frac{1}{2}}\varphi$ and $\tau = (F - F_c)^{\frac{1}{2}}t$. The coefficients α and β are positive because $g'(U_i) > 0$ for $i = 1, 3$ and $g''(u_0) < 0$ since $u_0 \in (U_1(0), U_2(0))$. For the FK potential, $\alpha = 1 + 2/\sqrt{A^2 - F_c^2}$ and $\beta = \sqrt{A^2 - F_c^2}/2$. Equation (3.3) has the (outer) solution

$$v_0(t) \sim \sqrt{\frac{\alpha(F - F_c)}{\beta}}\,\tan\left(\sqrt{\alpha\beta\,(F - F_c)}\,(t - t_0)\right), \tag{3.6}$$

which is very small most of the time, but it blows up when the argument of the tangent function approaches $\pm\pi/2$. Thus the outer approximation holds over a time interval $(t - t_0) \sim \pi/\sqrt{\alpha\beta\,(F - F_c)}$, which equals $\pi\sqrt{2/\alpha}(A^2 - 4)^{-\frac{1}{4}}(F - F_c)^{-\frac{1}{2}}$ for the FK potential. The reciprocal of this time interval yields an approximation for the wave front velocity,

$$c(A, F) \sim -\frac{\sqrt{\alpha\beta\,(F - F_c)}}{\pi}\,, \tag{3.7}$$

or $c \sim -(A^2 - 4)^{\frac{1}{4}}(1 + 2/\sqrt{A^2 - F_c^2})^{\frac{1}{2}}(F - F_c)^{\frac{1}{2}}/(\pi\sqrt{2})$ for an FK potential. The minus sign reminds us that wave fronts move towards the left for $F > F_c$. In Figures 3.3(a) and (b) we compare this approximation with the numerically computed velocity for $A = 100$ and $A = 10$.

When the solution begins to blow up, the outer solution (3.6) is no longer a good approximation, for $u_0(t)$ departs from the stationary value $u_0(A, F_c)$. We must go back to (3.1) and obtain an inner approximation to this equation. As F is close to F_c and $u_0(t) - u_0(A, F_c)$ is of order 1, we numerically solve (3.1) at $F = F_c$ with the matching condition that $u_0(t) - u_0(A, F_c) \sim 2/[\pi\sqrt{\beta/[\alpha\,(F - F_c)]} - 2\beta\,(t - t_0)]$ as $(t - t_0) \to -\infty$. This inner solution describes the jump of u_0 from $u_0(A, F_c)$ to values on the largest stationary solution of (3.1), which is close to U_3. During this jump, the motion of u_0 forces the other points to move. Thus, $u_{-1}(t)$ can be calculated by using the inner solution in (1.1) for u_0, with $F = F_c$ and $u_{-2} \approx U_1$. A composite expansion [2] constructed with these inner and outer solutions is compared to the result of direct numerical simulations in Figure 3.4.

Notice that (3.3) is the normal form associated with a saddle-node bifurcation in a one-dimensional phase space. The wave front depinning transition is a *global* bifurcation with generic features: each individual point $u_n(t)$ spends a long time, which scales as $|F - F_c|^{-\frac{1}{2}}$, near discrete values $u_n(A, F_c)$, and then jumps to the next discrete value on a time scale of order 1. The traveling wave ceases to exist for $F \le F_c$.

3.2. Theory with several active points. The approximations to $F_c(A)$ and the wave front speed provided by the previous asymptotic theory break down for small

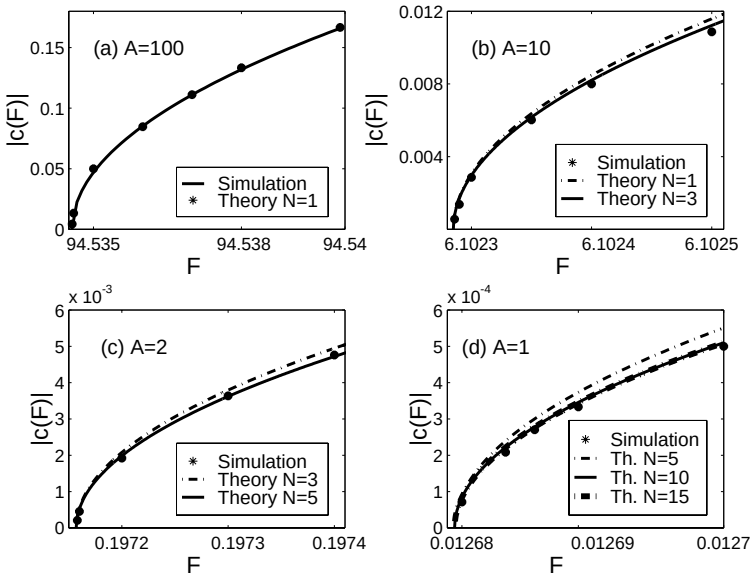


FIG. 3.3. *Comparison of theoretically predicted and numerically calculated wave front velocities near F_c for the FK model with N active points and the following values of the parameter A:* (a) $A = 100$, (b) $A = 10$, (c) $A = 2$, (d) $A = 1$.

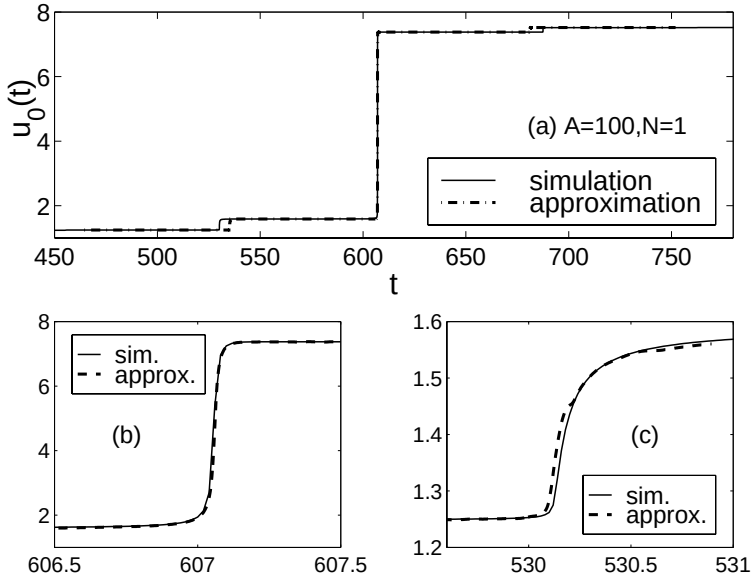


FIG. 3.4. *Comparison of asymptotic and numerically calculated wave front profiles near F_c:* (a) *Complete wave front profile as indicated by the trajectory $u_0(t)$.* (b) *Zoom near the largest jump in the profile.* (c) *Zoom near the jump preceding the largest one after translating the asymptotic profile. This last has been calculated by inserting the approximate $u_0(t)$ in the equation for $u_{-1}(t)$.*

A. In particular, for the FK potential and $A < 2$, no double zeroes of $2x + A\sin(x) - (F + U_1 + U_3)$ are found for $F = F_c$. What happens is that we need more than one point to approximate wave front motion. Depinning is then described by a reduced

system of more than one degree of freedom corresponding to active points. There is a saddle-node bifurcation in this reduced system whose normal form is of the same type as (3.3). The jump of the active points after blow up is found by solving the reduced system with a matching condition [10].

We explain our procedure for a finite number of active points. The front is formed by two tails, very close to $U_1(F/A)$ and $U_3(F/A)$, respectively, and several intermediate points $u_{-L}, \ldots, u_M$. The reduced system describing the dynamics of the front for $F > F_c$ is

$$\begin{aligned} &\frac{du_i}{dt} = u_{i+1} - 2u_i + u_{i-1} - Ag(u_i) + F, \quad i = -L, \ldots, M, \\ &u_{-L-1} = U_1(F/A), \qquad u_{M+1} = U_3(F/A). \end{aligned} \tag{3.8}$$

At $F = F_c$ this system has a stationary solution $u_{-L}, \ldots, u_M$. Writing $u_i(t) = u_i + v_i(t)$, we obtain for v_i

$$\begin{aligned} &\frac{dv_i}{dt} = v_{i+1} - 2v_i + v_{i-1} - Ag'(u_i)v_i - \frac{A}{2}g''(u_i)v_i^2 + F - F_c, \quad i = -L, \ldots, M, \\ &v_{-L-1} \sim \frac{F - F_c}{Ag'(U_1(F_c/A))}, \qquad v_{M+1} \sim \frac{F - F_c}{Ag'(U_3(F_c/A))}. \end{aligned} \tag{3.9}$$

The tridiagonal matrix $\mathcal{M}$ defined as

$$\begin{pmatrix} 2 + Ag'(u_{-L}) & -1 & & & & & \\ -1 & 2 + Ag'(u_{1-L}) & -1 & & & & \\ & \cdot & \cdot & \cdot & & & \\ & & \cdot & \cdot & \cdot & & \\ & & & \cdot & \cdot & \cdot & \\ & & & & -1 & 2 + Ag'(u_{M-1}) & -1 \\ & & & & & -1 & 2 + Ag'(u_M) \end{pmatrix} \tag{3.10}$$

has $L + M$ strictly positive eigenvalues plus a smallest eigenvalue $\lambda \sim 0$ with an associate positive eigenfunction V. We choose V such that $\sum_{i=-L}^{M} V_i^2 = 1$ and write an eigenfunction expansion for v as follows:

$$v(t) = \varphi\, V + \sum_{i=-L, i\neq 0}^{M} W_i \varphi_i \exp(-\lambda_i t). \tag{3.11}$$

Thus $v(t) \sim V\varphi(t)$ as time increases. Let $D = -\frac{A}{2}$ diag $(g''(u_{-L}), \ldots, g''(u_M))$, $V^2 = (V_{-L}^2, \ldots, V_M^2)$, and

$$w = (F - F_c)\left(1 + \frac{1}{Ag'(U_1(F_c/A))}, 1, \ldots, 1, 1 + \frac{1}{Ag'(U_3(F_c/A))}\right).$$

Then system (3.9) becomes

$$\frac{d\varphi}{dt}\, V \sim \varphi\, \mathcal{M}V + \varphi^2\, DV^2 + w \sim \varphi^2\, DV^2 + w \tag{3.12}$$

as time increases. Multiplying by the transpose of V, we get an evolution equation for the amplitude $\varphi(t)$:

$$\frac{d\varphi}{dt} = \alpha\,(F - F_c) + \beta\,\varphi^2, \tag{3.13}$$

where now α and β are

$$\alpha = \sum_{i=-L}^{M} V_i + \frac{V_{-L}}{Ag'(U_1(F_c/A))} + \frac{V_M}{Ag'(U_3(F_c/A))} > 0,$$

$$\beta = -\frac{A}{2} \sum_{i=-L}^{M} g''(u_i) V_i^3 > 0.$$

The coefficient α is positive because $g'(U_i) > 0$ for $i = 1, 3$. We have checked numerically that $\beta > 0$ for different nonlinearities and values of A. An intuitive explanation follows. First, notice that $g''(u) > 0$ for $u \in (U_2(0), U_3(0))$, and $g''(u) < 0$ for $u \in (U_1(0), U_2(0))$. For large A, the largest component is V_0, the others are negligible, and we have one active point as in the previous subsection; see Figure 3.4. Then $\beta \sim -g''(u_0(A, F_c))V_0^3 > 0$ because $u_0 < U_2(0)$, which implies $g''(u_0) < 0$. As A decreases, V_0 is still the largest component and $g''(u_0(A, F_c)) < 0$. Now there may be other terms with $g''(u_i(A, F_c)) > 0$, and we have only numerical evidence that $\beta > 0$, not a proof.

Notice that (3.13) is the normal form of a saddle-node bifurcation. Its solution is again (3.6), which blows up at times $(t - t_0) = \pm 1/(2c)$, where

$$c(A, F) \sim -\frac{1}{\pi} \sqrt{\alpha\beta(F - F_c)}, \tag{3.14}$$

as discussed before. c is the wave front speed near F_c, approximately given by the reciprocal of the time during which the outer solution holds.

Figures 2.1, 3.3, and 3.5 show the critical field, wave front velocities, and profiles for different values of $A \in (1, 10)$ corresponding to the FK model. We have compared results of direct numerical simulations to those of our theory for $N = L+M+1$ active points. Provided that $N = L + M + 1$ active points have been selected, we find the smallest eigenvalue of the matrix $\mathcal{M}$ and move F until $\lambda(F, A; N) = 0$, $N = L+M+1$, which yields an approximation for $F_c(A)$; see Figure 2.1. The wave front velocities can be calculated by means of (3.14) and have been depicted in Figure 3.3.

The wave front profiles near F_c can be determined as follows. We start with an initial condition, $u_n(0) \approx u_n(A, F_c)$ or $\varphi(0) = 0$ in (3.11). The active points blow up at $t \sim \pm(2c)^{-1}$, for example as

$$u_n(t) \sim u_n(A, F_c) + \frac{1}{\beta(\pm\frac{1}{2c} - t)} V, \tag{3.15}$$

provided $t \to \pm 1/(2c)$. At these times, we should insert a fast stage during which the $u_n(t)$ are no longer close to $u_n(A, F_c)$, as an inner layer. The inner layer variables $u_n(t)$ obey (1.1) with $F = F_c$ and the boundary conditions $u_n(t) \to u_n(A, F_c)$ (according to (3.15)) as $t \to -\infty$, and $u_n(t) \to u_{n+1}(A, F_c)$ as $t \to \infty$. To get a uniform approximation, we notice that the blow up times are $t_m = (2c)^{-1} + m/c$, $m \in \mathbb{Z}$. Let us denote by $u_n^{(m)}(\tau)$, $\tau = (t - t_m)$, the solution of (1.1) with $F = F_c$ and the boundary conditions $u_n^{(m)}(\tau) \to u_{n+m}(A, F_c)$ as $\tau \to -\infty$, and $u_n^{(m)}(\tau) \to u_{n+m+1}(A, F_c)$ as $\tau \to \infty$. During the time interval $(t_{-L-n-1}, t_{M-n}) = (-(2c)^{-1} - (L+n)/c, (2c)^{-1} + (M - n)/c)$ that $u_n(t)$ needs to go from $U_1(F_c/A)$ to $U_3(F_c/A)$, the uniform approximation

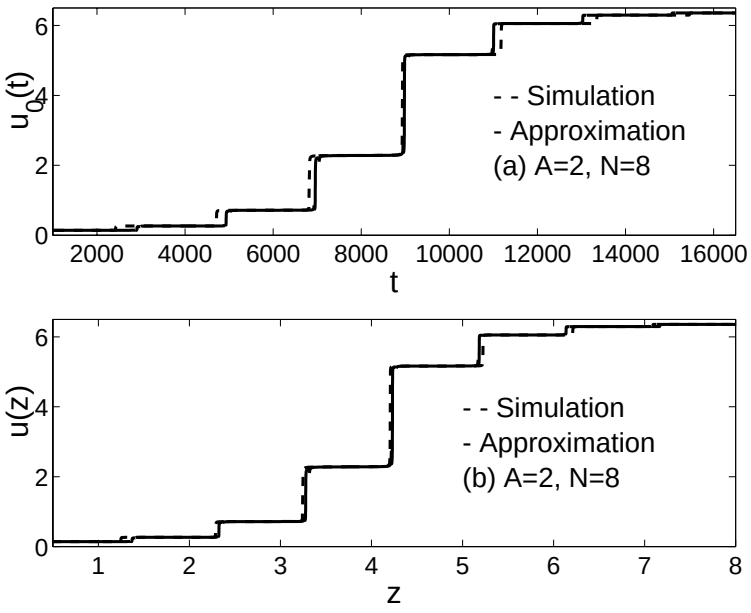


FIG. 3.5. *Comparison of theoretically predicted and numerically calculated wave fronts near F_c for $A = 2$ using $N = 8$ active points:* (a) *trajectory of one point,* (b) *wave front profile,* $u(z) = u_0(z/|c|)$.

to the wave front is

$$u_n(t) \sim \sum_{m=-n-L-1}^{M-n} \Big\{ u_n^{(m)}(t-t_m) + u_n^{(m-1)}(t-t_{m-1}) - u_{n+m}(A,F_c) + \left[\varphi\left(t - \frac{m}{c}\right) - \frac{1}{\beta\,(t_m - t)} + \frac{1}{\beta\,(t-t_{m-1})} \right] V \Big\} \chi_{(t_{m-1},t_m)}. \tag{3.16}$$

Then $u_n(t_{-L-n-1}) \sim U_1(F_c/A)$ and $u_n(t_{M-n}) \sim U_3(F_c/A)$. In (3.16), the indicator function $\chi_{(t_{m-1},t_m)}$ is 1 if $t_{m-1} < t < t_m$ and 0 otherwise. Therefore, $\chi_{(t_{m-1},t_m)} = \theta(t_m - t) - \theta(t_{m-1} - t)$, where $\theta(x) = 1$ if $x > 0$ and 0 otherwise. Written in terms of the variable $z = n - ct$ such that $u_n(t) = u(z)$, $u(z) = u_n((n-z)/c) = u_0(-z/c)$. Then (3.16) becomes

$$\begin{aligned} u(z) \sim \sum_{m=-L-1}^{M} & \Big\{ u_0^{(m)}\left(-\frac{z+m+\frac{1}{2}}{c}\right) + u_0^{(m-1)}\left(-\frac{z+m-\frac{1}{2}}{c}\right) - u_m(A,F_c) \\ & + \left[\varphi\left(-\frac{z+m}{c}\right) - \frac{c}{\beta\,(z+m+\frac{1}{2})} - \frac{c}{\beta\,(z+m-\frac{1}{2})} \right] V \Big\} \\ & \times \left[\theta\left(z+m+\frac{1}{2}\right) - \theta\left(z+m-\frac{1}{2}\right) \right] \end{aligned} \tag{3.17}$$

for $-M - 1/2 < z < L + 1/2$. We have $u(L + 1/2 + 0) \sim U_1(F_c/A)$ and $u(-M - 1/2 - 0) \sim U_3(F_c/A)$, and therefore (3.17) approximates the wave front profile. In Figure 3.5, we have depicted the wave front profile in two ways, by drawing $u_0(t)$ and $u(z) = u_0(-z/c)$. Notice that the largest source of discrepancy between numerical

calculations and our asymptotic approximation is the error in determining the wave speed. The discrepancies are more evident for $u_0(t)$ because of the different horizontal scale used to depict $u(z)$.

How do we determine the optimal number of active points? For large enough $N = L + M + 1$ and a given A, the eigenvector V corresponding to the smallest eigenvalue of the matrix $\mathcal{M}$ in (3.10) has a certain number of components that are of order one, whereas all others are very small. The number of components of normal size determines the optimal number of active points: only one point if A is larger than 10, five if $A = 2$, etc. Keeping less active points than the optimal number results in larger errors, whereas keeping more active points than optimal does not result in a significantly better approximation. The eigenvector of the reduced system of equations for the active points is a good approximation to the large components of the eigenvector corresponding to the complete system. As we approach the continuum limit, more and more points enter the reduced system of equations, and exponential asymptotic methods become a viable alternative to our methods.

3.3. Depinning transition as a global bifurcation. We have shown that the depinning transition is a global bifurcation in a reduced system of equations corresponding to the active points. Starting from a stable stationary solution, the smallest eigenvalue of the system linearized about the stationary solution becomes zero at the (approximate) critical field, and its associated eigenfunction is positive. The stationary solution disappears as the critical field is surpassed. Beyond it, the active points $u_n(t)$ spend a long time, of order $(F - F_c)^{-\frac{1}{2}}$, near the stationary values $u_n(A, F_c)$, and then jump to $u_{n+1}(A, F_c)$ on an order 1 time scale. Near the critical field, the depinning transition is described *locally* by the normal form of a saddle-node bifurcation. For $F > F_c$ (or $F < -F_c$), the bifurcation amplitude blows up in finite time, on a time scale of order $||F| - F_c|^{-\frac{1}{2}}$. The construction of the wave front profile is completed, matching the outer solution given by the saddle-node normal form to a solution of the reduced system of active points at $F = F_c$. (A mathematically related phenomenon occurs in a mean-field model of sliding charge-density waves [2].)

We conjecture that the depinning transition in the infinite system (1.1) is a global bifurcation of the same type as for the reduced system of active points. At the critical field, two stationary solutions of (1.1) (one stable, the other unstable) coalesce and disappear. For $F > F_c$ (or $F < -F_c$), the wave front profile is constructed as indicated above for the reduced system. To prove this conjecture, we could repeat our construction in section 3.2 for an infinite number of points. This is possible because we know that the infinite system, linearized about the "stable" steady solution $u_i(A, F_c)$ at $F = F_c$, has a zero eigenvalue and an associated positive exponentially decaying eigenfunction V. Using V, we obtain the normal form equation (3.13), where now

$$\alpha = \sum_{i=-\infty}^{\infty} V_i > 0, \qquad \beta = -\frac{A}{2} \sum_{i=-\infty}^{\infty} g''(u_i) V_i^3. \tag{3.18}$$

We should now prove that the coefficient β is positive and that the infinite system has solutions connecting $u_n(A, F_c)$ to $u_{n+1}(A, F_c)$ and satisfying the matching condition. We justified that $\beta > 0$ for the finite system in subsection 3.2, and we show in Proposition A.4 (in the appendix) that the eigenfunction for the infinite system can be approximated by the corresponding eigenfunction of the reduced system with a finite number of active points. The existence of traveling wave solutions for $F > F_c$ ensures that the infinite system has solutions connecting $u_n(A, F_c)$ to $u_{n+1}(A, F_c)$.

The velocity of a wave front in the infinite system is again given by (3.14) with the coefficients (3.18).

Remark 5. By using comparison techniques, it is possible to prove that solutions of discrete RD equations with finitely many points and Dirichlet boundary conditions approximate solutions of the same equations with infinitely many points. In the continuum limit, the wave fronts approach constant values exponentially fast as $i \to \pm\infty$. This exponential decay justifies the active point approximation in two ways. First, the number of active points needed to approximate well the wave fronts of the infinite system decreases as A increases. It is usually better to add another active point to the approximate system than to patch rigid tails to the last active points of a wave front by generalizing Kladko, Mitkov, and Bishop's active site approximation [28]. Secondly, exponential decay at the ends of a wave front causes the operator of the linearized problem about the wave front to be compact and therefore to have a discrete spectrum (see the appendix). This fact justifies that the normal form we calculate by using active points approximates the correct local normal form of the depinning global bifurcation.

4. Conclusions. In this paper, we have studied depinning of wave fronts in discrete RD equations. Pinned (stationary) and traveling wave fronts cannot coexist for the same value of the forcing term. There are two different depinning transitions, i.e., two different ways in which a pinned front may start moving. The normal depinning transition can be viewed as a loss of continuity of traveling front profiles as the critical field is approached: below the critical field, the fronts become pinned stationary profiles with discontinuous jumps at discrete values u_n. The wave front velocity scales as $|F - F_c|^{\frac{1}{2}}$ near the critical field F_c. For sufficiently large A (far from the continuum limit), the critical field and these fronts can be approximated by singular perturbation methods which show excellent agreement with numerical simulations. These methods are based upon the fact that the wave front motion can be described by a reduced system of equations corresponding to the dynamics of only a finite number of points, the active points.

Besides the normal depinning transition, certain nonlinearities present anomalous pinning (pinning failure): the velocity of the wave fronts is not zero except at zero forcing, just as for continuous RD equations. These nonlinearities are characterized by smooth profiles of stationary and moving wave fronts, by having zero critical field, and by a linear scaling of wave front velocity with field.

Appendix. Characterization of the depinning threshold. In this section we establish the "depinning criterion," which provides a characterization of $F_c(A)$ as follows.

THEOREM A.1. *Set $F = 0$ and $A > 0$. Assume that the nonlinearity $g \in C^3$ has three zeroes U_i, $U_1 < U_2 < U_3$, is odd about U_2, and satisfies $g'(U_1) = g'(U_3) > 0$. Let u_n be a stationary increasing solution of* (1.1)*, symmetric about U_2 and such that $u_{-\infty} = U_1$ and $u_\infty = U_3$. Let $\lambda_1(A, 0)$ be the smallest eigenvalue of the zero field operator $\mathcal{L}_0$ of* (2.8) *at $F = 0$:*

$$\begin{aligned} -(v_{n+1} - 2v_n + v_{n-1}) + Ag'(u_n)v_n &= \lambda_1(A,0)v_n, \\ v_{\pm n} \to 0 \text{ exponentially as } n \to \infty. \end{aligned} \tag{A.1}$$

If $\lambda_1(A, 0) > 0$, then $F_c(A) > 0$, and for $|F| \le F_c(A)$ there exist increasing stationary solutions $u_n(A, F)$ of (1.1) *with $u_{-\infty} = U_1(F/A)$ and $u_\infty = U_3(F/A)$. Moreover, the smallest eigenvalues of the operator $\mathcal{L}(F)$, corresponding to the linearization of* (1.1)

about $u_n(A,F)$,

$$\begin{aligned} -(v_{n+1}-2v_n+v_{n-1})+Ag'(u_n(A,F))v_n &= \lambda_1(A,F)v_n, \\ v_{\pm n} &\to 0 \text{ exponentially as } n\to\infty, \end{aligned} \tag{A.2}$$

are strictly positive for $|F|<F_c(A)$. *We can characterize* $F_c(A)$ *as the zero of the smallest eigenvalue,* $\lambda_1(A,F_c(A))=0$.

This theorem will be proved in subsection A.2. To calculate $\lambda_1(A,F)$ and $F_c(A)$, we approximate the infinite tridiagonal matrix in (A.2) by an $N\times N$ matrix, where N is the number of active points. Similar truncation approximations were used in [29] to calculate the lowest eigenvalue of an infinite tridiagonal matrix. For values of A which are not too small, numerical simulations show that the matrices $\mathcal{M}$ in (3.10) have positive eigenvalues. The eigenvector $V(A,F,N)$ (chosen to have norm 1), corresponding to the smallest eigenvalue $\lambda(A,F,N)$, is positive, and it is "concentrated" in the central components $V_{-m(A)},\ldots,V_{m(A)}$. All other components are very small. The number of significant components $m(A)$ does not change as N increases, but it increases as A decreases. For large A, $m(A)=0$, and only V_0 is significant. Provided N is large enough, the eigenvalues $\lambda(A,F,N)$ and the eigenvectors $V(A,F,N)$ approximate well the smallest eigenvalue and associated eigenfunction of the infinite problem, as we indicate in the next subsection. For a fixed value of N, the eigenvalues $\lambda(A,F,N)$ decrease as A decreases. For fixed N and A, they decrease as F increases from $F=0$ to values close to $F_c(A)$. In the next subsection, we collect several results on eigenvalues for this type of problem.

A.1. Eigenvalue problems. Before proving Theorem A.1, we should make sure that our linear operators do have eigenfunctions and eigenvalues. We consider the real valued and symmetric operators $\mathcal{L}(F)\,v_n = Ag'(u_n)v_n-(v_{n+1}-2v_n+v_{n-1})$ in spaces of sequences decaying exponentially at infinity. Their spectra are discrete and real (these operators are compact), and we would like to make sure that they are not empty. Since we are interested mainly in the smallest eigenvalue, we shall use its variational characterization, prove that this eigenvalue exists, and characterize its dependence on the parameters A and F. We shall also describe finite-dimensional approximations of eigenvalues and eigenfunctions.

Let us first look for necessary conditions for $\lambda(A,F)$ to exist. Let $\lambda\in\mathbb{R}$ be an eigenvalue of $\mathcal{L}(F)$ with eigenfunction V_n. Multiplying $\mathcal{L}(F)\,V_n-\lambda V_n$ by V_n and summing over n, we obtain

$$\begin{aligned} 0 &= \sum_n (V_{n+1}-V_n)^2+[Ag'(u_n)-\lambda]V_n^2 \\ &\geq \sum_n (V_{n+1}-V_n)^2+([A\min_n g'(u_n)]-\lambda)\sum_n V_n^2. \end{aligned} \tag{A.3}$$

Thus, $\lambda=\sum_n[(V_{n+1}-V_n)^2+Ag'(u_n)V_n^2]/\sum V_n^2 > A\min_n g'(u_n)$. This inequality implies that λ is positive if u_n does not take on values in the region where g' is negative, which occurs for large enough $A>0$. In general, we can say only that $\lambda>Ag'(U_2)$ for g' attains its minimum value in $[U_1,U_3]$ at U_2. Similarly,

$$\begin{aligned} 0 &= \sum_n (V_{n+1}-V_n)^2+[Ag'(u_n)-\lambda]V_n^2 \\ &\leq [4+A\max_n(g'(u_n))-\lambda]\sum_n V_n^2. \end{aligned} \tag{A.4}$$

Therefore, $\lambda < Ag'(U_1) + 4 = Ag'(U_3) + 4$ for $g'(u)$ attains its maximum value in $[U_1, U_3]$ at the end points, $u = U_1$ and $u = U_3$.

The smallest eigenvalue λ is given by the Rayleigh formula

$$\lambda = \min \frac{\sum_n (w_{n+1} - w_n)^2 + Ag'(u_n) w_n^2}{\sum_n w_n^2}, \tag{A.5}$$

where the infimum is taken over our space of exponentially decaying functions. That minimum is attained at an eigenfunction V_n solving (A.1). Now, (A.1) may have solutions decaying at $\pm\infty$ only if the difference equation

$$V_{n+1} + (-2 - Ag'(U_1) + \lambda) V_n + V_{n-1} = 0 \tag{A.6}$$

has solutions of the form r^n with $r < 1$. This happens when $(-2 - Ag'(U_1) + \lambda)^2 > 4$. Thus either $\lambda > 4 + Ag'(U_1) > 0$ (excluded above) or $\lambda < Ag'(U_1)$. We conclude that $\lambda < Ag'(U_1)$ is a necessary condition to attain the minimum (A.5) at a positive eigenfunction decaying exponentially at infinity.

We now establish sufficient conditions for the minimum (A.5) to exist.

LEMMA A.2 (Conditions for the existence of positive decaying eigenfunctions). *Let $F = 0$, $A > 0$, and let the nonlinearity g satisfy the hypotheses in Theorem* A.1. *Let u_n be a stationary increasing solution of* (1.1) *such that $u_{-\infty} = U_1(0)$ and $u_\infty = U_3(0)$. Given an exponentially decaying sequence $w = w_n$, we define*

$$J(w) = \frac{\sum_n [(w_{n+1} - w_n)^2 + Ag'(u_n) w_n^2]}{\sum_n w_n^2}.$$

Let us suppose that there is a sequence w_n such that $J(w_n) < Ag'(U_1)$. Then the infimum

$$\lambda = \inf_{\sum_n r_0^{-2|n|} v_n^2 < \infty} \frac{\sum_n [(v_{n+1} - v_n)^2 + Ag'(u_n) v_n^2]}{\sum_n v_n^2} \tag{A.7}$$

is attained at a positive function V_n which decays as $r(A,\lambda)^{|n|}$ at infinity, with $0 < r(A,\lambda) = [2 + Ag'(U_1) - \lambda - \sqrt{(-2 - Ag'(U_1) + \lambda)^2 - 4}]/2 < r_0 < 1$. Now $\lambda = \lambda_1(A, 0)$, and v_n solves (A.1).

Remark 6. The value $0 < r_0 < 1$ is determined in the proof. Note that $r(A,\lambda)$ is a decreasing function of A but an increasing function of λ.

Remark 7. We have shown above that $Ag'(U_2) < \lambda < Ag'(U_1)$. Thus the smallest eigenvalue shrinks to zero as $A \to 0$, although we do not have proof that it does so monotonically.

Proof. Clearly $J(w)$ is bounded from below by $A \min_n(g'(u_n))$. We choose $r_0 = r(A, J(w_n)) \in (0,1)$ and define $\|w\|_0 = \sum r_0^{-2|n|} |w_n|^2$. Let $w^m = w_n^m$, $m > 0$, be a sequence minimizing $J(w)$: $\|w^m\|_0 < \infty$ and let $J(w^m) \to \lambda$ when $m \to \infty$. We replace w^m with $v^m = v_n^m = w_n^m / \|w^m\|_0$. Then, $\|v^m\|_0 = 1$ and $J(v^m) = J(w^m) \to \lambda$. $\|v^m\|_0 = 1$ implies that, uniformly in m, $|v_n^m| \le r_0^{|n|}$ and $\sum_{n > n(\epsilon)} |v_n^m|^2 < \epsilon$ for $n(\epsilon)$ large enough. Thus, a subsequence v^m tends to some limit $V = V_n$ such that $|V_n| \le r_0^{|n|}$ and $\sum_n |v_n^m - V_n|^2$ tends to zero as m tends to infinity. Therefore, $J(v^m) \longrightarrow J(V) = \lambda$ and the infimum is attained at the sequence $V = V_n$. Moreover, $V = V_n$ satisfies the Euler equation (A.1) for the minimization problem, which then implies that V_n decays as stated in the lemma.

On the other hand, $J(|V_n|) \le J(V_n)$, and we can choose nonnegative V_n. But then λ has to be the smallest eigenvalue $\lambda_1(A,0)$.

LEMMA A.3 (Choice of the sequence w_n with $J(w_n) < Ag'(U_1)$). *Let $F = 0$, $A > 0$, and u_n be as in Lemma* A.2. *Let $u(x)$ be the solution of the boundary value problem $d^2u/dx^2 = g(u)$, with $u(-\infty) = U_1$, $u(\infty) = U_3$ such that $u(0) = U_2$.*

- *For sufficiently small $A < 1$, we have*

$$|u_{n+1} - u_n| \max_{[u_n, u_{n+1}]} |g''| < 2g'(U_1) \quad \forall n, \tag{A.8}$$

and $w_n = u_{n+1} - u_n$ satisfies $J(w_n) < Ag'(U_1)$. An estimation of the appropriate values of A indicates that they should be smaller than

$$A < \left(\frac{2g'(U_1)}{\max_{[U_1,U_3]}|g''| \max_{\mathbb{R}}|du/dx|} \right)^2 .$$

- *For $A > 2\,g'''(U_1)$, we can choose $w_n = 0$ for $|n| > M \ge 0$, $w_n = r^n$ for $n \ge 0$, and $w_n = r^{n+1}$ for $n < 0$.*

Proof. The function $w_n = u_{n+1} - u_n$ is a solution of

$$w_{n+1} - 2w_n + w_{n-1} = A \frac{g(u_{n+1}) - g(u_n)}{u_{n+1} - u_n} w_n$$

that decays at infinity as $r(A,0)^{|n|}$. Multiplying this equation by w_n and adding over n, we obtain

$$\sum_n \left((w_{n+1} - w_n)^2 + A \frac{g(u_{n+1}) - g(u_n)}{u_{n+1} - u_n} w_n^2 \right) = 0. \tag{A.9}$$

This result can be used to calculate $J(w_n)$:

$$J(w_n) = A \frac{\sum \left(g'(u_n) - \frac{g(u_{n+1}) - g(u_n)}{u_{n+1} - u_n} \right) w_n^2}{\sum w_n^2} = -\frac{A}{2} \frac{\sum g''(\xi_n) w_n^3}{\sum w_n^2} , \tag{A.10}$$

by the mean value theorem. Thus, $J(w_n) \le (A/2) \max_n (|u_{n+1} - u_n| \max_{[u_n, u_{n+1}]} |g''|)$. For sufficiently small A, $|u_{n+1} - u_n| \le C\sqrt{A}$, so that $J(w_n) < CA^{\frac{3}{2}} \max|g''|/2 < Ag'(U_1)$. More precisely, for small A, $w_n = u_{n+1} - u_n \simeq u((n+1)\sqrt{A}) - u(n\sqrt{A}) \simeq \sqrt{A} u'(\xi)$. Then $|w_n| \le \max |du/dx| \sqrt{A}$.

To prove the other case, we observe that $J(w_n) = (r-1)^2 + A \sum_{i=0}^{\infty} g'(u_n) r^{2n}$ is smaller than $Ag'(u_1)$, provided that $(1-r)/(1+r) < A\,g'''(U_1)/[2\,(1 - (r(A,0)r)^2)]$, with $r(A,\lambda)$ defined as in Lemma A.2. The last inequality holds if $A > 2\,g'''(U_1)$.

Remark 8. For the FK nonlinearity, the first condition of the lemma holds for $A < 0.9$, and the second condition for $A > 2$. For intermediate values, numerical simulations show that $w_n = u_{n+1} - u_n$ satisfies $J(w_n) < Ag'(U_1)$.

PROPOSITION A.4 (Finite-dimensional approximations). *Let $u_n(A,F)$ be a stationary solution of* (1.1) *under the hypotheses in Theorem* A.1 *for $|F| \le F_c(A)$. Let $\lambda_1(A,F)$ be the smallest eigenvalue of the operator $\mathcal{L}(F)$ (linearized about $u_n(A,F)$) and $\lambda(A,F,N)$ be the smallest eigenvalues of the matrices* (3.10). *Then, $\lambda(A,F,N) \to \lambda_1(A,F)$ as $N \to \infty$. As a consequence, if $V > 0$ is an eigenfunction associated to $\lambda_1(A,F)$ with $\sum_n V_n^2 = 1$ and if $V(N) > 0$ are eigenvectors associated to $\lambda(A,F,N)$ such that $\sum_n V_n(N)^2 = 1$, then $V(N) \to V$ as $N \to \infty$.*

Proof. It follows from the Rayleigh characterizations for the smallest eigenvalues:

$$\lambda_1(A,F) = \min_{\sum r^{-2|n|} w_n^2} \frac{\sum_{-\infty}^{\infty} [(w_{n+1}-w_n)^2 + Ag'(u_n(A,F))w_n^2]}{\sum_{-\infty}^{\infty} w_n^2}, \tag{A.11}$$

$$\lambda(A,F,N) = \min \frac{\sum_{-L}^{M} [(v_{n+1}-v_n)^2 + Ag'(u_n(A,F))v_n^2]}{\sum_{-L}^{M} v_n^2}. \tag{A.12}$$

Letting $w_n = v_n$ for $n = -L, \ldots, M$ and $w_n = 0$ otherwise, we see that $\lambda_1(A,F) \leq \lambda(A,F,N)$, $N = L + M + 1$. Now let w_n be an eigenfunction for $\lambda_1(A,F)$ such that $\sum_{-\infty}^{\infty} w_n^2 = 1$. Then,

$$\lambda(A,F,N) \leq \frac{\lambda_1(A,F) - \sum_{n<-L,\, n>M} [(w_{n+1}-w_n)^2 + Ag'(u_n(A,F))w_n^2]}{\sum_{-L}^{M} w_n^2}.$$

We conclude that $\lambda(A,F,N) \to \lambda_1(A,F)$ as $N \to \infty$. This and the exponential decay of V prove the convergence of the eigenvectors.

A.2. Proof of Theorem A.1. The theorem will be proved in two steps and, for simplicity, in the particular case of periodic g. In this case, $U_3(F/A) - U_3(0) = U_1(F/A) - U_1(0)$, which allows us to use symmetric sub- and supersolutions. Small modifications are required in the general case.

Step 1: $F_c(A) > 0$. We use the existence of a positive eigenfunction v_n associated with a positive eigenvalue $\lambda_1(A,0)$ to construct stationary supersolutions for (1.1) when $F > 0$ is small. The known solution u_n provides a stationary subsolution.

We look for a supersolution of the form

$$w_n = u_n + (1+\delta)\left(U_1\left(\frac{F}{A}\right) - U_1(0)\right) + \epsilon v_n, \tag{A.13}$$

with $\delta > 0$ to be chosen and ϵ, F small to be determined. Let us check that

$$w_{n+1} - 2w_n + w_{n-1} \leq g(w_n) - F, \qquad w_\infty > U_3(F/A), w_{-\infty} > U_1(F/A) \tag{A.14}$$

holds. The conditions at infinity are satisfied for any $\delta > 0$. Provided

$$\left|(1+\delta)\left(U_1\left(\frac{F}{A}\right) - U_1(0)\right)\right| \leq k\epsilon, \tag{A.15}$$

inequality (A.14) holds if

$$\epsilon\,(v_{n+1} - 2v_n + v_{n-1}) < Ag'(u_n)\left[\epsilon v_n + (1+\delta)\left(U_1\left(\frac{F}{A}\right) - U_1(0)\right)\right] - F + O(A\epsilon^2).$$

Using (A.1), we are left with

$$F < \epsilon\lambda_1(A,0)v_n + A(1+\delta)g'(u_n)\left(U_1\left(\frac{F}{A}\right) - U_1(0)\right).$$

Now, $U_1(F/A) = g^{-1}(F/A)$, the inverse being taken near $U_1(0)$, in the region with $g' > 0$. Using $g^{-1}(x) \sim g^{-1}(x_0) + (g^{-1})'(x_0)(x - x_0)$, we obtain

$$U_1\left(\frac{F}{A}\right) = g^{-1}\left(\frac{F}{A}\right) \sim U_1(A,0) + \frac{F}{A\,g'(U_1(0))}. \tag{A.16}$$

Thus, the condition for w_n to be a supersolution is

$$F < \epsilon\lambda_1(A,0)v_n + g'(u_n)\frac{1+\delta}{g'(U_1(0))}F.$$

Let M be sufficiently large. We distinguish two different ranges of indices n:

- For $|n| > M$, $g'(u_n) > 0$ and $v_n \ll 1$. Then the right-hand side of the previous inequality is dominated by the second term. We choose δ large enough to ensure $F < g'(u_n)(1+\delta)\,F/g'(U_1(0))$, that is, $1+\delta > g'(U_1(0))/g'(u_M)$.
- For small $|n|$, $g'(u_n) < 0$. The previous inequality is satisfied, provided we choose F so small that

$$\left(1 + |g'(u_n)|\frac{1+\delta}{g'(U_1(0))}\right) F < \epsilon\lambda_1(A,0)v_n$$

  for a fixed value of δ.

With these choices, w_n satisfies (A.14). Note that these choices are compatible with condition (A.15). Using (A.16), (A.15) becomes $(1+\delta)\,F/[A\,g'(U_1(0))] < k\,\epsilon$. This holds for small enough F.

Let $F > 0$ be small enough for a w_n defined in (A.13) to be a supersolution with δ, ϵ adequately selected. Now, let $h_n(t)$ be a solution to (1.1) for such $F > 0$ with initial datum $h_n(0)$ satisfying $u_n < h_n(0) < w_n$. Then, $u_n < h_n(t) < w_n$ for all $t > 0$. Therefore, propagation is excluded and the solutions are pinned.

Stationary solutions $u_n(A,F)$ for such $F > 0$ can be obtained as long time limits of solutions $h_n(t)$ to (1.1) when $h_n(0)$ is increasing, tends exponentially to $U_1(A,F)$ (resp., $U_3(A,F)$) at $-\infty$ (resp., ∞), and $u_n < h_n(0) < w_n$. We conclude that $F_c(A) > 0$.

Step 2: $\lambda_1(A,F) > 0$ for $|F| < F_c(A)$ and $\lambda_1(A,F_c(A)) = 0$. To fix ideas, we take $F > 0$. The case $F < 0$ follows by symmetry. From Step 1, we know that $F_c(A) > 0$, and there are stationary solutions $u_n(A,F)$ of (1.1) existing for $F > 0$ small that are increasing from $U_1(F/A)$ to $U_3(F/A)$.

In an analogous way as we did for $F = 0$, we get

$$\lambda_1(A,F) = \min_{\sum r_0^{-2|n|} w_n^2 < \infty} \frac{\sum_n [(w_{n+1} - w_n)^2 + A g'(u_n(A,F)) w_n^2]}{\sum w_n^2}\,. \tag{A.17}$$

This formula defines $\lambda_1(A,F)$ as a continuous function of F. That $\lambda_1(A,0) > 0$ implies $\lambda_1(A,F) > 0$ up to some F_c at which $\lambda_1(A,F_c) = 0$. As long as $\lambda_1(A,F_1) > 0$, we can obtain stationary solutions for $F > F_1$ (close to F_1), as done in Step 1. This procedure cannot continue forever since such stationary solutions do not exist for F close to A: eventually $g(U) = F/A$ ceases to have three solutions, and the stationary wave fronts cannot be constructed. Thus, we must reach a value F_c at which $\lambda_1(A,F_c) = 0$.

Acknowledgments. The authors are indebted to J. M. Vega for a critical reading of the manuscript and helpful comments, and to V. Hakim for pointing out to them the relevance of [7]. A. C. thanks S. P. Hastings and J. B. McLeod for fruitful discussions.

REFERENCES

[1] A. Amann, A. Wacker, L. L. Bonilla, and E. Schöll, *Dynamic scenarios of multistable switching in semiconductor superlattices*, Phys. Rev. E, 63 (2001), paper 066207.

[2] L. L. Bonilla, *Stable probability densities and phase transitions for mean-field models in the thermodynamic limit*, J. Statist. Phys., 46 (1987), pp. 659–678.
[3] L. L. Bonilla, J. Galán, J. A. Cuesta, F. C. Martínez, and J. M. Molera, *Dynamics of electric field domains and oscillations of the photocurrent in a simple superlattice model,* Phys. Rev. B, 50 (1994), pp. 8644–8657.
[4] O. M. Braun and Yu. S. Kivshar, *Nonlinear dynamics of the Frenkel-Kontorova model,* Phys. Rep., 306 (1998), pp. 1–108.
[5] P. C. Bressloff and G. Rowlands, *Exact travelling wave solutions of an "integrable" discrete reaction-diffusion equation,* Phys. D, 106 (1997), pp. 255–269.
[6] A. E. Bugrim, A. M. Zhabotinsky, and I. R. Epstein, *Calcium waves in a model with a random spatially discrete distribution of Ca^{2+} release sites*, Biophys. J., 73 (1997), pp. 2897–2906.
[7] J. W. Cahn, *Theory of crystal growth and interface motion in crystalline materials*, Acta Metallurgica, 8 (1960), pp. 554–562.
[8] A. Carpio, S. J. Chapman, S. Hastings, and J. B. McLeod, *Wave solutions for a discrete reaction-diffusion equation*, European J. Appl. Math., 11 (2000), pp. 399–412.
[9] A. Carpio, L. L. Bonilla, A. Wacker, and E. Schöll, *Wave fronts may move upstream in doped semiconductor superlattices*, Phys. Rev. E, 61 (2000), pp. 4866–4876.
[10] A. Carpio and L. L. Bonilla, *Wave front depinning transition in discrete one-dimensional reaction-diffusion systems*, Phys. Rev. Lett., 86 (2001), pp. 6034–6037.
[11] A. Carpio, L. L. Bonilla, and G. Dell'Acqua, *Motion of wave fronts in semiconductor superlattices*, Phys. Rev. E, 64 (2001), paper 036204.
[12] A. Carpio, L. L. Bonilla, and A. Luzón, *Effects of disorder on the wave front depinning transition in spatially discrete systems*, Phys. Rev. E, 65 (2002), paper 035207(R)
[13] P. M. Chaikin and T. C. Lubensky, *Principles of Condensed Matter Physics,* Cambridge University Press, Cambridge, UK, 1995.
[14] T. Erneux and G. Nicolis, *Propagating waves in discrete reaction-diffusion systems*, Phys. D, 67 (1993), pp. 237–244.
[15] G. Fáth, *Propagation failure of traveling waves in a discrete bistable medium*, Phys. D, 116 (1998), pp. 176–190.
[16] S. Flach, Y. Zolotaryuk, and K. Kladko, *Moving lattice kinks and pulses: An inverse method*, Phys. Rev. E, 59 (1999), pp. 6105–6115.
[17] J. Frenkel and T. Kontorova, *On the theory of plastic deformation and twinning*, J. Phys. USSR, 13 (1938), pp. 1–10.
[18] E. Gerde and M. Marder, *Friction and fracture*, Nature, 413 (2001), pp. 285–288.
[19] G. Grüner, *The dynamics of charge-density waves*, Rev. Modern Phys., 60 (1988), pp. 1129–1181.
[20] V. Hakim and K. Mallick, *Exponentially small splitting of separatrices, matching in the complex plane and Borel summation*, Nonlinearity, 6 (1993), pp. 57–70.
[21] R. Hobart, *Peierls-barrier minima*, J. Appl. Phys., 36 (1965), pp. 1948–1952.
[22] V. L. Indenbom, *Mobility of dislocations in the Frenkel-Kontorova model*, Soviet Phys.-Crystallogr., 3 (1959), pp. 193–201 (translated from Kristallografiya, 3 (1958), pp. 197–206).
[23] J. P. Keener, *Propagation and its failure in coupled systems of discrete excitable cells*, SIAM J. Appl. Math., 47 (1987), pp. 556–572.
[24] J. P. Keener and J. Sneyd, *Mathematical Physiology,* Springer, New York, 1998.
[25] P. G. Kevrekidis, C. K. R. T. Jones, and T. Kapitula, *Exponentially small splitting of heteroclinic orbits: From the rapidly forced pendulum to discrete solitons*, Phys. Lett. A, 269 (2000), pp. 120–129.
[26] P. G. Kevrekidis, I. G. Kevrekidis, and A. R. Bishop, *Propagation failure, universal scalings and Goldstone modes*, Phys. Lett. A, 279 (2001), pp. 361–369.
[27] J. R. King and S. J. Chapman, *Asymptotics beyond all orders and Stokes lines in nonlinear differential-difference equations*, European J. Appl. Math., 12 (2001), pp. 433–463.
[28] K. Kladko, I. Mitkov, and A. R. Bishop, *Universal scaling of wave propagation failure in arrays of coupled nonlinear cells*, Phys. Rev. Lett., 84 (2000), pp. 4505–4508.
[29] J. Miles, *On Faraday resonance of a viscous liquid*, J. Fluid Mech., 395 (1999), pp. 321–325.
[30] I. Mitkov, K. Kladko, and J. E. Pearson, *Tunable pinning of bursting waves in extended systems with discrete sources*, Phys. Rev. Lett., 81 (1998), pp. 5453–5456.
[31] F. R. N. Nabarro, *Dislocations in a simple cubic lattice*, Proc. Phys. Soc. London, 59 (1947), pp. 256–272.
[32] F. R. N. Nabarro, *Theory of Crystal Dislocations,* Oxford University Press, Oxford, UK, 1967.

[33] R. Peierls, *The size of a dislocation*, Proc. Phys. Soc. London, 52 (1940), pp. 34–37.
[34] V. H. Schmidt, *Exact solution in the discrete case for solitons propagating in a chain of harmonically coupled particles lying in double-minimum potential wells,* Phys. Rev. B, 20 (1979), pp. 4397–4405.
[35] L. I. Slepyan, *Dynamics of a crack in a lattice*, Sov. Phys. Dokl., 26 (1981), pp. 538–540 (translated from Dokl. Akad. Nauk SSSR, 258 (1981), pp. 561–564).
[36] J. M. Speight and R. S. Ward, *Kink dynamics in a novel discrete sine-Gordon system*, Nonlinearity, 7 (1994), pp. 475–484.
[37] J. M. Speight, *A discrete ϕ^4 system without a Peierls-Nabarro barrier*, Nonlinearity, 10 (1997), pp. 1615–1625.
[38] J. M. Speight, *Topological discrete kinks*, Nonlinearity, 12 (1999), pp. 1373–1387.
[39] H. S. J. van der Zant, T. P. Orlando, S. Watanabe, and S. H. Strogatz, *Kink propagation in a discrete system: Observation of phase locking to linear waves*, Phys. Rev. Lett., 74 (1995), pp. 174–177.
[40] B. Zinner, *Existence of traveling wave front solutions for the discrete Nagumo equation*, J. Differential Equations, 96 (1992), pp. 1–27.

SIAM J. APPL. MATH.
Vol. 63, No. 3, pp. 1083–1103

AN ASYMPTOTIC MODEL OF NONADIABATIC CATALYTIC FLAMES IN STAGNATION-POINT FLOW*

STEPHEN B. MARGOLIS† AND TIMOTHY J. GARDNER‡

Abstract. A formal asymptotic model is derived for a nonadiabatic catalytic flame in stagnation-point flow. In the present context, the premixed reaction in the bulk gas is augmented by a surface catalytic reaction on the stagnation plane, where conductive heat losses are allowed to occur. In addition, the thermal effects of a finite-volume combustor are accounted for by allowing for volumetric heat losses from the bulk gas. The analysis exploits the near-equidiffusional limit corresponding to near-unity Lewis numbers, and yields a general nonsteady nonplanar model for the reactionless outer flow subject to boundary conditions that reflect both surface catalysis and distributed chemical reaction in a thin boundary layer. For the case of steady planar combustion, the surface-temperature response indicates the possibility of multiple solution branches, which are shown to be linearly stable, and a corresponding extension of the extinction limit that demonstrates how the presence of a surface catalyst can counterbalance the extinguishing effects of heat loss and stretch in nonadiabatic strained flames. The present model is particularly relevant for small-volume combustors, where the increased surface-to-volume ratio can lead to extinction of the nonadiabatic flame in the absence of a catalyst.

Key words. combustion, catalysis, catalytic flames, nonadiabaticity, extinction limits, stability, asymptotic analysis, matched asymptotic expansions

AMS subject classifications. 80A25, 80A32, 80M35, 41A60

PII. S0036139902402452

1. Introduction.

The effects of catalysis in combustion problems have long been of interest because of a catalyst's ability to enhance what would otherwise be slow and/or incomplete chemical reactions. Combustion applications include the use of catalysts to increase fuel efficiency, accelerate the conversion of intermediate pollutant species, increase the rate of production of desirable products, and allow flames to be sustained in nonadiabatic environments. The last of these is of particular concern in the present work, which is motivated by a growing interest in the use of catalysts to extend extinction limits in small-volume reactors, or microcombustors. Because such a combustor is characterized by relatively large surface-to-volume ratios, the degree of nonadiabaticity associated with conductive and/or radiative heat losses through the walls of the device is a limiting factor in determining its minimum size. Depending on the geometry, coating one or more surfaces with a catalyst can allow combustion to proceed at lower temperatures than would otherwise be possible.

The geometry of the model problem considered here corresponds to a nonadiabatic stretched flame in stagnation-point flow against a catalytic surface. This is illustrated in Figure 1, where the nonadiabatic effects associated with the finite volume of an actual microcombustor are modeled by a volumetric heat-loss term in the energy

*Received by the editors February 11, 2002; accepted for publication (in revised form) July 30, 2002; published electronically February 25, 2003. This work was supported by the United States Department of Energy under contract DE-AC04-94AL85000 as part of Sandia's Laboratory Directed Research and Development (LDRD) Program. The U.S. Government retains a nonexculsive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes.

http://www.siam.org/journals/siap/63-3/40245.html

†Combustion Research Facility, Sandia National Laboratories, Livermore, CA 94551-0969 (margoli@saudia.gov).

‡Advanced Materials Laboratory, Sandia National Laboratories, Albuquerque, NM 87185-1349 (tjgardn@sandia.gov).

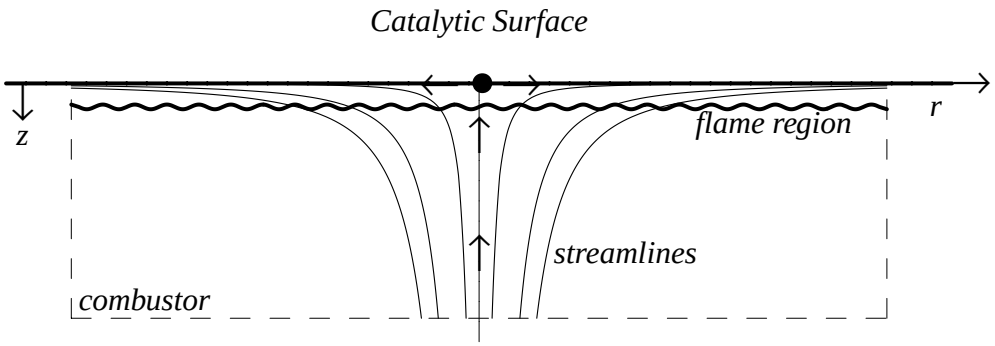


FIG. 1. *Model geometry for a nonadiabatic premixed flame in stagnation-point flow. Heat loss is modeled both volumetrically and by conduction at the catalytic surface. Near extinction, the reaction region is assumed to lie adjacent to the catalytic surface.*

equation and a conductive loss at the catalytic surface. In the absence of heat losses, the semi-infinite geometry depicted in Figure 1 is similar to that considered by others (cf. Law and Sivashinsky [1], Giovangigli and Candel [2], Warnatz et al. [3]), who have analyzed the corresponding adiabatic problem both analytically and numerically. In addition to previous experimental investigations (cf. Law, Ishizuka, and Mizomoto [4], Ikeda, Sato, and Williams [5]), this geometry is also suggested by more recent experiments (Mowery et al. [6]) on small-volume combustors. In the latter application, such a combustor (nominally $2500\mu \times 2500\mu \times 400\mu$) is fed by an inlet tube that blows against a catalytic surface (platinum mounted on a titanium/silicon wafer) and is vented by one or more outlet ports on either the opposite ($2500\mu \times 2500\mu$) face or sides. In those experiments, it was demonstrated, following ignition by the heated catalytic surface, that a nearly flat flame could be sustained under the inlet port, close to the catalytic surface, without further heat addition. In the absence of the catalyst, the level of heat loss was apparently sufficient to extinguish the flame.

The above experimental result was supported by our recent study (Margolis and Gardner [7]) that analyzed the steady-state response of the nonadiabatic semi-infinite problem. In that work, which was restricted to volumetric heat losses, it was shown that a sufficiently strong catalytic effect produced an extension of the extinction limit, allowing a degree of nonadiabaticity that would otherwise extinguish the flame. The present work extends these results by deriving a formal asymptotic model valid for a general nonsteady nonplanar flame. In addition, nonadiabaticity in the present model arises not only from a volumetric heat-loss term in the conservation equation for temperature, but also from a conductive loss term in the boundary condition at the catalytic surface. An analysis of the solution response corresponding to steady planar burning again predicts the possibility of multiple solutions, which are now shown to be linearly stable, and an extension of the extinction limit for sufficiently strong catalytic influences. In the current model, this extinction limit is shown to depend on two parameters associated with the effects of catalysis and on one other parameter that incorporates the combined effects of volumetric heat losses, conductive heat losses at the catalytic surface, and thermal/diffusive effects associated with the deviation in the Lewis number from unity.

2. Model formulation. Based on the above overall description of the model problem, the flame is assumed to be stabilized in a cylindrically symmetric stagnation-point flow field that occupies the domain $0 < \tilde{z} < \infty$, $0 < \tilde{r} < \infty$, where $\tilde{z}$ and $\tilde{r}$ are

the axial and radial coordinates, respectively, and the tildes denote dimensional quantities. The catalytic surface, across which conductive heat transfer is allowed, thus corresponds to the plane $\tilde{z} = 0$, and the effects of heat loss arising from the remaining finite dimensions of the actual combustor are represented in a volumetric fashion. Although one may consider the portion of the flow field of interest to be governed by a boundary-layer formulation (cf. [1]), it turns out that qualitatively identical results are obtained if potential flow and weak thermal expansion are assumed [7]. We thus make these assumptions for simplicity, resulting in the specified classical flow field $(\tilde{u}, \tilde{w}) = \tilde{\nabla}\tilde{\phi}$, where $\tilde{u}$ and $\tilde{w}$ are the radial and axial velocities, respectively, $\tilde{\phi} = -\tilde{a}(\tilde{z}^2 - \tilde{r}^2/2)$ is the velocity potential, and $\tilde{a}$ is the strain rate. Equivalently, $\tilde{u}(\tilde{r}, \tilde{z}) = -\tilde{r}^{-1}\partial\tilde{\psi}/\partial\tilde{z} = \tilde{a}\tilde{r}$ and $\tilde{w}(\tilde{r}, \tilde{z}) = \tilde{r}^{-1}\partial\tilde{\psi}/\partial\tilde{r} = -2\tilde{a}\tilde{z}$, where $\tilde{\psi}(\tilde{r}, \tilde{z}) = -\tilde{a}\tilde{r}^2\tilde{z}$ is the stream function.

Given this flow field, a closed problem for the temperature $\tilde{T}$ and mass fraction Y of the deficient component of the mixture (i.e., the mass fraction of fuel if the initial composition is lean, and the mass fraction of oxidizer if it is rich) in the region $0 < \tilde{z} < \infty$ can be specified. This is conveniently written in dimensionless form by first introducing a characteristic flame temperature $\tilde{T}_f$ (to be determined) and the nondimensional quantities

$$z = \sqrt{\frac{\tilde{a}}{\tilde{\lambda}}}\,\tilde{z}, \quad t = \tilde{a}\tilde{t}, \quad y = \frac{Y}{Y_u}, \quad \Theta = \frac{\tilde{T} - \tilde{T}_u}{\tilde{T}_f - \tilde{T}_u}, \quad \sigma = \frac{\tilde{T}_{sub} - \tilde{T}_u}{\tilde{T}_f - \tilde{T}_u},$$

$$T_f = \frac{\tilde{T}_f}{\tilde{T}_u}, \quad Le = \frac{\tilde{\lambda}}{\tilde{\lambda}_m}, \quad q = \frac{Y_u\tilde{Q}}{\tilde{T}_f - \tilde{T}_u}, \quad H = \frac{\tilde{H}}{\tilde{a}}, \quad K = \frac{\tilde{K}}{\sqrt{\tilde{a}\tilde{\lambda}}}, \tag{1}$$

$$\beta = \frac{\tilde{E}_g}{\tilde{R}^\circ\tilde{T}_f}\big(1 - \tilde{T}_u/\tilde{T}_f\big), \quad \nu = \frac{\tilde{E}_s}{\tilde{E}_g},$$

where ϑ is the angular coordinate, $\tilde{T}_u$ and Y_u are the unburned (ambient) temperature and mass fraction of the fresh mixture, $\tilde{\lambda}$ and $\tilde{\lambda}_m$ are the thermal and mass diffusivities, respectively, $\tilde{Q}$ is the heat release (in units of temperature), $\tilde{E}_g$ is the activation energy of the gas-phase reaction, $\tilde{A}_g$ and n are the rate coefficient and reaction order, respectively, $\tilde{R}^\circ$ is the gas constant, and $\tilde{H}$ is a volumetric heat-loss rate coefficient. The last of these, which is represented in the last term of (3) below, reflects a phenomenological volumetric representation of heat losses across the noncatalytic surfaces of an actual finite-volume combustor. An approximation for $\tilde{H}$ may be obtained from a knowledge of the corresponding surface heat-transfer coefficients and the surface-to-volume ratio of the combustor. Other quantities that arise in the specification of the boundary conditions include Y_s and $\tilde{T}_s$, which represent values at $\tilde{z} = 0$ that are to be determined. (It is assumed that the catalytic surface is highly conductive in the transverse direction, so that $\tilde{T}_s$ and Y_s are independent of $\tilde{r}$ and ϑ.) The temperature $\tilde{T}_{sub}$, on the other hand, is the temperature of the catalyst-coated substrate and is allowed to differ from the ambient temperature $\tilde{T}_u$. The boundary conditions given below also model the catalyst as an exothermic reaction at the surface $\tilde{z} = 0$, distinguished from the reaction rate in the bulk gas by a surface rate coefficient $\tilde{A}_s$ and a different activation energy $\tilde{E}_s$. In this work, the catalytic surface is explicitly allowed to be nonadiabatic ($\tilde{K}$ is the corresponding surface heat-transfer coefficient), with $\tilde{T}_s > (<)\ \tilde{T}_{sub}$ corresponding to a thermal loss (gain). The latter scenario can occur if, for example, the substrate is heated, but aside from this loss (gain), heat produced by the catalytic surface reaction is conducted normal to the

surface into the bulk gas. (There is no convective contribution since $\tilde{w} = 0$ at the surface.) The catalytic effect itself is modeled by assuming that $\tilde{E}_s < \tilde{E}_g$, thus allowing the surface reaction to take place at lower temperatures and consequently raising the temperature of the surrounding region such that the gas-phase reaction, if relatively weak in the absence of catalysis, is further encouraged. Finally, it is also useful to define the reaction-rate parameters Λ_s, Λ_g and their ratio τ as

$$\Lambda_s = \frac{\tilde{A}_s Y_u^{n-1}}{\sqrt{\tilde{a}\tilde{\lambda}}}\, e^{-\tilde{E}_s/\tilde{R}^\circ \tilde{T}_f}, \quad \Lambda_g = \frac{\tilde{A}_g Y_u^{n-1}}{\tilde{a}}\, e^{-\tilde{E}_g/\tilde{R}^\circ \tilde{T}_f},$$
$$\tau = \frac{\Lambda_s}{\Lambda_g} = \sqrt{\frac{\tilde{a}}{\tilde{\lambda}}}\, \frac{\tilde{A}_s}{\tilde{A}_g}\, e^{(\tilde{E}_g - \tilde{E}_s)/\tilde{R}^\circ \tilde{T}_f}. \tag{2}$$

With these definitions, the nondimensional forms of the energy and species conservation equations are given by

$$\frac{\partial \Theta}{\partial t} + r\frac{\partial \Theta}{\partial r} - 2z\frac{\partial \Theta}{\partial z} = \nabla^2 \Theta + q\Lambda_g y^n\, e^{\beta(\Theta-1)/[T_f^{-1} + (1 - T_f^{-1})\Theta]} - H\Theta, \quad 0 < z < \infty, \tag{3}$$

$$\frac{\partial y}{\partial t} + r\frac{\partial y}{\partial r} - 2z\frac{\partial y}{\partial z} = Le^{-1}\nabla^2 y - \Lambda_g y^n\, e^{\beta(\Theta-1)/[T_f^{-1} + (1 - T_f^{-1})\Theta]}, \quad 0 < z < \infty, \tag{4}$$

subject to the boundary conditions

$$\Theta \to 0, \quad y \to 1 \quad \text{as} \ \ z \to \infty, \tag{5}$$

$$\left.\frac{\partial \Theta}{\partial z}\right|_{z=0} = -q\tau\Lambda_g y_s^n\, e^{\nu\beta(\Theta_s - 1)/[T_f^{-1} + (1 - T_f^{-1})\Theta_s]} + K(\Theta_s - \sigma),$$
$$Le^{-1}\left.\frac{\partial y}{\partial z}\right|_{z=0} = \tau\Lambda_g y_s^n\, e^{\nu\beta(\Theta_s - 1)/[T_f^{-1} + (1 - T_f^{-1})\Theta_s]}, \tag{6}$$

where Θ_s and y_s denote the values of Θ and y at $z = 0$ and $\nabla^2 = \partial^2/\partial r^2 + r^{-1}\partial/\partial r + r^{-2}\partial^2/\partial \vartheta^2 + \partial^2/\partial z^2$. Thus, in addition to the distributed chemical reaction and volumetric heat losses represented by the nondiffusive terms on the right-hand sides of (3) and (4), surface combustion and conductive thermal losses across the catalytic surface are also specifically accounted for in the present model by the terms on the right-hand sides of the boundary conditions (6).

3. Asymptotic analysis of the model. It is assumed in the present work that the Zel'dovich number $\beta \gg 1$ and that $\nu \sim \mathrm{O}(1)$. Thus, the activation energies of both the bulk-gas and catalytic reactions are taken to be large, although it is logically expected that $\nu < 1$, implying that the catalytic surface reaction can be sustained at lower temperatures than the distributed gas-phase reaction. In this realistic asymptotic limit, there are two possible burning regimes, corresponding to a thin gaseous reaction zone that is either adjacent to the catalytic surface or an O(1) distance away. Both regimes have been discussed in the context of the adiabatic problem (cf. [1], [2]), where it has been heuristically argued, based on the strained nature of the flow field, that as the strain rate increases toward extinction, the gas flame will tend to either intrude onto the catalytic surface for $Le < 1$ or remain at an O(1) standoff distance for $Le > 1$ [1]. This conclusion is based on the physical argument that a

planar flame gains chemical energy and loses thermal energy by diffusion in the normal direction with respect to the flame, whereas convection, which is nonnormal to the flame, supplies chemical energy but removes thermal energy. Consequently, if energy losses from the flame to the diverging flow via thermal diffusion outweigh energy gains to the flame from mass diffusion ($Le > 1$), the flame temperature T_f would be expected to decrease and the flame would tend to extinguish prior to being pushed against the catalytic surface. On the other hand, if $Le < 1$, T_f would tend to increase and extinction would only tend to occur (due to sufficiently reduced residence time) after the flame had intruded against the surface. However, a detailed analysis [2] suggests more generally that the flame will lie adjacent to the stagnation surface prior to extinction either when the Lewis number is less than a critical value that is somewhat greater than unity or when the activation-energy ratio $\nu < 1/2$, corresponding to a sufficiently low surface activation energy and hence a more active catalytic reaction. It is only in this regime that catalysis can play a significant role because, if the bulk-gas reaction goes to completion at a finite standoff distance, reactants are depleted before reaching the surface and the catalyst then plays no role.

As in our previous study [7], the present work seeks in part to investigate the role of catalysis in counterbalancing the extinguishing effects of heat loss. We thus restrict our analysis to the intrusive regime in which the thin distributed reaction zone lies adjacent to the catalytic surface. Based on previous asymptotic studies of nonadiabatic combustion problems (cf. Matkowsky and Olagunju [8], Booty, Margolis, and Matkowsky [9], Kaper et al. [10], Margolis and Johnston [11]), it is clear that extinction then occurs for $O(\beta^{-1})$ values of the heat-loss parameters H and K. Accordingly, we define the corresponding scaled parameters h and k as

$$H = \frac{h}{\beta}, \qquad K = \frac{k}{\beta}, \tag{7}$$

where extinction is expected to occur for sufficiently large values of h and k. In addition, it will prove useful, for the purpose of deriving a closed asymptotic model (cf. Matkowsky and Sivashinsky [12]), to realistically restrict consideration to the near-equidiffusional regime

$$Le = 1 + \frac{l}{\beta}, \tag{8}$$

where l is the scaled departure of the Lewis number from unity.

Based on the preceding discussion, there are now two regions to consider in the asymptotic limit $\beta \gg 1$; namely, an outer reactionless region $z > 0$, in which temperatures are sufficiently low that the reaction terms in (3) and (4) become exponentially small, and a thin distributed reaction zone adjacent to the surface $z = 0$, on which the catalytic reaction occurs. Considering first the outer region, where z and $1 - \Theta$ are both $O(1)$, we seek solutions in the expanded form

$$\Theta^{(o)} \sim \Theta_0 + \beta^{-1}\Theta_1 + \cdots, \qquad y^{(o)} \sim y_0 + \beta^{-1}y_1 + \cdots. \tag{9}$$

Substituting these scalings and expansions into the reactionless version of (8)–(10), we obtain, at the zeroth order, the equations

$$\begin{aligned} &\frac{\partial \Theta_0}{\partial t} + r\frac{\partial \Theta_0}{\partial r} - 2z\frac{\partial \Theta_0}{\partial z} = \nabla^2 \Theta_0, \\ &\frac{\partial y_0}{\partial t} + r\frac{\partial y_0}{\partial r} - 2z\frac{\partial y_0}{\partial z} = \nabla^2 y_0, \quad 0 < z < \infty, \end{aligned} \tag{10}$$

subject to

$$\Theta_0 \to 0, \quad y_0 \to 1 \ \text{ as } \ z \to \infty, \qquad \Theta_0 \to 1, \quad y_0 \to 0 \ \text{ as } \ z \to 0, \tag{11}$$

where the boundary conditions at $z = 0$ are immediately deduced from the form of the inner expansions given below. Defining the enthalpy variable $S_0 = \Theta_0 + y_0$ and summing (10) and (11) thus determines a closed problem for S_0 as

$$\frac{\partial S_0}{\partial t} + r\frac{\partial S_0}{\partial r} - 2z\frac{\partial S_0}{\partial z} = \nabla^2 S_0, \qquad 0 < z < \infty, \tag{12}$$

$$S_0 \to 1 \ \text{ as } \ z \to 0, \ \infty. \tag{13}$$

Assuming compatible initial conditions (those that are consistent with the result (14) below), the solution to (12) and (13) is simply $S_0 = 1$, and thus

$$S_0 = y_0 + \Theta_0 = 1. \tag{14}$$

At the first order, the equations for Θ_1 and y_1 are given by

$$\begin{aligned} \frac{\partial \Theta_1}{\partial t} + r\frac{\partial \Theta_1}{\partial r} - 2z\frac{\partial \Theta_1}{\partial z} &= \nabla^2 \Theta_1 - h\Theta_0, \\ \frac{\partial y_1}{\partial t} + r\frac{\partial y_1}{\partial r} - 2z\frac{\partial y_1}{\partial z} &= \nabla^2 y_1 - l\nabla^2 y_0, \qquad 0 < z < \infty, \end{aligned} \tag{15}$$

subject to

$$\Theta_1 \to 0, \quad y_1 \to 0 \ \text{ as } \ z \to \infty, \tag{16}$$

and appropriate matching conditions, given below, as $z \to 0$. Defining the second-order enthalpy variable $S_1 = \Theta_1 + y_1$, the equation for S_1, obtained from summing the equations of (15) and the use of (14), is given by

$$\frac{\partial S_1}{\partial t} + r\frac{\partial S_1}{\partial r} - 2z\frac{\partial S_1}{\partial z} = \nabla^2 S_1 + l\nabla^2 \Theta_0 - h\Theta_0, \qquad 0 < z < \infty, \tag{17}$$

where

$$S_1 \to 0, \quad \Theta_0 \to 0 \ \text{ as } \ z \to \infty. \tag{18}$$

In the inner region adjacent to the surface $z = 0$, we introduce the stretched coordinate η and the scaled rate coefficients $\hat{\lambda}$ and $\hat{\tau}$ according to

$$\eta = \beta z, \quad \Lambda_g = \beta^{n+1}\hat{\lambda}, \quad \tau = \beta^{-1}\hat{\tau}, \tag{19}$$

and seek solutions in the expanded form

$$\Theta^{(i)} \sim 1 + \beta^{-1}\theta_1 + \beta^{-2}\theta_2 + \cdots, \qquad y^{(i)} \sim \beta^{-1}\zeta_1 + \beta^{-2}\zeta_2 + \cdots. \tag{20}$$

Substituting these expansions and scalings into (3), (4), and (6), we obtain the first-order inner problem as

$$\frac{\partial^2 \theta_1}{\partial \eta^2} + q\hat{\lambda}\zeta_1^n e^{\theta_1} = 0, \qquad \frac{\partial^2 \zeta_1}{\partial \eta^2} - \hat{\lambda}\zeta_1^n e^{\theta_1} = 0, \tag{21}$$

subject to the inner boundary and matching conditions

$$\left.\frac{\partial\theta_1}{\partial\eta}\right|_{\eta=0} = -q\hat{\tau}\hat{\lambda}\zeta_s^n e^{\nu\theta_s}, \quad \left.\frac{\partial\zeta_1}{\partial\eta}\right|_{\eta=0} = \hat{\tau}\hat{\lambda}\zeta_s^n e^{\nu\theta_s}, \tag{22}$$

$$\theta_1 \sim \Theta_1\big|_{z=0} + \eta\left.\frac{\partial\Theta_0}{\partial z}\right|_{z=0}, \quad \zeta_1 \sim y_1\big|_{z=0} + \eta\left.\frac{\partial y_0}{\partial z}\right|_{z=0}, \tag{23}$$

where θ_s and ζ_s, respectively, denote θ_1 and ζ_1 evaluated at $\eta = 0$. Combining the equations of (21) so as to eliminate the reaction term and integrating the result, we thus obtain

$$\frac{\partial\theta_1}{\partial\eta} + q\frac{\partial\zeta_1}{\partial\eta} = 0, \tag{24}$$

where the boundary conditions (22) have been used to evaluate the constant of integration. Applying the matching conditions (23) and using the result (14) then gives the requirement

$$\left.\frac{\partial\Theta_0}{\partial z}\right|_{z=0} + q\left.\frac{\partial y_0}{\partial z}\right|_{z=0} = (q-1)\left.\frac{dy_0}{dz}\right|_{z=0} = 0. \tag{25}$$

Since y_0 must satisfy the problem given by (10) and (11), the last equality can only be satisfied in general if $q = 1$, which, based on the definition of q in (1), determines the characteristic flame temperature $\tilde{T}_f$ as the adiabatic flame temperature $\tilde{T}_f = \tilde{T}_u + Y_u\tilde{Q}$. Using this result and integrating (24) then gives, upon applying the matching conditions (23), the relation

$$\theta_1 + \zeta_1 = S_1\big|_{z=0}. \tag{26}$$

Returning to the first equation in (21), we substitute (26) for ζ_1 into the reaction-rate expression to obtain a scalar equation for θ_1 as

$$\frac{\partial^2\theta_1}{\partial\eta^2} + \hat{\lambda}\big(S_1\big|_{z=0} - \theta_1\big)^n e^{\theta_1}. \tag{27}$$

Multiplying this result by $\partial\theta_1/\partial\eta$ and integrating, we obtain, upon use of the matching condition for θ_1, the first integral

$$\left(\frac{\partial\theta_1}{\partial\eta}\right)^2 + 2\hat{\lambda}\int_{-\infty}^{\theta_1}\big(S_1\big|_{z=0} - \bar{\theta}_1\big)^n e^{\bar{\theta}_1}d\bar{\theta}_1 = \left(\left.\frac{\partial\Theta_0}{\partial z}\right|_{z=0}\right)^2. \tag{28}$$

Thus, evaluating (28) at $\eta = 0$ according to the first boundary condition in (22) yields an implicit relation for θ_s given by

$$\hat{\tau}^2\hat{\lambda}^2\big(S_1\big|_{z=0} - \theta_s\big)^{2n} e^{2\nu\theta_s} + 2\hat{\lambda}G_n\big(\theta_s; S_1\big|_{z=0}\big) = \left(\left.\frac{\partial\Theta_0}{\partial z}\right|_{z=0}\right)^2, \tag{29}$$

where (26) implies (since the surface mass fraction $\zeta_1 \geq 0$) the physical restriction $\theta_s \leq S_1\big|_{z=0}$ and

$$G_n\big(\theta_s; S_1\big|_{z=0}\big) = \int_{-\infty}^{\theta_s}\big(S_1\big|_{z=0} - \bar{\theta}_1\big)^n e^{\bar{\theta}_1}d\bar{\theta}_1 = \int_{-\theta_s}^{\infty}\big(\chi + S_1\big|_{z=0}\big)^n e^{-\chi}\,d\chi. \tag{30}$$

An additional relation that will be needed is obtained from a consideration of the second-order inner problem, which, since $q = 1$, is given by

$$\begin{aligned} &\frac{\partial^2\theta_2}{\partial\eta^2} + \hat{\lambda} f(\zeta_1, \zeta_2, \theta_1, \theta_2; n) e^{\theta_1} = 0, \\ &\frac{\partial^2\zeta_2}{\partial\eta^2} - l\frac{\partial^2\zeta_1}{\partial\eta^2} - \hat{\lambda} f(\zeta_1, \zeta_2, \theta_1, \theta_2; n) e^{\theta_1} = 0, \end{aligned} \tag{31}$$

$$\begin{aligned} &\left.\frac{\partial\theta_2}{\partial\eta}\right|_{\eta=0} = -\hat{\tau}\hat{\lambda} g(\zeta_s, \hat{\zeta}_s, \theta_s, \hat{\theta}_s; n) e^{\nu\theta_s} + k(1-\sigma), \\ &\left.\frac{\partial\zeta_2}{\partial\eta}\right|_{\eta=0} - l\left.\frac{\partial\zeta_1}{\partial\eta}\right|_{\eta=0} = \hat{\tau}\hat{\lambda} g(\zeta_s, \hat{\zeta}_s, \theta_s, \hat{\theta}_s; n) e^{\nu\theta_s}, \end{aligned} \tag{32}$$

$$\begin{aligned} &\theta_2 \sim \Theta_2\big|_{z=0} + \eta\left.\frac{\partial\Theta_1}{\partial z}\right|_{z=0} + \frac{1}{2}\eta^2\left.\frac{\partial^2\Theta_0}{\partial z^2}\right|_{z=0}, \\ &\zeta_2 \sim y_2\big|_{z=0} + \eta\left.\frac{\partial y_1}{\partial z}\right|_{z=0} + \frac{1}{2}\eta^2\left.\frac{\partial^2 y_0}{\partial z^2}\right|_{z=0}, \end{aligned} \tag{33}$$

where $\hat{\theta}_s$ and $\hat{\zeta}_s$, respectively, denote θ_2 and ζ_2 evaluated at $\eta = 0$, $f(\zeta_1, \zeta_2, \theta_1, \theta_2; n) = n\zeta_2\zeta_1^{n-1} + \zeta_1^n\big[\theta_2 - (1 - T_f^{-1})\theta_1^2\big]$, and $g(\zeta_s, \hat{\zeta}_s, \theta_s, \hat{\theta}_s; n) = n\hat{\zeta}_s\zeta_s^{n-1} + \nu\zeta_s^n\big[\hat{\theta}_s - (1 - T_f^{-1})\theta_s^2\big]$. Integrating the sum of (31) and applying the boundary conditions (32), we obtain

$$\frac{\partial\theta_2}{\partial\eta} + \frac{\partial\zeta_2}{\partial\eta} + l\frac{\partial\theta_1}{\partial\eta} = k(1-\sigma), \tag{34}$$

where we have used the result (24) with $q = 1$. Applying the matching conditions (33) thus yields the condition

$$\left.\frac{\partial S_1}{\partial z}\right|_{z=0} = -l\left.\frac{\partial\Theta_0}{\partial z}\right|_{z=0} + k(1-\sigma). \tag{35}$$

Collecting the preceding results, we have thus succeeded in deriving a closed asymptotic model for the outer temperature- and enthalpy-perturbation variables Θ_0 and S_1, and the surface-temperature coefficient θ_s. In particular, solutions for these variables are determined from the first equality of (10) and from (17) in the half-space $0 < z < \infty$, the corresponding boundary conditions (11) and (18) at $z = 0$ and as $z \to \infty$, and the derived conditions (29) and (35) at $z = 0$. In what follows, we apply this model to obtain multiple basic solutions corresponding to steady planar flames, as well as corresponding extinction limits beyond which no solutions of this type exist.

4. Steady plane-flame solutions. The basic solution $\bar{\Theta} = \Theta_0(z)$, $\bar{S} = S_1(z)$, and $\bar{\theta}_s = \theta_s = const.$, corresponding to steady planar combustion, is governed by the time-independent, one-dimensional version of (17) and the first equality of (10), subject to (11), (18), (29), and (35). In particular,

$$-2z\frac{d\bar{\Theta}}{\partial z} = \frac{d^2\bar{\Theta}}{dz^2}, \quad -2z\frac{d\bar{S}}{dz} = \frac{d^2\bar{S}}{dz^2} + l\frac{d^2\bar{\Theta}}{dz^2} - h\bar{\Theta}, \quad 0 < z < \infty, \tag{36}$$

$$\bar{S} \to 0, \quad \bar{\Theta} \to 0 \ \text{ as } \ z \to \infty. \tag{37}$$

$$\bar{\Theta}(0) = 1, \quad \left.\frac{d\bar{S}}{dz}\right|_{z=0} = -l \left.\frac{d\bar{\Theta}}{dz}\right|_{z=0} + k(1-\sigma), \tag{38}$$

$$\hat{\tau}^2 \hat{\lambda}^2 \left[\bar{S}(0) - \bar{\theta}_s\right]^{2n} e^{2\nu\bar{\theta}_s} + 2\hat{\lambda} G_n\left(\bar{\theta}_s; \bar{S}(0)\right) = \left(\left.\frac{d\bar{\Theta}}{dz}\right|_{z=0} \right)^2, \quad \bar{\theta}_s \le \bar{S}(0). \tag{39}$$

We remark that while (36)–(38) are sufficient to determine $\bar{\Theta}$ and $\bar{S}$, the requirement that (39) produce physical solutions for the surface-temperature coefficient $\bar{\theta}_s$ will yield parameter limits associated with extinction.

The solution for $\bar{\Theta}$ satisfying the first equation of (36) and the boundary conditions in (37) and (38) is easily obtained as

$$\bar{\Theta}(z) = \operatorname{erfc}(z) = 1 - \frac{2}{\sqrt{\pi}} \int_0^z e^{-\bar{z}^2} d\bar{z}. \tag{40}$$

Substituting this result into the second equation of (36) then gives an inhomogeneous equation for $\bar{S}$ as

$$\frac{d^2\bar{S}}{dz^2} + 2z \frac{d\bar{S}}{dz} = h \operatorname{erfc}(z) - \frac{4l}{\sqrt{\pi}}\, z\, e^{-z^2}. \tag{41}$$

Homogeneous solutions of (41) are 1 and $\operatorname{erfc}(z)$, and thus a particular solution of (41) is obtained, using the variation-of-parameters formula, as

$$\begin{aligned} \bar{S}_p = \frac{\sqrt{\pi}}{2} \Bigg\{ & - \int_z^\infty \operatorname{erfc}(\bar{z}) \left[h \operatorname{erfc}(\bar{z}) - \frac{4l}{\sqrt{\pi}} \bar{z}\, e^{-\bar{z}^2} \right] e^{\bar{z}^2} d\bar{z} \\ & - \operatorname{erfc}(z) \int_0^z \left[h \operatorname{erfc}(\bar{z}) - \frac{4l}{\sqrt{\pi}} \bar{z}\, e^{-\bar{z}^2} \right] e^{\bar{z}^2} d\bar{z} \Bigg\}. \end{aligned} \tag{42}$$

Hence, using the fact that $4\int_z^\infty \bar{z} \operatorname{erfc}(\bar{z})\, d\bar{z} = (1-2z^2)\operatorname{erfc}(z) + (2/\sqrt{\pi})\, z \exp(-z^2)$, the complete solution of (41) can be written as

$$\begin{aligned} \bar{S}(z) = c_1 + c_2 \operatorname{erfc}(z) - \frac{\sqrt{\pi}}{2} h \left[\int_z^\infty e^{\bar{z}^2} \operatorname{erfc}^2(\bar{z})\, d\bar{z} + \operatorname{erfc}(z) \int_0^z e^{\bar{z}^2} \operatorname{erfc}(\bar{z})\, d\bar{z} \right] \\ + \frac{l}{2} \left[\operatorname{erfc}(z) + \frac{2}{\sqrt{\pi}}\, z\, e^{-z^2} \right], \end{aligned} \tag{43}$$

where application of the boundary conditions (37) and (38) for $\bar{S}$ determines c_1 and c_2 as

$$c_1 = 0, \qquad c_2 = -l - \frac{\sqrt{\pi}}{2}\, k(1-\sigma). \tag{44}$$

The solutions $\bar{\Theta}$ and $\bar{S}$ are displayed in Figure 2.

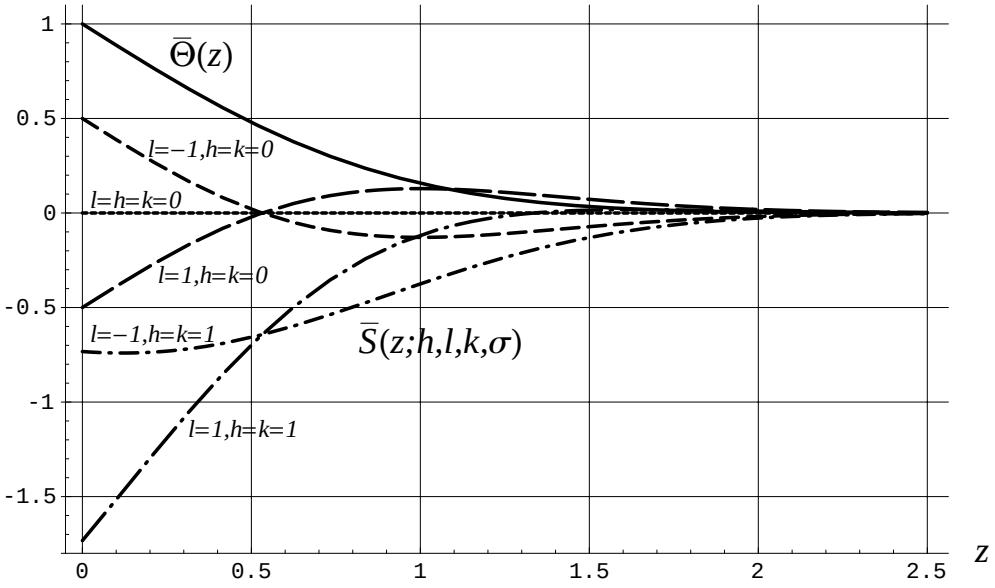


FIG. 2. *Outer solution profiles* $\bar{\Theta}(z)$ *and* $\bar{S}(z;h,l,k,\sigma)$. *The several curves for* $\bar{S}$ *were drawn for* $\sigma = 0$ *and various values of the parameters* h, l, *and* k *as indicated.*

From (40), (43), and (44), we note that $\bar{S}(0)$ is given by

$$\bar{S}(0) = -\frac{l}{2} - \frac{\sqrt{\pi}}{2}\big[k(1-\sigma)+\gamma h\big], \qquad \gamma = \int_0^\infty e^{\bar{z}^2}\operatorname{erfc}^2(\bar{z})\,d\bar{z} \doteq 0.391066. \tag{45}$$

Substituting this result and $d\bar{\Theta}/dz|_{z=0} = -2/\sqrt{\pi}$ into (39) then gives the condition for $\bar{\theta}_s$ as

$$\frac{2}{\pi\hat{\lambda}} = \left[G_n(\bar{\theta}_s;-\Psi) + \frac{1}{2}\hat{\tau}^2\hat{\lambda}(-\bar{\theta}_s-\Psi)^{2n}e^{2\nu\bar{\theta}_s}\right], \quad -\bar{\theta}_s \geq \Psi, \tag{46}$$

where

$$\Psi \equiv \frac{l}{2} + \frac{\sqrt{\pi}}{2}\big[k(1-\sigma)+\gamma h\big] = -\bar{S}(0) \tag{47}$$

and the restriction $-\bar{\theta}_s \geq \Psi$ follows from (26) evaluated at $z = 0$ and the fact that the leading-order mass-fraction coefficient $\zeta_1 \geq 0$. The parameter group Ψ represents a sum of parametric influences arising from conductive losses ($k > 0$) at the catalytic surface, volumetric heat losses ($h > 0$) associated with the finite size of the combustor, and unequal rates of thermal and mass diffusion ($l \neq 0$). We note that Ψ is generally positive for sufficiently large values of the heat-loss coefficients, which, given our present focus on nonadiabaticity, is the main regime of interest here. However, for sufficiently small values of these parameters and a Lewis number less than unity ($l < 0$), Ψ can take on negative values as well. In that case, the minimum value $-\bar{\theta}_s = \Psi$ corresponds to a surface temperature that exceeds the adiabatic flame temperature for freely-propagating flames, consistent with the known dependence of the flame temperature on Lewis number in the present strained geometry (cf. [1], [7]).

Equation (46) provides a relationship for the surface-temperature coefficient $\bar{\theta}_s$, which, according to (20), reflects the scaled leading-order temperature perturbation

at the nonadiabatic catalytic surface from the normalized flame temperature of unity. As discussed in detail in the following section, (46) also admits a physical solution for $\bar{\theta}_s$ only for parameter values that do not exceed a critical condition, and thus (46) also defines an extinction criterion for steady planar burning, beyond which the present solution does not exist. As will also be shown below, the result (46) is consistent with our previous analysis [7], which was restricted to a steady, planar flame at the outset, with $k = 0$ and $Le \sim \mathrm{O}(1)$.

5. Extinction limits for steady planar combustion. The integral represented by $G_n(\bar{\theta}_s; -\Psi)$, which can also be expressed in terms of the incomplete gamma function $\Gamma(b, x) = \int_x^\infty t^{b-1}e^{-t}\,dt$ as $G_n(\bar{\theta}_s; -\Psi) = e^{-\Psi}\Gamma(n+1, -\bar{\theta}_s - \Psi)$, can be evaluated explicitly for integer values of the reaction order n. In particular, we have

$$G_0(\bar{\theta}_s; \Psi) = e^{\bar{\theta}_s}, \quad G_1(\bar{\theta}_s; -\Psi) = \big(1 - \bar{\theta}_s - \Psi\big)e^{\bar{\theta}_s}, \qquad G_2(\bar{\theta}_s; -\Psi) = \big[1 + \big(1 - \bar{\theta}_s - \Psi\big)^2\big]e^{\bar{\theta}_s}, \ldots. \tag{48}$$

Hence, restricting further consideration to the case $n = 1$, (46) may be written as

$$\alpha_1 = (1 - \bar{\theta}_s - \Psi)e^{\bar{\theta}_s} + \alpha_2(-\bar{\theta}_s - \Psi)^2 e^{2\nu\bar{\theta}_s}, \quad -\bar{\theta}_s \geq \Psi, \tag{49}$$

where we have defined the parameters α_1 and α_2 as

$$\alpha_1 = \frac{2}{\pi\hat{\lambda}}, \qquad \alpha_2 = \frac{1}{2}\hat{\tau}^2\hat{\lambda}. \tag{50}$$

Here, α_1, which is inversely proportional to $\hat{\lambda}$ and hence Λ_g, may be regarded, according to the definition of Λ_g in (2), as a measure of either the strain rate $\tilde{a}$ or the reciprocal of the gas-phase reaction rate. On the other hand, α_2, which is proportional to $\hat{\tau}^2\hat{\lambda}$ or to $(\Lambda_s/\Lambda_g)^2\Lambda_g$, is independent of $\tilde{a}$ but does represent a relative scaled measure of the surface reaction-rate coefficient with respect to that of the bulk gas (in units of the gas-phase rate). Equation (49) is thus an implicit relation for $\bar{\theta}_s$ as a function of the four parameter groups α_1, α_2, ν, and Ψ.

Before proceeding, we note that (49) is similar in form to that derived previously in the absence of heat losses (cf. [1]) and in the presence of volumetric heat losses only [7]. In the latter case, the results are equivalent, provided we set $k = 0$ in the definition of Ψ and restrict Le, which was regarded as O(1) in [7], to be an $\mathrm{O}(1/\beta)$ perturbation of unity, as in the present study. Indeed, setting $k = 0$ and defining $\theta_s^* = \bar{\theta}_s + l/2$, $\hat{\alpha}_1 = \alpha_1 e^{l/2}$, and $\hat{\alpha}_2 = \alpha_2 e^{l/2-\nu l}$, we recover, to O(1), the corresponding expression obtained in [7]. In deriving this result, we note that because $T_f = 1 + Q/\sqrt{Le}$ (cf. [1], [7]), where $Q = \tilde{Y}_u\tilde{Q}/\tilde{T}_u$, the Lewis number enters into the definition of both Λ_g and Λ_s (or τ) such that, to O(1), $\hat{\alpha}_1$ and $\hat{\alpha}_2$ are equivalent to the α_1 and α_2 defined in [7].

The surface-temperature response predicted by (49) is briefly summarized as follows. In particular, rather than consider the implicit solution of (49) for $\bar{\theta}_s$, it is convenient to instead analyze the explicit algebraic behavior of $\alpha_1(\bar{\theta}_s; \alpha_2, \Psi, \nu)$.Thus, in the absence of catalysis ($\alpha_2 = 0$), equation (49) is reduced to

$$\alpha_1 = (1 - \bar{\theta}_s - \Psi)e^{\bar{\theta}_s}, \tag{51}$$

where we again note that physical solutions are always restricted to $-\bar{\theta}_s \geq \Psi$, where the lower limit corresponds, according to (26), to complete consumption of reactants

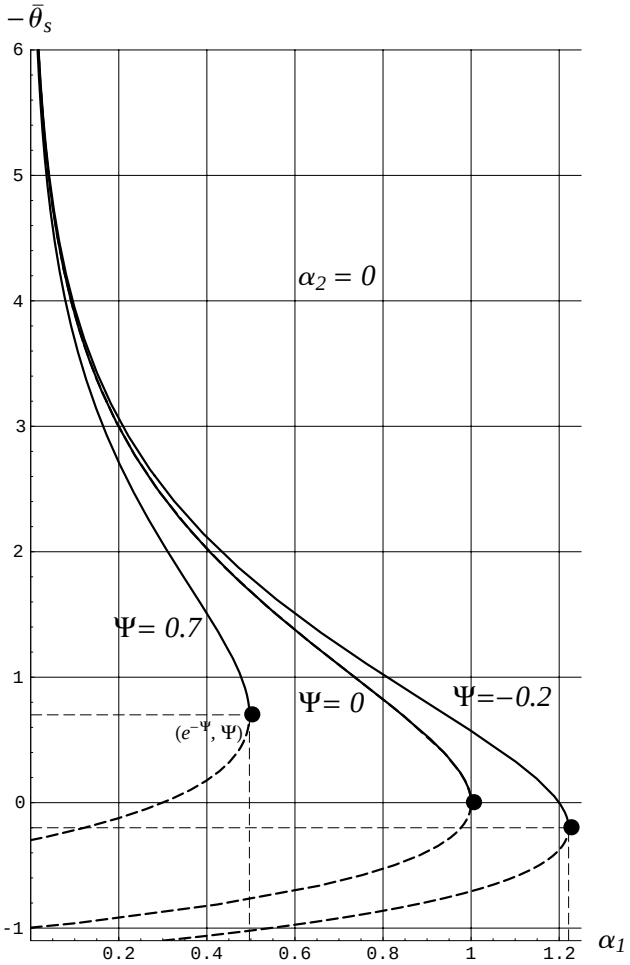


FIG. 3. *Solution response in the absence of catalysis ($\alpha_2 = 0$) for reaction order $n = 1$. Physical solutions (solid curve) are restricted to $-\bar{\theta}_s \geq \Psi$. A steady planar solution does not exist for $\alpha_1 > \alpha_1^e$, where α_1^e is the extinction limit.*

by the gas-phase reaction (i.e., $\zeta_s = 0$). The solution curve $\alpha_1(-\bar{\theta}_s)$, plotted as $\bar{\theta}_s(\alpha_1)$, is shown in Figure 3. Since no steady planar solution exists for $\alpha_1 > e^{-\Psi}$, we interpret this critical value of the strain-rate parameter α_1 as an extinction limit. According to the definition (47) of Ψ, this limit decreases exponentially with increasing values of the heat-loss parameters h and k, and with increasing values of the Lewis-number perturbation from unity. As indicated above, the latter reflects the decrease in flame temperature associated with increased thermal losses via diffusion to the strained flow field. In terms of Ψ, the critical condition for extinction is thus $\Psi > -\ln \alpha_1$, so that smaller values of the strain rate allow the flame to tolerate larger thermal losses.

For the more general response in the presence of catalysis ($\alpha_2 > 0$), we first calculate $d\alpha_1/d(-\bar{\theta}_s)$ from (49) as

$$\frac{d\alpha_1}{d(-\bar{\theta}_s)} = (-\bar{\theta}_s - \Psi)e^{-(-\bar{\theta}_s)} \left\{ -1 + 2\alpha_2 \left[1 - \nu(-\bar{\theta}_s - \Psi) \right] e^{(1-2\nu)(-\bar{\theta}_s)} \right\}. \tag{52}$$

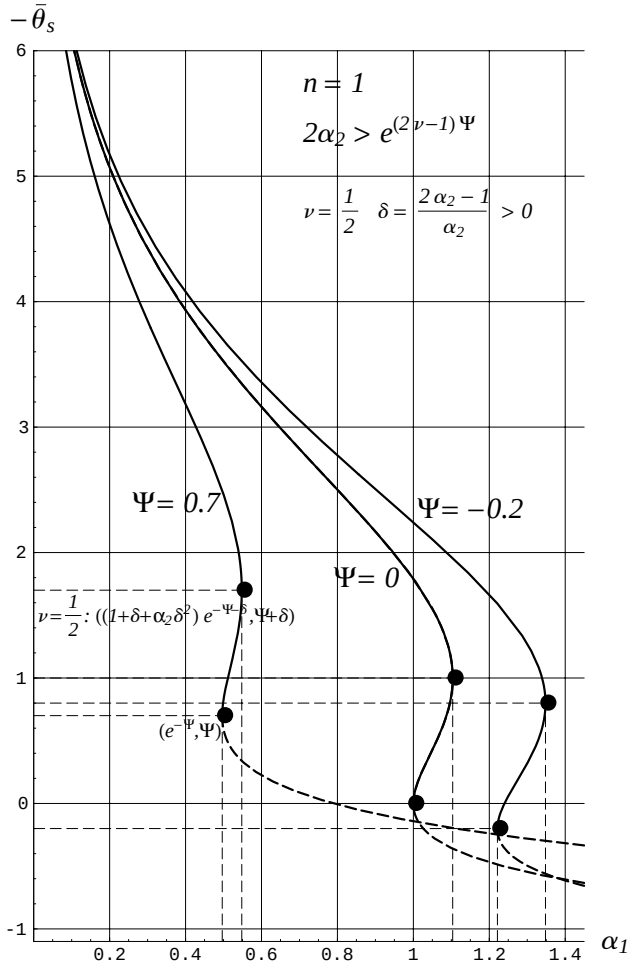


FIG. 4(a). *Solution response for activation-energy ratio $\nu = 1/2$ and $\alpha_2 > 1/2$. For $\alpha_2 > 1/2$, corresponding to a relatively strong surface reaction, an extension of the extinction limit to the higher value $\alpha_1^e = (1 + \delta + \alpha_2\delta^2)\, e^{-\Psi-\delta}$ is realized. The curves were drawn for $\alpha_2 = 1$ and the indicated values of Ψ.*

From (52) we thus conclude that $d\alpha_1/d(-\bar{\theta}_s) = 0$ at $-\bar{\theta}_s = \Psi$, corresponding to $\alpha_1 = e^{-\Psi}$, and may also be zero at value(s) of $-\bar{\theta}_s$ that satisfy the condition

$$2\alpha_2\big[1 - \nu(-\bar{\theta}_s - \Psi)\big] = e^{(2\nu-1)(-\bar{\theta}_s)}. \tag{53}$$

For example, if $\nu = 1/2$, corresponding to the case in which the activation energy of the catalytic surface reaction is half that of the distributed reaction in the bulk gas, (53) is satisfied when $-\bar{\theta}_s = \Psi + \delta$, where $\delta = (2\alpha_2 - 1)/\alpha_2$. Thus, for $\alpha_2 > 1/2$ (i.e., for $\delta > 0$), there exists a second physical solution of (49), corresponding to $\alpha_1 = \alpha_1^e = (1 + \delta + \alpha_2\delta^2)e^{-\Psi-\delta}$, for which $d\alpha_1/d(-\bar{\theta}_s) = 0$. On the other hand, for $\alpha_2 < 1/2$ (i.e., for $\delta < 0$), this additional solution is unphysical since it occurs for $-\bar{\theta}_s < \Psi$. The consequences of a physical root of (53) are evident from Figures 4(a) and 4(b), which are drawn for the case $\nu = 1/2$ just described. For $\alpha_2 > 1/2$, corresponding to a sufficiently vigorous surface reaction, the extinction limit is increased (since

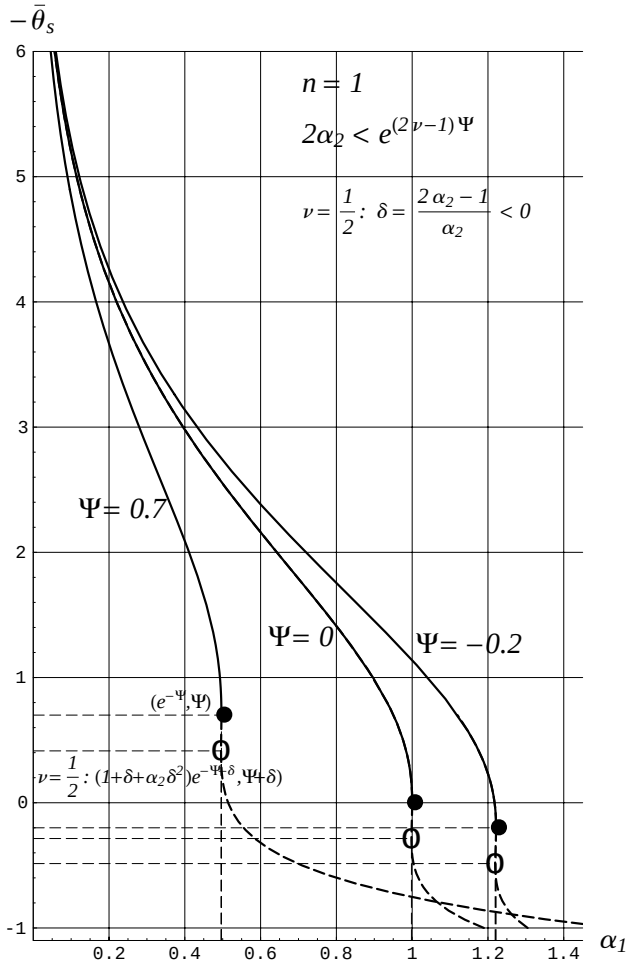


FIG. 4(b). *Solution response for activation-energy ratio $\nu = 1/2$ and $2\alpha_2 < 1$. For $\alpha_2 < 1/2$, corresponding to a relatively weak catalytic influence, the solution response is modified accordingly, but the extinction limit is the same as that obtained in the absence of catalysis. The curves were drawn for $\alpha_2 = 7/16$ and the indicated values of Ψ.*

$d^2\alpha_1/d(-\bar{\theta}_s)^2|_{-\bar{\theta}_s=\Psi} > 0$) to the value $\alpha_1 = \alpha_1^e$ given above (Figure 4(a)). In addition, the solution becomes multivalued for $e^{-\Psi} < \alpha_1 < \alpha_1^e$, implying both a high- and low-temperature solution (corresponding to a small and large value of $-\bar{\theta}_s$, respectively) for α_1 within this range. On the other hand, for $\alpha_2 < 1/2$ (Figure 4(b)), which corresponds to a relatively weak surface reaction, the extinction limit $\alpha_1 = e^{-\Psi}$ remains the same as that in the complete absence of a catalytic reaction. That is, even though a weak catalytic reaction does modify the solution response relative to the noncatalytic case, the maximum possible value of α_1 is unchanged.

The preceding discussion for the case $\nu = 1/2$ implies that the effects of a sufficiently active surface reaction at a reduced activation energy allows for a lower flame temperature (i.e., a larger value of $-\bar{\theta}_s$), thereby extending the extinction limit. That is, reactants that pass through the gas-phase reaction region due to higher rates of strain and/or lower gas-phase reactivity are still able to undergo at least partial conversion at the catalytic surface and thus contribute to the overall heat release. If the

catalytic reaction is weak or absent altogether, this additional opportunity for conversion is reduced or eliminated, and consequently, the gas flame cannot sustain itself at values of the strain-rate parameter α_1 that are larger than the critical value corresponding to extinction in the noncatalytic case. In terms of the parameter group Ψ, the critical value corresponding to extinction is raised from $\Psi = -\ln \alpha_1$ for $\alpha_2 < 1/2$ to $\Psi = \Psi^e = -\ln \alpha_1 + \ln(1+\delta+\alpha_2\delta^2) - \delta$. As δ decreases to zero from above (i.e., as α_2 approaches $1/2$), the maximum rate of heat loss that can be tolerated for a given value of α_1 is reduced to the same limit as for the noncatalytic problem.

These results are readily extended beyond the special case $\nu = 1/2$ as follows. First, it is useful to differentiate (52) to obtain

$$\left.\frac{d^2\alpha_1}{d(-\bar{\theta}_s)^2}\right|_{-\bar{\theta}_s=\Psi} = e^{-\Psi}\left[2\alpha_2\, e^{(1-2\nu)\Psi} - 1\right]. \tag{54}$$

Thus, at the minimum value $-\bar{\theta}_s = -\bar{\theta}_s^0 = \Psi$, it is seen that $d^2\alpha_1/d(-\bar{\theta}_s)^2$ is either positive or negative, depending on whether $2\alpha_2$ is greater or less than $e^{(2\nu-1)\Psi}$. In the first case, since $d\alpha_1/d(-\bar{\theta}_s)$ is zero at $-\bar{\theta}_s = -\bar{\theta}_s^0$, α_1 will increase with increasing $-\bar{\theta}_s$ until it reaches a maximum at the value of $-\bar{\theta}_s$ that corresponds to the single root of (53). This argument holds for all values of ν since, at $-\bar{\theta}_s = -\bar{\theta}_s^0$, the left-hand side of (53) exceeds the value of the right-hand side, leading to a single intersection when the linear left-hand and exponential right-hand sides of (53) are plotted against $-\bar{\theta}_s$. Consequently, for $2\alpha_2 > e^{(2\nu-1)\Psi}$, the qualitative behavior will be identical to Figure 4(a), indicating an extension of the extinction limit relative to the noncatalytic case. On the other hand, for $2\alpha_2 < e^{(2\nu-1)\Psi}$, we have that $d^2\alpha_1/d(-\bar{\theta}_s)^2$ is negative at $-\theta_1 = -\bar{\theta}_s^0$ and hence α_1 decreases as $-\bar{\theta}_s$ increases from that value. The qualitative nature of the solution response then depends on whether $\nu \geq 1/2$ or $\nu < 1/2$.

If, for the case $2\alpha_2 < e^{(2\nu-1)\Psi}$, we have $\nu \geq 1/2$, the right-hand side of (53) is either exponentially increasing or constant, whereas the left-hand side is a linearly decreasing function of $-\bar{\theta}_s$. In this instance, the left-hand side of (53) is less than the value of the right-hand side at $-\bar{\theta}_s = -\bar{\theta}_s^0$ and there are no physical roots of (53). The solution response is then qualitatively similar to Figure 4(b) and there is no catalytic extension of the extinction limit. This situation persists as ν decreases below the value $1/2$ (at which point the right-hand side of (53) transitions from a growing to a decaying exponential) until at some point the linearly decreasing left-hand side of (53) intersects the exponentially-decaying right-hand side tangentially in at first one, and then two, places. The first of these roots, if it occurs for $-\bar{\theta}_s > -\bar{\theta}_s^0$, then corresponds to a relative minimum in the $\alpha_1(-\bar{\theta}_s)$ response, while the second corresponds to a relative maximum and hence an extension of the extinction limit, provided that this root occurs in the physical range $-\bar{\theta}_s > -\bar{\theta}_s^0$ and the relative maximum value of α_1 exceeds the value at $-\bar{\theta}_s^0$. In the two-root case just described, the solution response is triple-valued for a range of α_1 values, corresponding to low-, intermediate- and high-temperature solution branches.

The various scenarios just described for $\nu < 1/2$, which illustrate how different effects can counterbalance one another, are illustrated in Figures 5(a)–(c). In particular, the $\Psi = 0$ curve in Figure 5(a) demonstrates both the aforementioned relative minimum and maximum for $2\alpha_2 < e^{(2\nu-1)\Psi}$, while the other two curves for $\Psi > 0$ exhibit only the relative maximum in the physical range $-\bar{\theta}_s > -\bar{\theta}_s^0$ as the increase in Ψ eventually leads to the parameter regime $2\alpha_2 > e^{(2\nu-1)\Psi}$. In Figure 5(b), which

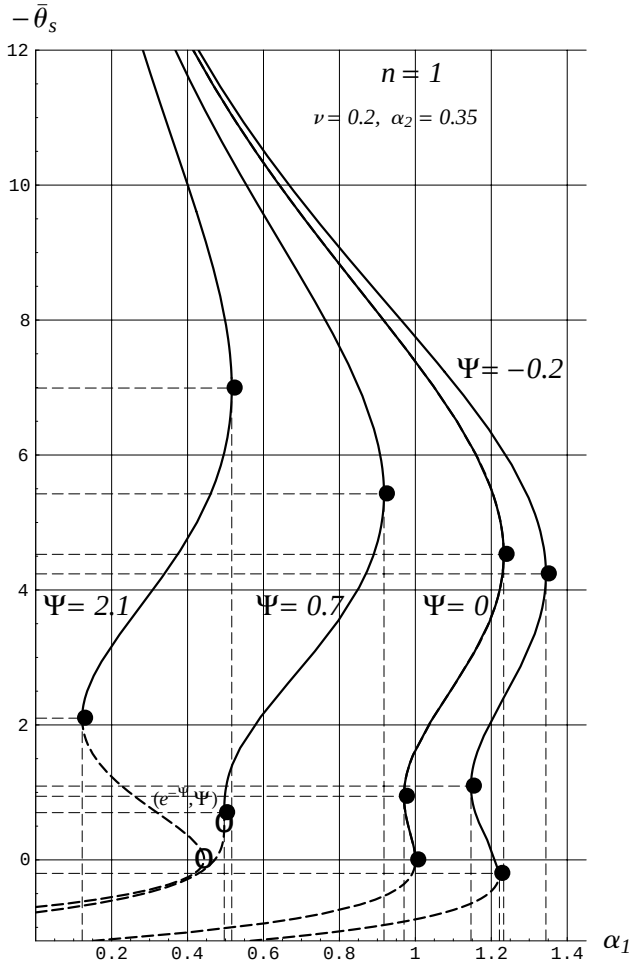


FIG. 5(a). *Solution response for $\nu < 1/2$ and several values of Ψ: $\nu = 0.2$ and $\alpha_2 = 0.35$. Larger values of α_2 and smaller values of ν have a tendency to extend the extinction limit and can thus compensate for the extinguishing effects of larger heat losses.*

is qualitatively similar to Figure 5(a), the surface activation-energy parameter ν has been decreased further with respect to its previous value, leading to a greater catalytic effect and a consequently greater extension of the extinction limit. This same effect is achieved by increasing the surface reaction-rate parameter α_2 to the value used in Figure 5(c), where there is now no relative minimum in any of the solution responses since the value of α_2 is now sufficiently large that $2\alpha_2 > e^{(2\nu-1)\Psi}$. Hence, decreasing ν and increasing α_2 have the same qualitative effect on the extension of the extinction limit. With respect to the heat-loss parameters h and k, we note from the definition (47) that the extinction limit corresponds to larger values of these parameters for smaller Lewis numbers. Thus, for example, flames whose Lewis numbers are less than unity ($l < 0$) can tolerate larger thermal losses than those whose Lewis numbers are greater then unity ($l > 0$).

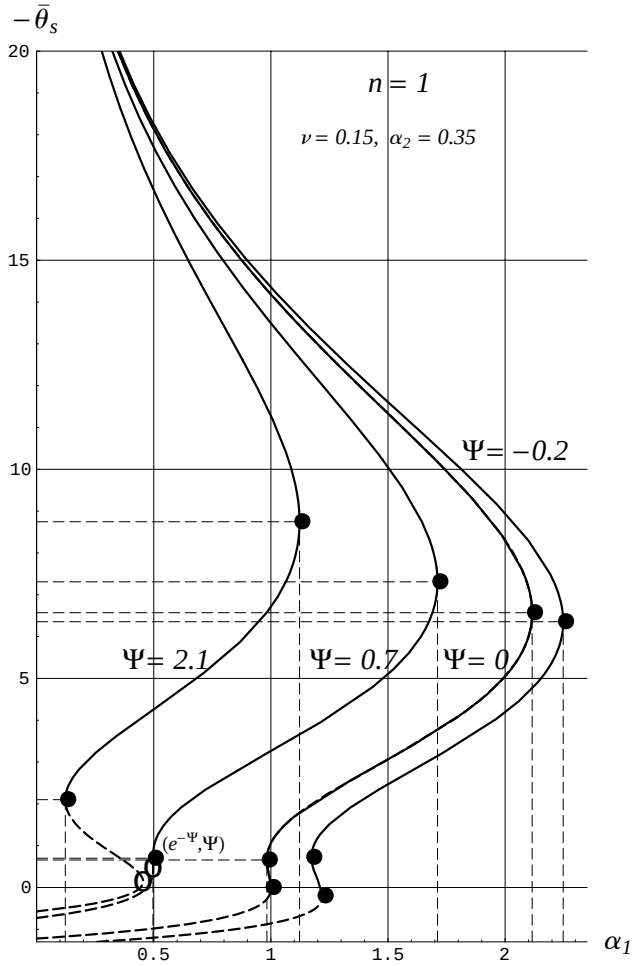


FIG. 5(b). *Solution response for $\nu < 1/2$ and several values of Ψ: $\nu = 0.15$ and $\alpha_2 = 0.35$. Larger values of α_2 and smaller values of ν have a tendency to extend the extinction limit and can thus compensate for the extinguishing effects of larger heat losses.*

6. Stability of the basic solution. The possibility of multiple steady plane-flame solutions, as illustrated in Figures 4(a) and 5(a)–(c), suggests the need for a stability analysis to determine which portions of the solution response correspond to stable combustion states, as well as to indicate the possible existence of additional nonsteady and/or nonplanar solutions. Accordingly, we introduce perturbations u, v, and τ about the basic state $\bar{\Theta}$, $\bar{S}$, and $\bar{\theta}_s$ according to

$$\Theta_0 = \bar{\Theta}(z) + u(r, \vartheta, z, t), \quad S_1 = \bar{S}(z) + v(r, \vartheta, z, t), \quad \theta_s = \bar{\theta}_s + \tau(r, \vartheta, t), \tag{55}$$

where $\bar{\Theta}$ and $\bar{S}$ are given by (40), (43), and (44), and $\bar{\theta}_s$ is determined from (46) and (47). Substituting these definitions into the closed problem given by (17) and the first equation of (10), subject to (11), (18), (29), and (35), we obtain, after linearizing

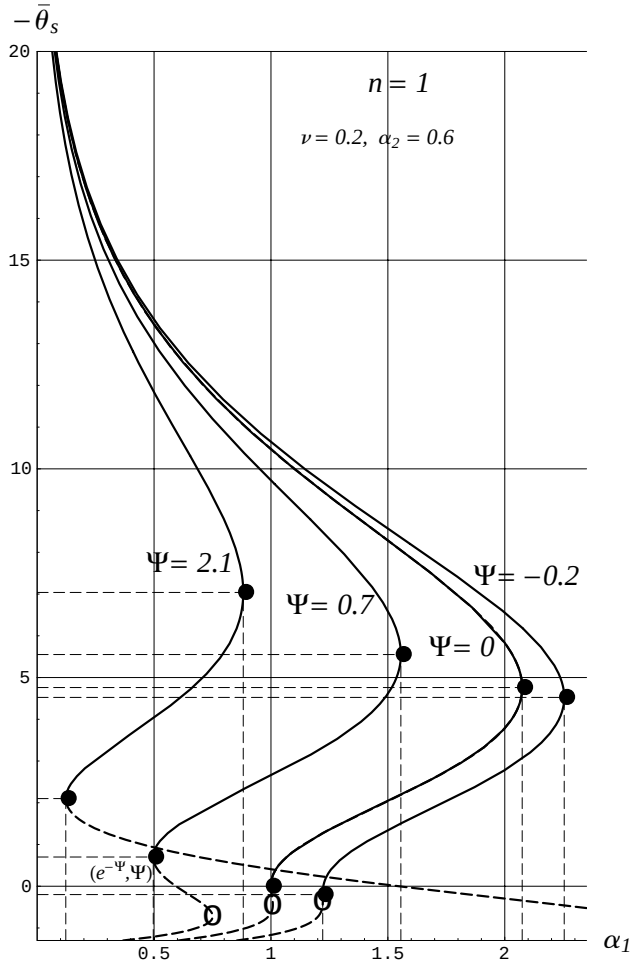


FIG. 5(c). *Solution response for $\nu < 1/2$ and several values of Ψ: $\nu = 0.2$ and $\alpha_2 = 0.6$. Larger values of α_2 and smaller values of ν have a tendency to extend the extinction limit and can thus compensate for the extinguishing effects of larger heat losses.*

about the basic solution, the linear stability problem

$$\begin{aligned} &\frac{\partial u}{\partial t} + r\frac{\partial u}{\partial r} - 2z\frac{\partial u}{\partial z} = \nabla^2 u, \\ &\frac{\partial v}{\partial t} + r\frac{\partial v}{\partial r} - 2z\frac{\partial v}{\partial z} = \nabla^2 v + l\nabla^2 u - hu, \quad 0 < z < \infty, \end{aligned} \tag{56}$$

$$u\big|_{z=0} = 0, \quad \left.\frac{\partial v}{\partial z}\right|_{z=0} = -l\left.\frac{\partial u}{\partial z}\right|_{z=0}, \qquad u,\ v \to 0 \ \text{ as } \ z \to \infty, \tag{57}$$

$$\begin{aligned} -\sqrt{\pi}\,\alpha_1 \left.\frac{\partial u}{\partial z}\right|_{z=0} &= \left[v\big|_{z=0} + \tau(-\Psi - \bar{\theta}_s)\right] e^{\bar{\theta}_s} + 2\alpha_2(-\Psi - \bar{\theta}_s) \\ &\quad \times \ \left[v\big|_{z=0} + \ \nu\tau(-\Psi - \bar{\theta}_s - \nu^{-1})\right] e^{2\nu\bar{\theta}_s}. \end{aligned} \tag{58}$$

In obtaining (58), the reaction order n has been taken to be unity, and the preceding results were used to evaluate $\bar{S}(0) = -\Psi$, $d\bar{\Theta}/dz\big|_{z=0} = -2/\sqrt{\pi}$ and, from (30), to calculate

$$\begin{aligned} G_1\big(\theta_s; S_1\big|_{z=0}\big) &= \big(1 + S_1\big|_{z=0} - \theta_s\big)\, e^{\theta_s} \\ &= \big(1 - \Psi + v\big|_{z=0} - \bar{\theta}_s - \tau\big)\, e^{\bar{\theta}_s + \tau} \\ &\sim \big(1 - \Psi - \bar{\theta}_s\big)\, e^{\bar{\theta}_s} + \big(1 - \Psi - \bar{\theta}_s\big)\, e^{\bar{\theta}_s}\tau \\ &\quad + \big(v\big|_{z=0} - \tau\big)\, e^{\bar{\theta}_s} + \cdots . \end{aligned} \tag{59}$$

An investigation of the linear stability of the basic solution may now proceed by seeking solutions of (56)–(58) proportional to $e^{i(\omega t \pm m\theta)}$ times appropriate functions of kr and z, where $i\omega$ is the complex growth rate and m and k are the angular and radial wavenumbers, respectively. However, because it is reasonable to assume a high conductivity for the catalytic surface relative to the gas, which would consequently not support a nonuniform temperature distribution at $z = 0$, we may, in that case, set the radial and angular wavenumbers to zero. That is, only planar perturbations of the basic state are admissible, and harmonic solutions are thus restricted to the form

$$u = f(z)\, e^{i\omega t}, \quad v = g(z)\, e^{i\omega t}, \quad \tau = c_1 e^{i\omega t}. \tag{60}$$

Here, choosing the preexponential coefficient to be unity in the last expression would normalize any nontrivial solution to the linear stability problem, with stability of the basic state determined by the sign of the real part of $i\omega$. Substituting (60) into (56)–(58) thus gives a closed problem for the complex unknowns $f(z)$, $g(z)$, and $i\omega$ as

$$i\omega f - 2z\frac{df}{dz} = \frac{d^2 f}{dz^2}, \quad i\omega g - 2z\frac{dg}{dz} = \frac{d^2 g}{dz^2} + l\frac{d^2 f}{dz^2} - hf, \qquad 0 < z < \infty, \tag{61}$$

$$f(0) = 0, \quad \frac{dg}{dz}\bigg|_{z=0} = -l\frac{df}{dz}\bigg|_{z=0}, \qquad f,\ g \to 0 \ \text{ as } \ z \to \infty, \tag{62}$$

$$\begin{aligned} -\sqrt{\pi}\,\alpha_1 \frac{df}{dz}\bigg|_{z=0} &= \big[g(0) + c_1(-\Psi - \bar{\theta}_s)\big]\, e^{\bar{\theta}_s} + 2\alpha_2(-\Psi - \bar{\theta}_s) \\ &\quad \times \ \big[g(0) + \nu c_1(-\Psi - \bar{\theta}_s - \nu^{-1})\big]\, e^{2\nu\bar{\theta}_s}. \end{aligned} \tag{63}$$

We proceed with an analysis of (61)–(63) by noting that the subproblems for f and g may be solved sequentially in terms of $i\omega$. We also consider only solutions for which $\Re(i\omega) \geq 0$, since solutions with $\Re(i\omega) < 0$ correspond to decaying perturbations and hence stability of the basic solution.

It is readily verified that the transformation $x = -z^2$ converts the first of (61) into the standard form of the confluent hypergeometric equation $x\, d^2f/dx^2 + (b - x)\, df/dx - af = 0$ with $b = 1/2$ and $a = -i\omega/4$. For noninteger values of b, two independent solutions for f are therefore ${}_1F_1(a; b; x)$ and $x^{1-b}{}_1F_1(1 + a - b; 2 - b; x)$, where ${}_1F_1(a; b; x) = \sum_{n=0}^{\infty}[(a)_n/(b)_n]x^n/n! = \big\{\Gamma(b)/[\Gamma(b-a)\Gamma(a)]\big\} \int_0^1 e^{xt}t^{a-1} \times (1 - t)^{b-a-1}dt$ is the confluent hypergeometric (Kummer) function (cf. Slater [13]). Thus, in terms of z, the general solution of the first equality in (61) may be written as

$$f(z) = C_1 \cdot {}_1F_1\left(-\frac{i\omega}{4}; \frac{1}{2}; -z^2\right) + C_2\, z \cdot {}_1F_1\left(\frac{1}{2} - \frac{i\omega}{4}; \frac{3}{2}; -z^2\right), \tag{64}$$

where C_1 and C_2 are constants of integration. Requiring that $f(0) = 0$ thus implies $C_1 = 0$, while the boundary condition $f \to 0$ as $z \to \infty$ requires, when $\Re(i\omega) \geq 0$, that $C_2 = 0$ as well, where the latter is deduced from the asymptotic behavior ${}_1F_1(a; b; x) \sim [\Gamma(b)/\Gamma(b-a)](-x)^{-a}[1 + \mathrm{O}(|x|^{-1})]$ as $|x| \to \infty$ for $\Re(x) < 0$ (cf. [13]). Hence, $f(z) \equiv 0$ so that no nontrivial solution for f corresponding to neutral or growing perturbations is admissible.

A similar argument can be applied to the determination of $g(z)$. In particular, because $f = 0$, the second equation of (61) for g is identical to the equation for f, and thus the general solution for $g(z)$ is also given by the right-hand side of (64). In this case, the boundary condition $\partial g/\partial z = l\partial f/\partial z = 0$ at $z = 0$ implies that $C_2 = 0$ in the corresponding general solution for g, while the condition that g vanish as $z \to \infty$ requires, based on the above asymptotic behavior of ${}_1F_1$, that $C_1 = 0$ as well. Thus, for $\Re(i\omega) \geq 0$, both $f(z)$ and $g(z)$ are identically zero. Equation (63) then reduces to

$$c_1(-\Psi - \bar{\theta}_s)\left[1 + 2\alpha_2\nu(-\Psi - \bar{\theta}_s - \nu^{-1})\, e^{(1-2\nu)(-\bar{\theta}_s)}\right] = 0, \tag{65}$$

which is generally satisfied only if $c_1 = 0$. An exception occurs at values of $-\bar{\theta}_s$ for which $d\alpha_1/d(-\bar{\theta}_s) = 0$ (i.e., at the relative extrema in Figures 3–5), in which case, based on (52) and (53), c_1 is indeterminate. However, all other points on the basic solution response correspond to either decaying harmonic perturbations or to the absence of a nontrivial solution to the linear stability problem. We thus conclude that, at least in the classical sense, all branches of the basic solution are linearly stable. Consequently, in parameter regimes where multiple solutions exist, the observed solution is likely to depend on the initial conditions, which appears to be the case in actual experiments [6]. We remark, however, that relaxing the assumption of high surface conductivity so as to allow for transverse perturbations might, as in the case of strictly gaseous flames in stagnation-point flow, permit instability for sufficiently small values of the strain-rate parameter (cf. Sivashinsky, Law, and Joulin [14]).

7. Conclusion. The present work has presented a formal asymptotic model of a nonadiabatic catalytic flame in stagnation-point flow. The thermal/diffusive model, which was derived under the assumptions of large activation energies and near-unity Lewis numbers, considers both surface and volumetric heat losses, where the former occurs at the catalytic surface and the latter approximates conductive and/or radiative losses across the remaining surfaces. Assuming combustion to occur in the near-surface region where catalytic effects are felt, it was shown that the presence of a catalytic surface has the potential to significantly extend the extinction limits arising from the effects of flame stretch and heat loss. In particular, reactants that leak through the distributed portion of the gas flame, due to larger strain rates and/or larger rates of heat loss that lower the reaction rate, have an additional opportunity to react under the influence of a catalyst at the surface. Such an influence is particularly desirable from the standpoint of building small combustors that have relatively large surface-to-volume ratios, and the present work has therefore focused on further extending earlier studies to the nonadiabatic regime. Indeed, recent experiments indicate that microcombustors on the order of a cubic millimeter or less are feasible when the stagnation surface is coated with a catalyst.

The solution response of the model problem is parameterized by the nondimensional strain rate, the ratios of the surface reaction rate and activation energy to those of the distributed reaction in the bulk gas, and a parameter that represents a linear combination of the effects of surface and volumetric heat losses and those associated with nonunity Lewis numbers. The main results demonstrate how, in certain

parameter regimes associated with a strongly catalytic effect, the solution response is modified from that of the noncatalytic problem to allow for larger values of the strain rate, rate of heat loss, and/or Lewis number than would be the case in the absence of catalysis. In such regimes, the solution response exhibits an extinction limit at a value of the strain rate (or heat-loss coefficient) that is larger than the corresponding value in the absence of catalysis, resulting in a catalytic extension of the extinction limit. In addition, multiple solution branches appear, corresponding to high, low, and, in some cases, intermediate surface temperatures. A linear stability analysis suggests, consistent with recent experiments, that in the limit of high surface conductivity each branch is locally stable so that it should be possible to observe different solutions depending on the initial conditions.

REFERENCES

[1] C. K. Law and G. I. Sivashinsky, *Catalytic extension of extinction limits of stretched premixed flames,* Combust. Sci. Technol., 29 (1982), pp. 277–286.

[2] V. Giovangigli and S. Candel, *Extinction limits of premixed catalyzed flames in stagnation point flows,* Combust. Sci. Technol., 48 (1986), pp. 1–30.

[3] J. Warnatz, M. D. Allendorf, R. J. Kee, and M. E. Coltrin, *A model of elementary chemistry and fluid mechanics in the combustion of hydrogen on platinum surfaces,* Combust. Flame, 96 (1994), pp. 393–406.

[4] C. K. Law, S. Ishizuka, and M. Mizomoto, *Lean-limit extinction of propane/air mixtures in the stagnation-point flow,* in Proceedings of the Combustion Institute, 18 (1981), pp. 1791–1798.

[5] H. Ikeda, J. Sato, and F. A. Williams, *Surface kinetics for catalytic combustion of hydrogen-air mixtures on platinum at atmospheric pressure in stagnation flows,* Surface Science, 326 (1995), pp. 11–26.

[6] D. L. Mowery, T. J. Gardner, G. C. Frye-Mason, R. Kottenstette, R. P. Manginell, and S. B. Margolis, *Development of a novel on-chip catalytic microcombustor device,* in Proceedings of the 17th North American Catalysis Society Meeting, Toronto, 2001.

[7] S. B. Margolis and T. J. Gardner, *Extinction limits of nonadiabatic, catalyst-assisted flames in stagnation-point flow,* Combust. Theory Model., 6 (2002), pp. 19–34.

[8] B. J. Matkowsky and D. O. Olagunju, *Pulsations in a burner-stabilized premixed plane flame,* SIAM J. Appl. Math., 40 (1981), pp. 551–562.

[9] M. R. Booty, S. B. Margolis, and B. J. Matkowsky, *Interaction of pulsating and spinning waves in nonadiabatic flame propagation,* SIAM J. Appl. Math., 47 (1987), pp. 1241–1286.

[10] H. G. Kaper, G. K. Leaf, S. B. Margolis, and B. J. Matkowsky, *On nonadiabatic condensed phase combustion,* Combustion Science and Technology, 53 (1987), pp. 289–314.

[11] S. B. Margolis and S. C. Johnston, *Multiplicity and stability of supercritical combustion in a nonadiabatic tubular reactor,* Combustion Science and Technology, 65 (1989), pp. 103–136.

[12] B. J. Matkowsky and G. I. Shivashinsky, *An asymptotic derivation of two models in flame theory associated with the constant density approximation*, SIAM J. Appl. Math., 37 (1979), pp. 686-699.

[13] L. J. Slater, *Confluent Hypergeometric Functions*, Cambridge University Press, London, 1960.

[14] G. I. Sivashinsky, C. K. Law, and G. Joulin, *On stability of premixed flames in stagnation-point flow,* Combust. Sci. Technol., 28 (1982), pp. 155–159.

SIAM J. APPL. MATH.
Vol. 63, No. 4, pp. 1105–1120

RADIAL STRUCTURE OF TRAVELING WAVES IN THE INNER EAR*

HONGXUE CAI† AND RICHARD CHADWICK†

Abstract. We develop a hybrid approach for modeling the cochlea, in which we let the WKB method determine the axial propagation of waves and restrict the numerics to transverse planes, where we solve a fluid-solid interaction eigenvalue problem. The cochlear fluid is treated as viscous and incompressible. Viscous effects are confined to oscillatory boundary layers and the thin gap between the reticular lamina (RL) and the lower surface of the tectorial membrane (TM). Our model includes axial fluid coupling and also axial elastic coupling via a basilar membrane (BM) modeled as an orthotropic clamped plate. Three-dimensional (3D) flow is solved in two-dimensional (2D) domains, with interactions with 2D elastic domains representing the organ of Corti (OC) and the TM. The OC contains inhomogeneities representing discrete cellular structures. We have computed the interaction between the BM, TM, OC, and the cochlear fluid to find the complex-valued wavenumber-frequency relation and vibrational modes. The details of the cochlear fluid flow and pressure fields are calculated, along with displacements of the elastic structures. Simulation of passive radial modes in the apical region of a guinea pig cochlea for frequencies less than 1 kHz indicates monophasic vibration of the BM and a synchronous rotation of three rows of outer hair cell stereocilia induced by a shearing motion between the RL and TM.

Key words. cochlear mechanics, traveling wave, fluid-solid interaction, radial modes

AMS subject classifications. 74F10, 92C10, 92C35

PII. S0036139901388957

1. Introduction.

Modeling the radial structure of traveling waves in the inner ear is timely because of new experimental observations that require interpretation. Experimental observations can now resolve the vibratory patterns across the width of the organ of Corti (OC). Both monophasic and multiphasic radial modal patterns have been reported, depending on the method of stimulation, axial location, and frequency [11, 16, 25, 26, 29]. These patterns, representing the radial structure of the traveling wave, are critical to understanding the stimulation of inner and outer hair cells (IHCs and OHCs). An excellent recent review of experimental cochlear mechanics can be found in Robles and Ruggero [31]. A reliable computational model would be very useful in sorting out the variety of reported responses. Here we begin by developing a passive model.

The detailed cochlear fluid flow and micromechanical movements of the tectorial membrane (TM) and cellular structures relative to one another within the OC have interested many auditory researchers [1, 3, 4, 12, 22, 24, 39, 40]. Most models in the literature use simple lumped elements (lever, spring, damper, transformer, etc.). While these lumped-parameter models have been very useful, they oversimplify the dynamics by neglecting fluid mass coupling in the micromechanics. They also make assumptions concerning the kinematics that can now be computed rather than as-

*Received by the editors May 7, 2001; accepted for publication (in revised form) October 2, 2002; published electronically March 26, 2003.

http://www.siam.org/journals/siap/63-4/38895.html

†Laboratory of Cellular Biology, Section on Auditory Mechanics/NIDCD, National Institutes of Health, Bethesda, MD 20892 (hongxuec@helix.nih.gov, chadwick@helix.nih.gov).

sumed. Toward this end we develop a *physically realistic* OC geometry and model the OC as an elastic body whose *inhomogeneities* are due to discrete cellular elements including IHCs, OHCs, Deiter's cells (DCs), pillar cells (PCs), and the reticular lamina (RL). Accordingly, the OC has a material property map with each subdomain having the Young's modulus (E) and Poisson's ratio (ϑ) of the corresponding cell structures (see Appendix B).

The excessive computational cost of incorporating radial structure and fluid-solid interaction into models of traveling waves complicates the study of cochlear mechanics. Straightened models with three-dimensional (3D) flow including radial variation of the traveling wave have been studied using the WKB method [8, 17, 33, 34] and by fully numerical methods [19, 28]. Manoussaki and Chadwick [23] developed a hybrid approach to study the effects of cochlear curvature. But these types of models considered the basilar membrane (BM) only and did not attempt to consider the complications of other true degrees of freedom in the cochlear partition (CP). Actually there are very few calculations for which the radial variation of BM transverse deflection is computed and not assumed. Other modeling efforts [7, 20, 38] used full 3D finite elements or finite differences, with some critique offered by Steele [32]. In his hybrid 3D model of the cochlea, which includes structural details of the OC modeled as orthotropic shell elements and discrete fluid domains modeled as equivalent rectangles, Steele reported different propagation modes in different fluid domains. Here we develop a hybrid approach that provides a single propagating mode with a single wavenumber applicable to the entire cross section. Like Steele [32], we let the WKB method deal with the axial propagation of the wave, and we restrict the numerics to the transverse plane. We use finite elements for both fluid and solid components, and we compute fluid-solid interaction in a physically realistic complex geometry of the cochlear cross section. Modes and displacements of the OC are analyzed, and the details of the cochlear fluid flow are calculated. This approach avoids a full 3D computation and allows for greater resolution in a cochlear cross section. In our cochlear model, we carry out only the first step in the reduction of the 3D hydroelastic problem to a sequence of eigenvalue problems in transverse planes. The WKB-numerical hybrid approach allows this reduction and provides the formalism for connecting the solution in different transverse planes via an energy transport equation. That part of the solution is beyond the scope of the present work.

2. Model description. The coordinates (X, Y, Z) respectively denote the "radial," "transverse," and "axial" (along the duct length) directions. The transverse plane of the cochlea is divided into fluid and elastic domains (Figure 2.1). Discrete structural elements are embedded in the two-dimensional (2D) continuum of the OC. The TM and OC solid domains are coupled by cochlear fluid and OHC stereocilia. The OC rests on an orthotropic clamped plate that represents the BM, whose axial coupling is considered via the plate Green's function (see section 3 for details). The axial fluid flow and coupling are also retained by WKB expansion on all fluid domains: scala tympani (ST) and the combined scala media and scala vestibuli (SM+SV). The cross section of the cochlea is bounded by rigid walls, represented by circular arcs and straight segments in Figure 2.1.

2.1. Fluid domains. The fluid velocity $\tilde{\mathbf{V}} = (\tilde{V_X}, \tilde{V_Y}, \tilde{V_Z})$ and pressure $\tilde{P}$ satisfy the linearized Navier–Stokes and mass conservation equations (see [5])

$$\rho \frac{\partial \tilde{\mathbf{V}}}{\partial t} = -\nabla \tilde{P} + \mu \nabla^2 \tilde{\mathbf{V}}, \tag{2.1}$$

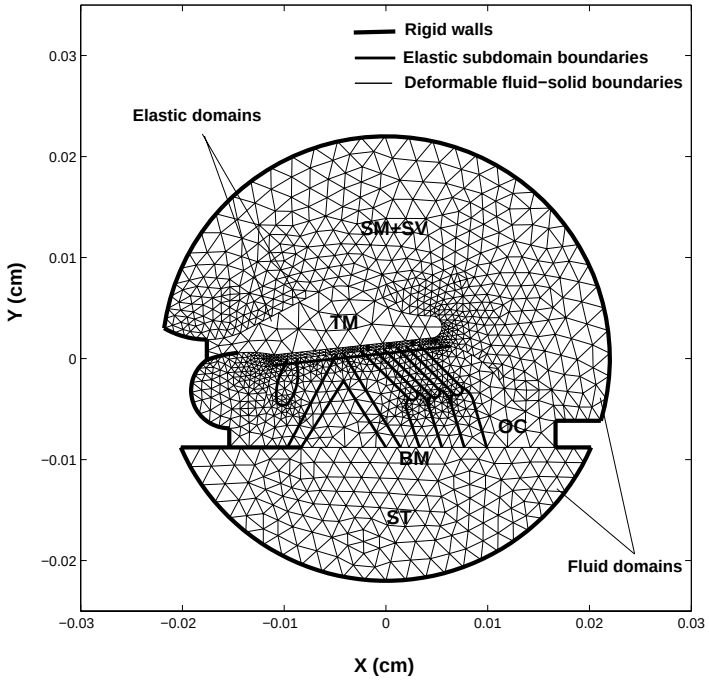


FIG. 2.1. *Model representation of a cochlear cross section. Tectorial membrane (TM) and organ of Corti (OC) are represented by 2D elastic domains. The former is homogeneous, while the latter contains different subdomains representing discrete cellular structures. The OC has the reticular lamina (RL, not labeled) as its top boundary and rests on the basilar membrane (BM), which is represented by an orthotropic clamped plate. A narrow fluid-filled gap exists between the RL and the lower surface of the TM. Stereocilia elastically couple the RL and TM. The scala tympani (ST) is the fluid compartment on the lower side, while the upper fluid region is the combined scala media and scala vestibuli (SM+SV). (For the present calculation without the Reissner's membrane, the scala media and scala vestibuli contain the same fluid.) Scale is in cm.*

$$\nabla \cdot \tilde{\mathbf{V}} = 0, \tag{2.2}$$

where ∇ is the gradient operator, ∇^2 is the Laplacian operator, and ρ and μ are the density and viscosity of the cochlear fluid. From (2.1) and (2.2), it follows that the pressure is harmonic:

$$\nabla^2 \tilde{P} = 0. \tag{2.3}$$

Boundary conditions for pressure can be obtained for a viscous fluid following Holmes and Cole [17]. For thin oscillatory boundary layers the dominant boundary conditions are those for an inviscid fluid,

$$\mathbf{n} \cdot \nabla \tilde{P} = 0 \tag{2.4}$$

on rigid walls, where $\mathbf{n}$ is the outward unit normal. Since $\tilde{\mathbf{V}} = \partial \tilde{\mathbf{U}}/\partial t$,

$$\mathbf{n} \cdot \nabla \tilde{P} = -\rho \frac{\partial \tilde{\mathbf{V}}_{\mathbf{n}}}{\partial t} = -\rho \frac{\partial^2 \tilde{\mathbf{U}}_{\mathbf{n}}}{\partial t^2} \tag{2.5}$$

on deformable boundaries, where $\tilde{\mathbf{U}} = (\tilde{U}_X, \tilde{U}_Y, 0)$ is the displacement vector of a deformable boundary.

2.2. Solid domains. We solved a structural mechanics plane strain problem for the TM and OC solid domains. The stress-strain relation can be written, assuming isotropic and isothermal conditions (see [35]), as

$$\begin{pmatrix} \tilde{\sigma}_X \\ \tilde{\sigma}_Y \\ \tilde{\tau}_{XY} \end{pmatrix} = \frac{E}{(1+\vartheta)(1-2\vartheta)} \begin{pmatrix} 1-\vartheta & \vartheta & 0 \\ \vartheta & 1-\vartheta & 0 \\ 0 & 0 & (1-2\vartheta)/2 \end{pmatrix} \begin{pmatrix} \tilde{\epsilon}_X \\ \tilde{\epsilon}_Y \\ \tilde{\gamma}_{XY} \end{pmatrix}, \tag{2.6}$$

where $\tilde{\sigma}_X$ and $\tilde{\sigma}_Y$ are the stresses in the X and Y directions and $\tilde{\tau}_{XY}$ is the shear stress. The material properties are expressed as a combination of E (Young's modulus) and ϑ (Poisson's ratio). The strains are defined as

$$\tilde{\epsilon}_X = \frac{\partial \tilde{U}_X}{\partial X}, \quad \tilde{\epsilon}_Y = \frac{\partial \tilde{U}_Y}{\partial Y}, \quad \tilde{\gamma}_{XY} = \frac{\partial \tilde{U}_X}{\partial Y} + \frac{\partial \tilde{U}_Y}{\partial X}. \tag{2.7}$$

For the vibratory 2D elastic domains, the equation of motion, neglecting gravity, is given by

$$-\nabla \cdot \tilde{\sigma} + \rho_s \frac{\partial^2 \tilde{\mathbf{U}}}{\partial t^2} = 0, \tag{2.8}$$

where ρ_s is the density of the solid domains. Combining (2.6), (2.7), and (2.8), we can arrive at a PDE system involving the displacements. The Young's modulus and Poisson's ratio are incorporated into the PDE coefficient matrices; thus we can handle the inhomogeneities of the OC by discretizing its mechanical properties (E and ϑ) on corresponding subdomains (see Appendix B).

There are four types of boundary conditions on the 2D solid domains. Boundary segments contiguous with rigid domains are subject to a homogeneous Dirichlet condition with zero displacements: $\tilde{U}_X = \tilde{U}_Y = 0$. The OC boundary segment contiguous with the BM is an inhomogeneous Dirichlet condition: $\tilde{U}_X = 0$ and $\tilde{U}_Y = \tilde{Y}_b$, where $\tilde{Y}_b$ is the BM displacement. $\tilde{Y}_b$ can be calculated via a plate Green's function integral (see section 3). Deformable boundary segments in contact with fluid are subject to the stress vector $\tilde{\sigma}_{\mathbf{n}}$:

$$\tilde{\sigma}_{\mathbf{n}} = \tilde{\sigma} \cdot \mathbf{n} = -\tilde{P}\mathbf{n} + \tilde{\tau}\mathbf{s}, \tag{2.9}$$

where $\mathbf{s}$ is the unit tangential vector in the transverse plane and $\tilde{\tau}$ is the tangential surface traction due to an oscillatory boundary layer. The stress vector $\tilde{\sigma}_{\mathbf{n}}$ can be expressed in terms of solid displacements via (2.6) and (2.7), so that (2.9) is an inhomogeneous Neumann condition; this will be discussed further in section 2.3. Finally, the thin gap region bounded by the lower surface of the TM and the upper surface of the RL is given special treatment in the present model. Here we adopt the analysis of Chadwick, Dimitriadis, and Iwasa [9], who showed using lubrication theory that there is no squeezing of the gap, but only a relative tangential motion between the TM and RL, as was originally proposed by Allen [1, 3]. In the present context this leads to the inhomogeneous Dirichlet condition $\tilde{U}_{TM,n} = \tilde{U}_{RL,n}$.

2.3. WKB expansion. The axial coordinate is normalized by the cochlear length L ($z = Z/L$), and the other two coordinates are normalized by a characteristic cross-sectional radius R_0 ($x = X/R_0$, $y = Y/R_0$). We express all dependent variables in the form $\tilde{\Phi}(x, y, z, t) = \Phi(x, y, z)\exp[i(\omega t - \epsilon^{-1} \int k(z)dz)]$, where k is the dimensionless complex-valued wavenumber normalized by R_0 (i.e., $k = \tilde{k}R_0$, where

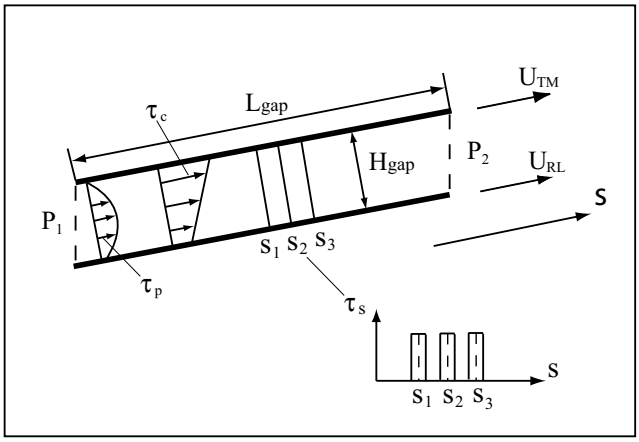


FIG. 2.2. *RL-TM gap tangential stresses. τ_p is the Poiseuille flow-induced tangential stress, τ_c is the Couette flow-induced tangential stress, and τ_s is the tangential stress contributed by hair bundle stiffness.*

$\tilde{k}$ is the dimensional wavenumber), $\epsilon = R_0/L_c \ll 1$, and ω is the radian frequency. The slowly varying wave approximation requires the expansion of the amplitude Φ of each dependent variable: $\Phi = \Phi_0 + \epsilon\Phi_1 + \cdots$. The dominant equation of the WKB expansion of (2.3) gives us

$$\nabla_T^2 P_0 = k^2 P_0, \tag{2.10}$$

where the Laplacian operator is defined in the normalized transverse plane: $\nabla_T^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2$. The WKB expansion of (2.4) and (2.5) gives the following dominant boundary condition for the fluid domains:

$$\frac{\partial P_0}{\partial n_T} = \rho\omega^2 R_0 U_{n_T 0}, \tag{2.11}$$

where the fluid-solid interface normal displacement $U_{n_T 0} = \mathbf{U_0} \cdot \mathbf{n_T}$. After solving the inviscid problem, the pressure acting on a deformable boundary can be corrected for viscous effects in an oscillatory boundary layer [5] by adding a small correction term to the boundary pressure [9]: $P_c = \rho\omega^2 U_{n_T 0}\sqrt{-i\nu/\omega}$, where $\nu = \mu/\rho$ is the kinematic viscosity.

The dominant WKB expansion of (2.8) is

$$\nabla_T \cdot \sigma_{\mathbf{0}} + R_0\rho_s\omega^2 \mathbf{U_0} = 0, \tag{2.12}$$

with boundary condition at a deformable surface

$$\sigma_{n_T 0} = \sigma_{\mathbf{0}} \cdot \mathbf{n_T} = -P_0 \mathbf{n_T} + \tau \mathbf{s}. \tag{2.13}$$

Solid damping is included in the model by assuming a Voigt solid and replacing E by $E + i\omega E'$, where E' is a damping parameter [19]. For the TM-RL gap boundary segments, $\tau = \tau_p + \tau_c + \tau_s$ (see Figure 2.2), where $\tau_p = \frac{1}{2}\Delta P_0 H_{gap}/L_{gap}$ is the Poiseuille flow-induced tangential stress due to the pressure difference ΔP_0 between the two ends of the gap of length L_{gap} and thickness H_{gap}, $\tau_c = i\mu\omega\Delta U_{s0}/H_{gap}$ is

the Couette flow-induced tangential stress due to the difference between the TM and RL tangential displacements ΔU_{s0}, and τ_s is the tangential stress contributed by the hair bundle stiffness (K_s). We model τ_s as a rectangular wave over the normalized OHC diameter with amplitude $K_s \Delta U_{s0}/A_s$, where A_s is the effective area of stress ($15 \times 15 \mu m^2$) of the hair bundle. Couette and Poiseuille flow in the gap was proposed by Allen [1, 3]. For the nongap boundary segments, $\tau = \sqrt{i\omega\nu}(\frac{1}{i\omega R_0}\frac{\partial P_0}{\partial s} + i\rho\omega U_{s0})$, where U_{s0} is either the TM or OC segment tangential displacement (see Appendix A). All boundary conditions are linearized and applied on their undeformed locations.

3. Computation. We fix ω and solve the above eigenvalue problem for the wavenumber, fluid pressure, and solid displacement fields using the MATLAB PDE Toolbox on an SGI workstation with an R12000 processor. We define the geometry and the boundary conditions by directly editing MATLAB geometry and boundary condition matrices. Each boundary segment (straight line or arc) corresponds to a column in the geometry matrix, and each column in the geometry matrix must correspond to a column in the boundary condition matrix. A boundary condition must be expressed in a special MATLAB string representation, which contains the 2D coordinates x and y, the outward normal vector components n_x and n_y, and a normalized arc length parameter s. Since the boundary string expressions must be continuous functions, we define the boundaries using polynomial curve-fits. Meshes of the 2D domains are generated and refined by MATLAB. For clarity, the meshes shown in Figure 2.1 are coarser than those we used in our actual calculations. Our algorithm iterates between elliptic, 2D plane strain, and eigenvalue solvers in the PDE Toolbox (see Appendix B). First we guess values of the wavenumber k and of the displacements of the TM, OC, and BM, and we solve the elliptic problems in the upper and lower fluid chambers (see Figure 2.1) for the fluid pressure field P_0. Then we solve the 2D plane strain problems sequentially for the OC and TM. We calculate the updated deflection of the BM via the plate Green's function G:

$$Y_b(x) = R_0 \int_{b_1}^{b_2} G(x, x')\Delta\sigma_n(x')dx', \tag{3.1}$$

where b_1 and b_2 are the radial coordinates of the BM endpoints and $\Delta\sigma_n$ is the difference between the pressure at the lower surface of the BM and the negative of the normal stress at the lower surface of the OC. Considering the axial coupling of the BM [2, 18, 21, 36], the orthotropic plate Green's function $G(x, x')$ satisfies

$$D_x \frac{\partial^4 G}{\partial x^4} - 2(D_x D_z)^{1/2} k^2 \frac{\partial^2 G}{\partial x^2} + (-R_0^4 \rho_b H_b \omega^2 + D_z k^4)G = \delta(x - x')R_0^4, \tag{3.2}$$

with the following boundary conditions for an orthotropic clamped plate: $G_L(b_1) = G_R(b_2) = 0$, $G'_L(b_1) = G'_R(b_2) = 0$, $G_L(x') = G_R(x')$, $G'_L(x') = G'_R(x')$, $G''_L(x') = G''_R(x')$, and $G'''_L(x') - G'''_R(x') = R_0^3/D_x$, where $G_L(x)$ and $G_R(x)$ are defined as $G_L(x) = G(x, x')$ for $x < x'$ and $G_R(x) = G(x, x')$ for $x > x'$. Note that $D_x = (E_b + i\omega E_b^{'})I_b$ is the complex radial rigidity of the BM, D_z is the axial rigidity of the BM, and $\rho_b H_b$ is the mass per unit area of the BM. The plate Green's function is determined analytically in Mathematica and imported into MATLAB. More specifically, $G_R(x)$ and $G_L(x)$ are each the sum of four exponential functions with multiplicative constants. The boundary conditions lead to an 8×8 linear system, which is solved analytically.

The impedances of deformable surfaces are defined to be $Z_n = -P_0/U_{n0}$. The impedances are also determined as continuous functions using polynomial fits. This

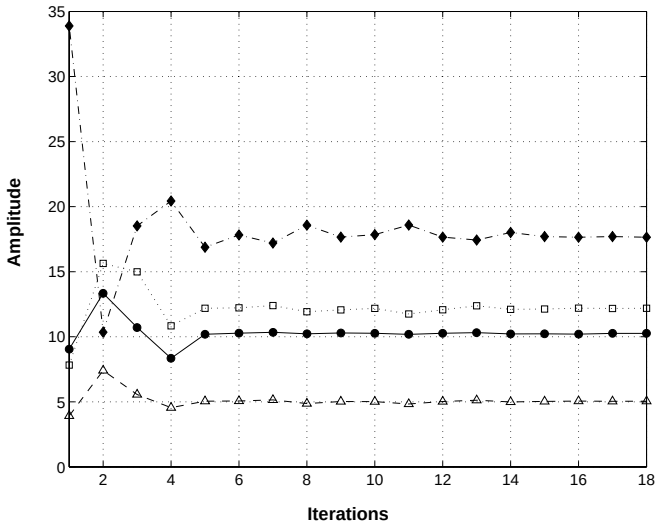


FIG. 3.1. *Convergence of the algorithm. Plots versus iteration number of the absolute values of the dimensional wavenumber* $\tilde{k}$ $[cm^{-1}]$ *(solid circle) and point impedances Z: middle of BM lower surface (open triangle), middle of RL (open square), and middle of TM lower surface (solid diamond). Impedances* $[dyn/cm^3]$ *are multiplied by the factor* 1×10^7 *for plotting convenience. Frequency is* 200 Hz.

provides the eigenvalue solver with the mixed homogeneous boundary condition: $\partial P_0/\partial n + \rho\omega^2 R_0 Z_n P_0 = 0$. The eigenvalue solver gives the updated wavenumber $k_{eig} = \xi + i\eta$. This value of k_{eig} is used in the elliptic solver for the next iteration of pressure, and so on. The algorithm selects the k_{eig} that corresponds to a propagating mode, and rejects eigenvalues that correspond to evanescent waves with η large and negative. Iteration is stopped when both the real part and imaginary part of k_{eig} given by the eigenvalue solver satisfy the convergence conditions: $(\xi_m - \xi_{m-1})/\xi_m < tolerance$ and $(\eta_m - \eta_{m-1})/\eta_m < tolerance$, where m is the iteration cycle number. For the present calculation, we use $tolerance = 0.01$. Figure 3.1 shows that the algorithm is convergent. We also note that k_{eig} in the upper and lower chambers agree within a few percent.

4. Results and discussion. In this paper we concentrate on developing an algorithm to determine the wavenumber and *relative* motions in fluid and elastic domains as a function of frequency in a single transverse plane. This problem is not without interest in itself, since experiments measuring the motions of the BM and other structures are typically made in a single plane, with the frequency of the stimulus being varied. Thus in our calculation, the cross-sectional geometry is chosen to model the apex of guinea pig cochlea (see Figure 2.1). Parameters used in the computation are listed in Table 4.1. An elasticity map for the model is depicted in Figure 4.1.

Figure 4.2 shows the real part and the imaginary part of the complex-valued dimensional wavenumber $\tilde{k}$ of the propagating wave as a function of frequency f. The real part of $\tilde{k}$ is related to phase or wavelength of the traveling wave, while the imaginary part of $\tilde{k}$ is related to the damping of the traveling wave: as the real part of $\tilde{k}$ increases, the axial wavelength (λ) decreases ($\lambda = 2\pi/Re[\tilde{k}]$), and as the imaginary part of $\tilde{k}$ becomes more negative, the axial damping of the wave increases. A traveling wave with a positive real part and a negative imaginary part of $\tilde{k}$ represents a damped

TABLE 4.1
Parameters used for the apical turn of the guinea pig cochlea.

Symbol	Value and unit	Meaning
ρ	1 g/cm^3	Fluid density
ρ_b	1 g/cm^3	BM density
ρ_t	1 g/cm^3	TM density
ρ_c	1 g/cm^3	OC density
μ	0.01 g/(cm $\cdot$ s)	Fluid viscosity
$E_b{}'$	6e4 dyn $\cdot$ s/cm^2	BM damping
$E_t{}'$	0.3 dyn $\cdot$ s/cm^2	TM damping
$E_c{}'$	0.01 dyn $\cdot$ s/cm^2	OC damping
H_b	0.00015 cm	BM thickness
E_b	2e9 dyn/cm^2	BM Young's modulus
E_t	1e4 dyn/cm^2	TM Young's modulus
E_c	4e2 dyn/cm^2	OC Young's modulus
E_i	4e4 dyn/cm^2	IHCs Young's modulus
E_o	6e4 dyn/cm^2	OHCs Young's modulus
E_d	1e5 dyn/cm^2	DCs Young's modulus
E_p	4e5 dyn/cm^2	PCs Young's modulus
E_r	3e5 dyn/cm^2	RL Young's modulus
ϑ_t	0.49	TM Poisson's ratio
ϑ_c	0.49	OC Poisson's ratio
$b_2 - b_1$	0.025 cm	BM width
H_{gap}	0.00058 cm	Gap thickness
L_{gap}	0.015 cm	Gap length
K_s	1 dyn/cm	Hair bundle stiffness
D_z/D_x	0.1	BM axial coupling parameter

FIG. 4.1. *Elasticity map of the model. Scale is* $log_{10}E$, *where* E (dyn/cm^2) *has the values listed in Table* 4.1.

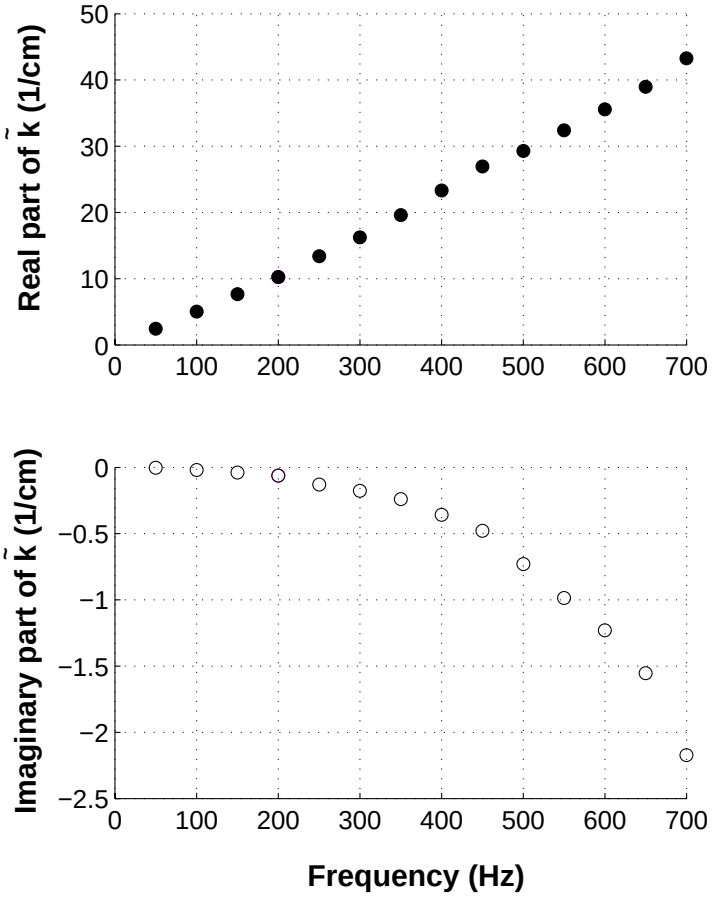


FIG. 4.2. *Dimensional wavenumber $\tilde{k}$. Real (solid circle) and imaginary (open circle) parts.*

right-running wave. von Békésy [6] shows 200 Hz waveforms on the cochlear partition inferred from measurements of amplitude and phase at the apex of human cadaver cochlea. Those waveforms have $\lambda \sim 0.6$cm. From Figure 4.2 we also find $\lambda \sim 2\pi/10 \sim 0.6$cm. At 700 Hz we find $\lambda \sim 2\pi/43 \sim 0.15$cm, which can be compared with the 0.155cm measurement made near the apex of the guinea pig cochlea at 1 kHz by Cooper and Rhode [10].

Modal pressure fields and CP displacements at 200 Hz and 700 Hz are shown in Figures 4.3 and 4.4, respectively. The spatial distribution of the fluid pressure is much more uniform at lower frequency (200 Hz) than at higher frequency (700 Hz). Our calculated pressure distribution in the ST at 700 Hz exhibits a pressure gradient similar to that measured by Olson [27]. In either case, the axial fluid velocity and pressure satisfy $\rho \frac{\partial \tilde{V}_Z}{\partial t} = -\frac{\partial \tilde{P}}{\partial Z}$ when the viscous boundary layer effect is neglected. The WKB approximation of this relation gives $\rho i\omega V_{0Z} = i\tilde{k}P_0$. Thus axial velocity is proportional to the product of wavenumber and fluid pressure at fixed frequency ($V_Z = \tilde{k}P_0/(\rho\omega)$). Because a single wavenumber $\tilde{k}$ exists in both the upper and lower chambers, the mass conservation condition for axial flow is $\langle P_{0u}\rangle A_u + \langle P_{0l}\rangle A_l = 0$, where $\langle\rangle$ denotes the area average. The pressure fields shown in Figures 4.3 and 4.4

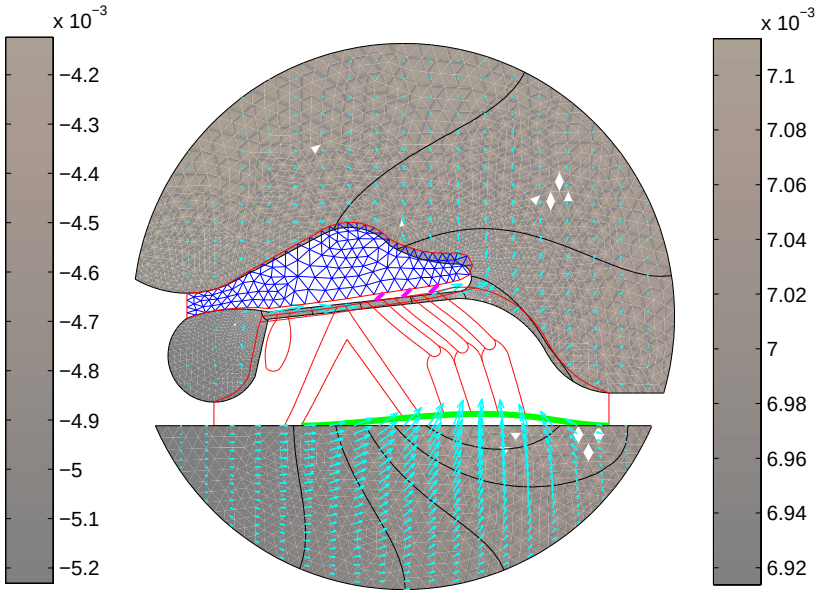


FIG. 4.3. *Snapshot of the real parts of modal pressure fields and CP displacements in an apical cross section at* 200 *Hz. The left and right grayscale maps represent the spatial distribution of the cochlear fluid pressure in the upper and lower chambers, respectively. Axial velocity is proportional to the pressure. The arrows represent the pressure gradient, and the solid lines represent the pressure contours. Stereocilia are shown as short solid lines traversing the thin gap. Fluid velocity leads pressure gradient by* $\pi/2$ *radians, except in the gap, where flow is opposite to pressure gradient.*

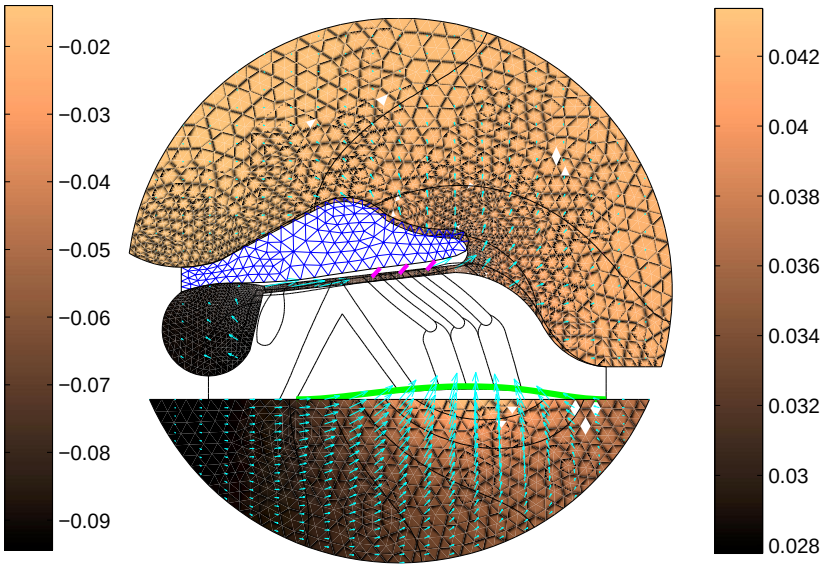


FIG. 4.4. *Snapshot of the real parts of modal pressure fields and CP displacements in an apical cross section at* 700 *Hz. The left and right grayscale maps represent the spatial distribution of the cochlear fluid pressure in the upper and lower chambers, respectively. Axial velocity is proportional to the pressure. The arrows represent the pressure gradient, and the solid lines represent the pressure contours. Stereocilia are shown as short solid lines traversing the thin gap. Fluid velocity leads pressure gradient by* $\pi/2$ *radians, except in the gap, where flow is opposite to pressure gradient.*

satisfy this relation. We can also estimate the axial flow in the spiral sulcus (SS) and the radial flow in the gap. A radial flow in the gap is induced by the pressure gradient existing along the gap. This flow would be important for the stimulation of the IHCs whose stereocilia (not shown) are not in contact with the TM. Note that this Poiseuille flow, estimated as $Q_R \simeq H_{gap}^3 \lambda \Delta P_0/(48\mu L_{gap})$, is much smaller than the axial fluid flow in SS ($Q_Z \simeq \tilde{k} P_{0ss} A_{ss}/(\rho\omega)$). Note also that ΔP_0 is the pressure difference across the gap length, and P_{0ss} and A_{ss} are spiral sulcus pressure and area. Using $\nu = 0.01\text{cm}^2/s, A_{ss} = 0.009^2\text{cm}^2, f = 700$ Hz, $\tilde{k} = 43\text{cm}^{-1}$ (Figure 4.2), $H_{gap} = 6 \times 10^{-4}\text{cm}, Lgap = 0.015\text{cm}$, and $P_{0ss}/\Delta P_0 \simeq 2$ in Figure 4.4, we get $Q_Z/Q_R \simeq 360$.

von Békésy [6] observed the passive BM radial vibrational mode of the guinea pig cochlea near the helicotrema. He found that no vibrational subdivision of the BM by the pillar cells occurred at low frequencies. Our calculated CP vibrational modes agree with his findings (see Figures 4.3 and 4.4). When we increase the Young's modulus of the pillar cells by an order of magnitude, the BM retains its monophasic mode shape.

Figure 4.5 shows the detailed movements within the OC and TM at 700 Hz. We have found similar patterns at lower frequencies. An upward transverse movement of the BM develops into a leftward radial component at the RL. We find that the TM essentially rotates like a rigid body about its attachment to the bone, with little radial motion. This leads to a shear movement between TM and RL which causes a clockwise rotation of the stereocilia, as is commonly believed. This pattern of motion agrees with the experimental findings of Ulfendahl, Khanna, and Heneghan [37] and Hemmert, Zenner, and Gummer [14] at the apex of the guinea pig cochlea, and with the observation of Richter et al. [30] in the hemicochlear preparation. Gummer, Hemmert, and Zenner [13] and Hemmert, Zenner, and Gummer [15] measured a resonant TM radial motion that was rather sharply tuned to a frequency $\sim 1/2$ octave below the BM characteristic frequency. This radial TM resonance was not found by Ulfendahl, Khanna, and Heneghan [37], so there is some experimental disagreement concerning this finding. Lumped-parameter micromechanical models by Allen [3], which involved TM radial stiffness, and by Zwislocki [39], which involved OHC stereocilia stiffness, introduced the idea of a radial TM resonance or "second filter" to reconcile neural and mechanical tuning curves. The present calculations do not show a significant radial TM motion. It would be premature, however, to rule out the possibility that such a motion might be found.

Allen and Sondhi [2] showed that a small amount of axial rigidity reduces the high-frequency slope of the tuning curve. We have included axial elastic coupling in the model via an orthotropic plate model of the BM (see (3.2)). However, it turns out that small elastic coupling ($D_z/D_x = 0.1$) has a negligible effect on the radial mode shape and only slightly decreases the wavenumber. This decrease was previously found by Holmes and Cole [17]. Increasing stereocilia stiffness slightly reduced the real part of the wavenumber, while increasing fluid viscosity slightly increased the imaginary part of the wavenumber. Neither stereocilia stiffness nor viscosity changes seemed to affect the radial mode shapes.

Figure 4.6 shows the shear stresses along the lower surface of the TM in the gap at 700 Hz. The shear stresses due to Couette flow (τ_c) and Poiseuille flow (τ_p) are small relative to the pressure along the gap surface, but the shear stress induced by stereocilia (τ_s) is of the same order as the gap pressure.

In the future we would like to extend the model to include fluid domains in the OC.

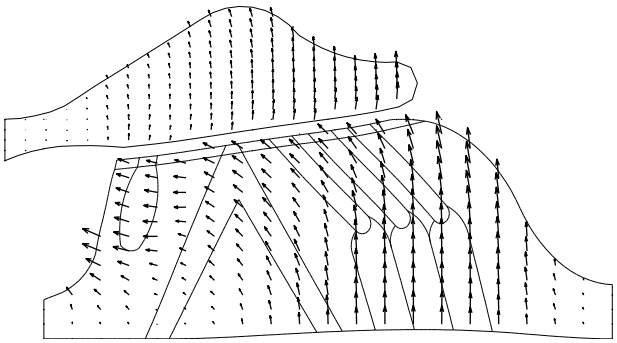


FIG. 4.5. *Detailed movements within the OC and TM. The direction and size of the arrows denote the velocity direction and amplitude of OC and TM movements. A shear movement is present between the TM and RL.*

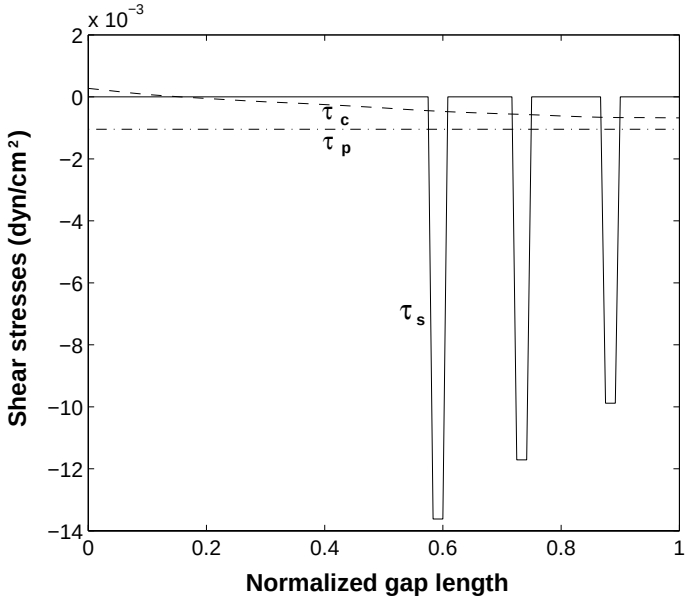


FIG. 4.6. *Shear stresses along the lower surface of the TM in the gap at* 700 *Hz.* τ_p*: Poiseuille flow-induced shear stress,* τ_c*: Couette flow-induced shear stress,* τ_s*: shear stress due to stereocilia stiffness.*

A preliminary effort in that direction has been made by Steele [32], who concluded that leakages between OC fluid domains must be important to mitigating the very high pressures that would otherwise develop inside the OC. A preliminary study of the effect of including the inner tunnel of Corti in our model corroborates that finding. Modeling the basal turn at high frequencies would obviously be desirable, but that presents difficulties in meshing the smaller OC and TM domains with the large fluid domains (ST, SM+SV).

Our results show that the hybrid WKB-numerical approach is a good choice for modeling the cochlea. It avoids a full 3D computation, and the number of transverse sections required for good axial resolution in our hybrid method is far less than the number of axial nodes required in full 3D numerics. This enables us to treat wave

propagation in the complex cochlear geometry with cellular resolution (10 microns) using a desktop workstation. Typical runtimes for a convergent solution at fixed frequency in a single cross section are in the range of 5–20 minutes, depending on initial guesses, while hours of computing time on a high-end computer are needed for a full 3D numerical model [28].

Appendix A. Tangential fluid stress on a deformable boundary segment. Within a thin fluid layer near a deformable boundary other than the TM-RL gap (Figure A.1), the tangential fluid velocity (q_0) and pressure (P_0) satisfy the linearized oscillatory boundary layer equation

$$\rho i\omega q_0 = -R_0^{-1}\frac{\partial P_0}{\partial s} + \mu R_0^{-2}\frac{\partial^2 q_0}{\partial n^2}, \tag{A.1}$$

where ρ and μ are the density and viscosity of the cochlear fluid, and n and s represent the outward normal and tangential direction of the deformable boundary, respectively. Let $q_0^* = V_{s0} - q_0$, where V_{s0} is the local inviscid tangential fluid velocity. Then

$$\mu R_0^{-2}\frac{\partial^2 q_0^*}{\partial n^2} - \rho i\omega q_0^* = -\frac{\partial P_0}{\partial s}R_0^{-1} - \rho i\omega V_{s0} = 0, \tag{A.2}$$

where ω is the radian frequency and $i = \sqrt{-1}$. Therefore,

$$R_0^{-2}\frac{\partial^2 q_0^*}{\partial n^2} - \frac{i\omega}{\nu}q_0^* = 0, \tag{A.3}$$

with $q_0^* \to 0$ as $n \to \infty$, and $q_0^* = V_{s0} - i\omega U_{s0}$ on $n = 0$, where U_{s0} is the boundary segment tangential displacement and $\nu = \mu/\rho$ is kinematic viscosity. Thus

$$q_0^* = (V_{s0} - i\omega U_{s0})e^{-nR_0\sqrt{i\omega/\nu}}, \tag{A.4}$$

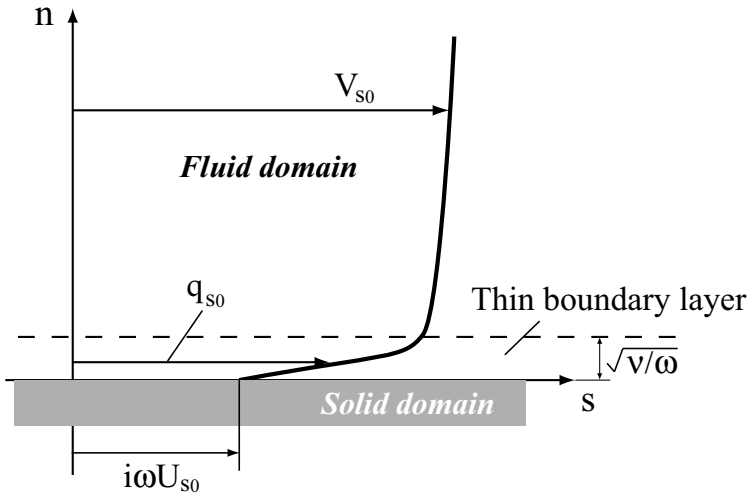


FIG. A.1. *Dominant WKB expansion of the tangential fluid velocity profile in a fluid domain with a deformable boundary segment.* $V_{s0} = -\frac{1}{\rho i\omega R_0}\frac{\partial P_0}{\partial s}$ *is the local inviscid tangential fluid velocity;* q_0 *is the tangential fluid velocity within the thin layer near the deforming boundary;* U_{s0} *is the boundary segment tangential displacement, and* $i\omega U_{s0}$ *is the tangential velocity.*

and the shear stress may be calculated from

$$\tau = \mu R_0^{-1} \frac{\partial q_0^*}{\partial n}\bigg|_{n=0} = \sqrt{i\omega\nu}\left(\frac{1}{i\omega R_0}\frac{\partial P_0}{\partial s} + i\rho\omega U_{s0}\right). \tag{A.5}$$

Appendix B. MATLAB numerical solvers. The MATLAB PDE Toolbox can solve elliptic and eigenvalue problems of the forms

$$-\nabla\cdot(c\nabla P_0) + aP_0 = f \tag{B.1}$$

and

$$-\nabla\cdot(c\nabla P_0) + aP_0 = \lambda d P_0 \tag{B.2}$$

in bounded domain Ω, where c, a, f, d, and the unknown P_0 are scalar, complex-valued functions on Ω, and λ is an unknown eigenvalue. The following boundary conditions are defined for scalar P_0:

(1) Dirichlet: $hP_0 = r$ on the boundary $\partial\Omega$.

(2) Generalized Neumann: $\mathbf{n}\cdot(\nabla P_0) + qP_0 = g$ on $\partial\Omega$,

where $\mathbf{n}$ is the outward unit normal. Here g, q, h, and r are complex-valued functions defined on $\partial\Omega$.

Note that, for solving the elliptic pressure P_0^{elli} of a fluid domain (see (2.10)), we let $c = -1$, $a = k^2$, and $f = 0$ in (B.1), and for solving eigenvalue k^2 and eigenvalue pressure P_0^{eigen} (see (2.10)), we let $c = -1$, $d = -1$, and $a = 0$ in (B.2). We define $q = 0$, and $g = 0$ and $g = -\rho\omega^2 R_0 U_{n_T 0}$ for rigid and deformable boundaries, respectively, to provide the elliptic solver boundary conditions (see (2.11)). For the eigenvalue solver we let $g = 0$ and $q = \rho\omega^2 Z_n$, where $Z_n = -P_0/U_{n0}$ is the impedance of deformable surfaces.

For the OC and TM solid domains, we use the following MATLAB Plane Strain solver:

$$-\nabla\cdot(\underline{\mathbf{c}}\otimes\nabla\mathbf{U_0}) + \mathbf{aU_0} = \mathbf{f}, \tag{B.3}$$

where $\underline{\mathbf{c}}$ is a rank-four tensor, which can be written as four 2-by-2 matrices c_{11}, c_{12}, c_{21}, c_{22}:

$$c_{11} = \begin{pmatrix} 2G+\zeta & 0 \\ 0 & G \end{pmatrix}, \qquad c_{12} = \begin{pmatrix} 0 & \zeta \\ G & 0 \end{pmatrix}, \tag{B.4}$$

$$c_{21} = \begin{pmatrix} 0 & G \\ \zeta & 0 \end{pmatrix}, \qquad c_{22} = \begin{pmatrix} G & 0 \\ 0 & 2G+\zeta \end{pmatrix},$$

where G, the shear modulus, is defined by $G = E/2(1+\vartheta)$, and ζ in turn is defined by $2G\vartheta/(1-2\vartheta)$. $\mathbf{f} = (f_x, f_y)^T$ are volume forces. By the notation $\nabla\cdot(\underline{\mathbf{c}}\otimes\nabla\mathbf{U_0})$, we mean the 2-by-1 vector with $(i,1)$-component:

$$\sum_{j=1}^{2}\left(\frac{\partial}{\partial x}c_{ij11}\frac{\partial}{\partial x} + \frac{\partial}{\partial x}c_{ij12}\frac{\partial}{\partial y} + \frac{\partial}{\partial y}c_{ij21}\frac{\partial}{\partial x} + \frac{\partial}{\partial y}c_{ij22}\frac{\partial}{\partial y}\right)U_{0j}. \tag{B.5}$$

When constructing the geometry of the OC, we define the left and right subdomain numbers of each segment in the geometry matrix. Three matrices of fixed format contain the information about mesh points, the boundary segments, and the triangles

when a triangular mesh is built on the domain Ω by MATLAB mesh generating and refining facilities. The MATLAB function pdesdt can provide indices of the triangles inside a subdomain. Thus we can discretize the above PDE coefficients c_{ij} and boundary conditions on Ω to obtain a linear system which will give the approximate solution at the mesh points of the unknown displacements (U_{0x} and U_{0y}). Because the PDE coefficients c_{ij} contain the Young's modulus E and the Poisson's ratio ϑ, the inhomogeneities of the OC are introduced by its E-map discretized on Ω. (Note that the Poisson's ratio ϑ is constant on all subdomains.)

Acknowledgments. The authors thank E. K. Dimitriadis, K. Iwasa, and B. Shoelson for their helpful comments.

REFERENCES

[1] J.B. Allen, *Cochlear micromechanics—A mechanism for transforming mechanical to neural tuning within the cochlea*, J. Acoust. Soc. Amer., 62 (1977), pp. 930–939.
[2] J.B. Allen and M.M. Sondhi, *Cochlear macromechanics: Time domain solutions*, J. Acoust. Soc. Amer., 66 (1979), pp. 123–132.
[3] J.B. Allen, *Cochlear micromechanics: A physical model of transduction*, J. Acoust. Soc. Amer., 68 (1980), pp. 1660–1670.
[4] J.B. Allen and S.T. Neely, *Micromechanical models of the cochlea*, Physics Today, 45 (1992), pp. 40–47.
[5] G.K. Batchelor, *An Introduction to Fluid Mechanics*, Cambridge University Press, London, 1967, pp. 353–358.
[6] G. von Békésy, *Experiments in Hearing*, McGraw–Hill, New York, 1960.
[7] F. Böhnke, J. von Mikusch-Buchberg, and W. Arnold, *3D finite elemente modell des cochleären verstärker*, Biomed. Tech., 42 (1996), pp. 311–312.
[8] R.S. Chadwick, *Three dimensional effects on low frequency cochlear mechanics*, Mech. Res. Comm., 12 (1985), pp. 181–186.
[9] R.S. Chadwick, E.K. Dimitriadis, and K.H. Iwasa, *Active control of waves in a cochlear model with subpartitions*, Proc. Natl. Acad. Sci. USA, 93 (1996), pp. 2564–2569.
[10] N.P. Cooper and W.S. Rhode, *Fast traveling waves, slow traveling waves and their interactions in experimental studies of apical cochlear mechanics*, Auditory Neurosci., 2 (1996), pp. 289–299.
[11] N.P. Cooper, *Radial variation in the vibrations of the cochlear partition*, in Recent Developments in Auditory Mechanics, H. Wada, T. Takasaka, K. Ikeda, K. Ohyama, and T. Koike, eds., World Scientific, River Edge, NJ, 1999, pp. 109–115.
[12] C.D. Geisler, *A model of the effect of outer hair cell motility on cochlear vibrations*, Hear. Res., 24 (1986), pp. 125–131.
[13] A.W. Gummer, W. Hemmert, and H. Zenner, *Resonant tectorial membrane motion in the inner ear: Its crucial role in frequency tuning*, Proc. Natl. Acad. Sci. USA, 93 (1996), pp. 8727–8732.
[14] W. Hemmert, H. Zenner, and A.W. Gummer, *Characteristics of the traveling wave in the low-frequency region of a temporal-bone preparation of the guinea pig cochlea*, Hear. Res., 142 (2000), pp. 184–202.
[15] W. Hemmert, H. Zenner, and A.W. Gummer, *Three-dimensional motion of the organ of Corti*, Biophys. J., 78 (2000), pp. 2285–2297.
[16] L.F. Hao and S.M. Khanna, *Vibration of the guinea pig organ of Corti in the apical turn*, Hear. Res., 148 (2000), pp. 47–62.
[17] M. Holmes and J.D. Cole, *Cochlear mechanics: Analysis for a pure tone*, J. Acoust. Soc. Amer., 76 (1984), pp. 767–778.
[18] Y. Kagawa, T. Yamabuchi, N. Watanabe, and N. Mizoguchi, *Finite element cochlear models and their steady state response*, J. Sound and Vibration, 119 (1987), pp. 291–315.
[19] H. Kolsky, *Stress Waves in Solids*, Dover, New York, 1963.
[20] P.J. Kolston and J.F. Ashmore, *Finite element micromechanical modeling of the cochlea in three dimensions*, J. Acoust. Soc. Amer., 93 (1996), pp. 455–467.
[21] M. Lien, *A Mathematical Model of the Mechanics of the Cochlea*, Sc.D. dissertation, Department of Electrical Engineering, Washington University in St. Louis, St. Louis, MO, 1973.

[22] F. Mammano and R. Nobili, *Biophysics of the cochlea: Linear approximation*, J. Acoust. Soc. Amer., 93 (1993), pp. 3320–3332.
[23] D. Manoussaki and R.S. Chadwick, *Effects of geometry on fluid loading in a coiled cochlea*, SIAM J. Appl. Math., 61 (2000), pp. 369–386.
[24] S.T. Neely and D.O. Kim, *An active cochlear model showing sharp tuning and high sensitivity*, Hear. Res., 9 (1983), pp. 123–130.
[25] K.E. Nilsen and I.J. Russell, *Timing of cochlear feedback: Spatial and temporal representation of tone across the basilar membrane*, Nature Neurosci., 2 (1999), pp. 642–648.
[26] A.L. Nuttall, M. Guo, and T. Ren, *The radial pattern of basilar membrane motion evoked by electric stimulation of the cochlea*, Hear. Res., 131 (1999), pp. 39–46.
[27] E.S. Olson, *Direct measurement of intra-cochlear pressure waves*, Nature, 402 (1999), pp. 526–529.
[28] A.A. Parthasarathi, K. Grosh, and A.L. Nuttall, *Three-dimensional numerical modeling for global cochlear dynamics*, J. Acoust. Soc. Amer., 107 (2000), pp. 474–485.
[29] C.P. Richter, B.N. Evans, R. Edge, and P. Dallos, *Basilar membrane vibration in the gerbil hemicochlea*, J. Neurophysiol., 79 (1998), pp. 2255–2264.
[30] C.P. Richter, B.N. Evans, R. Edge, and P. Dallos, *Basilar membrane micro-mechanics measured in the gerbil inner ear*, Assoc. Res. Otolaryngol. Abs., G.R. Popelka, ed., St. Petersburg, FL, 1988, p. 181.
[31] L. Robles and M.A. Ruggero, *Mechanics of the mammalian cochlea*, Physiolog. Rev., 81 (2001), pp. 1305–1351.
[32] C.R. Steele, *Toward three-dimensional analysis of cochlear structure*, J. Oto-rhino-laryngology, 61 (1999), pp. 238–251.
[33] C.R. Steele and L.A. Taber, *Three dimensional model calculation for the guinea pig cochlea*, J. Acoust. Soc. Amer., 69 (1981), pp. 1107–1118.
[34] L.A. Taber and C.R. Steele, *Cochlear modeling including three dimensional fluid and four modes of partition flexibility*, J. Acoust. Soc. Amer., 70 (1982), pp. 426–436.
[35] S. Timoshenko and J.N. Goodier, *Theory of Elasticity*, McGraw–Hill, New York, 1951.
[36] S. Timoshenko and S. Woinowsky-Krieger, *Theory of Plates and Shells*, McGraw–Hill, New York, 1959.
[37] M. Ulfendahl, S.M. Khanna, and C. Heneghan, *Shearing motion in the hearing organ measured by confocal laser heterodyne interferometry*, Neuroreport, 6 (1995), pp. 1157–1160.
[38] L. Zhang, D.C. Mountain, and A.E. Hubbard, *Shape and stiffness changes of the organ of Corti from the base to apex cannot predict characteristic frequencies changes: Are multiple modes the answer?*, in Diversity in Auditory Mechanics, E.R. Lewis, G.R. Long, R.F. Lyon, P.M. Narins, C.R. Steele, and E. Hecht-Poinar, eds., World Scientific, Singapore, 1997, pp. 472–478.
[39] J.J. Zwislocki, *Five decades of research of cochlear mechanics*, J. Acoust. Soc. Amer., 67 (1980), pp. 1679–1685.
[40] J.J. Zwislocki and E.J. Kletzki, *What basilar membrane tuning says about cochlear micromechanics*, Am. J. Otolaryngol., 3 (1982), pp. 48–52.

SIAM J. APPL. MATH.
Vol. 63, No. 4, pp. 1121–1140

SIMULATION OF WAVE INTERACTIONS AND TURBULENCE IN ONE-DIMENSIONAL WATER WAVES*

KURT M. BERGER† AND PAUL A. MILEWSKI‡

Abstract. The weak- or wave-turbulence problem consists of finding statistical states of a large number of interacting waves. These states are obtained by forcing and dissipating a conservative dispersive wave equation at disparate scales to model physical forcing and dissipation, and by predicting the spectrum, often as a Kolmogorov-like power law, at intermediate scales. The mechanism for energy transfer in such systems is usually triads or quartets of waves. Here, we first derive a small-amplitude nonlinear dispersive equation (a finite-depth Benney–Luke-type equation), which we validate, analytically and numerically, by showing that it correctly captures the main deterministic aspects of gravity wave interactions: resonant quartets, Benjamin–Feir-type wave-packet stability, and wave-mean flow interactions. Numerically, this equation is easier to integrate than either the full problem or the Zakharov integral equation. Some additional features of wave interaction are discussed such as harmonic generation in shallow water. We then perform long time computations on the forced-dissipated model equation and compute statistical quantities of interest, which we compare to existing predictions. The forward cascade yields a spectrum close to the prediction of Zakharov, and the inverse cascade does not.

Key words. water waves, wave turbulence, finite depth, quartets

AMS subject classifications. 74J15, 74J30, 76B15, 76F99

PII. S0036139902402063

1. Introduction.

The weak- or wave-turbulence problem consists of finding statistical states of a large number of interacting waves. These states are obtained by forcing and dissipating a conservative dispersive wave problem at disparate scales and predicting the spectrum, often as a Kolmogorov-like power law, at intermediate scales. In dispersive waves, the energy transfer between waves occurs mostly amongst *resonant* sets of waves, usually triads or quartets of waves. Here we consider only quartets since triads do not exist in surface gravity waves. Quartet resonances occur when the product of a pair of waves has a component with the same frequency and wavenumber as the product of two other waves. For simple waves $e^{i(\mathbf{k}_j\cdot\mathbf{x}-\omega(\mathbf{k}_j)t)}$, this means

$$\mathbf{k}_1+\mathbf{k}_2=\mathbf{k}_3+\mathbf{k}_4, \tag{1.1}$$

$$\omega(\mathbf{k}_1)+\omega(\mathbf{k}_2)=\omega(\mathbf{k}_3)+\omega(\mathbf{k}_4). \tag{1.2}$$

In dispersive problems, these resonant sets are sparse, in contrast to nondispersive problems, where interactions are dense in Fourier space ((1.2) is always satisfied). The deterministic dynamics of *isolated* resonant quartets are modeled by sets of coupled nonlinear differential equations for the wave amplitudes and are well understood (see [10]). The dynamics of quartets which are not isolated (allowed to interact with other quartets) are poorly understood. In the limit in which all possible quartets are active, statistical theories of wave turbulence apply.

*Received by the editors February 4, 2002; accepted for publication (in revised form) September 16, 2002; published electronically March 26, 2003.
http://www.siam.org/journals/siap/63-4/40206.html

†Department of Mathematics, The Ohio State University, 231 W. 18th Ave., Columbus, OH 43210 (berger@oblon.com).

‡Department of Mathematics, University of Wisconsin, 480 Lincoln Dr., Madison, WI 53706 (milewski@math.wisc.edu). The research of this author was partially supported by NSF-DMS and a Sloan Research Fellowship.

The initial work on wave turbulence was done by Hasselmann [14], Benney and Saffmann [7], and Benney and Newell [5], who introduced the statistical closures based on the resonant wave interactions. Zakharov [29], through conformal transformations, solved the resulting kinetic equation and obtained the power law for the Kolmogorov spectrum. The particular physical context for these initial results was the ocean surface gravity wave spectrum.

Majda, McLaughlin, and Tabak [19] started the numerical investigation of the predictions of weak turbulence theory using a nonlinear Schrödinger-like (NLS) model equation. Adding large-scale forcing and dissipation to their one-dimensional model, they investigated the turbulent cascades of energy and initially showed that Zakharov's [29] prediction for the energy spectrum did not hold, proposing a simpler, yet unrigorous quartet-based scaling to explain their results. More recent work by Cai et al. [9] and Zakharov et al. [30] shows that several meta-stable spectra can coexist in the system, with the Zakharov spectra being among those observed.

Classical, small-amplitude periodic gravity waves, discovered by Stokes, are unstable to small modulations through the Benjamin–Feir instability. This result was derived independently by Lighthill [15], Benjamin [1], and Whitham [26], and confirmed experimentally by Benjamin and Feir [2]. One can obtain the result by analyzing the slow modulation of gravity waves and deriving an NLS equation for the evolution of the wave envelope (see Hasimoto and Ono [13] and Zakharov [28], among others). A plane wave solution of the NLS equation corresponds to the Stokes wave, and it can be shown for waves of wavelength $\frac{2\pi}{k}$ in water of depth H to be unstable when $kH > 1.363$. (Waves in deeper water are unstable, and the NLS switches from "defocusing" to "focusing.") Davey and Stewartson [11] generalized this result to two spatial dimensions, deriving a more complicated NLS-type equation. This result for two-dimensional waves was derived independently a few years earlier, however, by Benney and Roskes [6], albeit in a slightly different form.

Here we investigate wave interaction and turbulence numerically for an equation describing small-amplitude gravity water waves. We perform wave interaction experiments and long time wave turbulence computations using a finite-depth Benney–Luke (fBL) equation [21]. To validate this model, we first show, analytically and numerically, that the fBL equation correctly captures the main deterministic aspects of resonant gravity wave interactions: resonant quartets and the Benjamin–Feir-type wave-packet stability. Some additional features of our numerical results are discussed: the generation of harmonics in shallow water and the long time frequency downshift of unstable wavepackets. For the wave-turbulence experiments, we compare the computed wave spectrum to predicted spectra. We note that the use of a single partial differential equation, rather than the full water wave equations, makes computing complex surface wave dynamics possible. All of our work is for a one-dimensional free surface. Although the computation of the two-dimensional free surface problem is not fundamentally different, we restrict our attention to the one-dimensional problem because of computational time constraints.

We note that there is a fundamental difference in the wave interaction problem between the one-dimensional and two-dimensional free surface. In two dimensions and infinite depth, the fundamental interaction mechanism is resonant quartets. The quartet interaction coefficients, however, vanish as the waves become parallel to each other, and therefore, for one-dimensional infinite depth, quartets are not important. Thus in a one-dimensional deep water system, the strongest mechanism for energy exchange between Fourier modes is the Benjamin–Feir instability, which is local in Fourier space, and the slower quintet interaction, which requires quartic terms in the

equations to be modeled correctly. The Benjamin–Feir instability is also relevant for two-dimensional free surface problems.

For a one-dimensional free surface over water of finite depth, there exist quartet interactions (which vanish as $|\mathbf{k}| \to \infty$ to agree with the deep water limit). Therefore the one-dimensional finite depth problem is a computationally accessible useful test for the more relevant two-dimensional problem. That is why we restrict our numerical calculations to waves that are long enough to be influenced by the bottom.

The remainder of this paper is organized as follows. In section 2, we derive the fBL equation. Next, in section 3, we derive the nonlinear Schrödinger equation from the one-dimensional fBL equation using a multiple-scales approach, in a manner similar to Hasimoto and Ono, who started from the full water wave equations. This NLS equation correctly predicts the Benjamin–Feir instability limit, which we verify numerically using the fBL equation. In section 4 we derive a set of new partial differential equations that describe the coupled evolution of quartets and the induced mean flow for one-dimensional finite-depth gravity waves. We then show that solutions to these quartet equations closely match numerical solutions of the fBL equation, when initialized with four waves that satisfy the resonance conditions. We also study a model of the interaction of a primary wave and its quasi-resonant second harmonic in shallow water to explain the quartet simulation results. Finally, in the last section, we investigate wave turbulence numerically using the fBL model.

2. The Benney–Luke equation for gravity waves in finite depth. The Benney–Luke equation [4] describes the evolution of three-dimensional, weakly nonlinear waves in shallow water. Recently Milewski and Keller [22] derived a more general Benney–Luke model for waves in water of finite depth, shown here in a slightly different (and corrected) form:

$$u_{tt} + \mathcal{L}u + \epsilon\mathcal{N}_1(u,u) + \epsilon^2\mathcal{N}_2(u,u,u) = 0 \tag{2.1}$$

with quadratic terms

$$\mathcal{N}_1 = (\nabla u)_t^2 + (\mathcal{L}u)_t^2 + u_t\Delta u - u_t\mathcal{L}u_{tt} \tag{2.2}$$

and cubic terms

$$\mathcal{N}_2 = \frac{1}{6}\nabla\cdot(\nabla u\,(\nabla u)^2) + (\Delta u - \mathcal{L}^2 u)\left(u_t\mathcal{L}u_t - \frac{1}{2}(\mathcal{L}u)^2\right) + 2u_t\Delta u_t\mathcal{L}u$$

$$- 2u_t(\nabla u\cdot\nabla\mathcal{L}u)_t + 2\mathcal{L}u(\nabla u\cdot\nabla\mathcal{L}u) + \frac{1}{2}\mathcal{L}^2u(\nabla u)^2. \tag{2.3}$$

In this equation, $\epsilon = a/H \ll 1$ is the ratio of wave amplitude a to depth H, $u(x,y,t)$ is the velocity potential at the undisturbed free surface $z = H$, and $\mathcal{L}$ is the operator $\mathcal{L} = (-\Delta)^{\frac{1}{2}}\tanh[(-\Delta)^{\frac{1}{2}}]$, resulting in the dispersion relation $\omega^2 = |\mathbf{k}|\tanh(|\mathbf{k}|)$. The water surface is given by $H + \eta(x,y,t)$, where, to leading order, $\eta = -u_t$. The fBL is derived as follows. Using the depth H as both the horizontal and vertical length scale, a as the scale for typical free surface displacements, $a\sqrt{gH}$ as the velocity potential scale, and $\sqrt{H/g}$ as the time scale, the dimensionless water wave equations can be written in terms of the velocity potential $\phi(x,y,z,t)$ and free

surface displacement $\eta(x,y,t)$ as

$$\Delta\phi + \phi_{zz} = 0, \quad 0 < z < 1 + \epsilon\eta, \tag{2.4}$$

$$\phi_z = 0, \quad z = 0, \tag{2.5}$$

$$\eta_t + \epsilon(\nabla\eta \cdot \nabla\phi) - \phi_z = 0, \quad z = 1 + \epsilon\eta, \tag{2.6}$$

$$\phi_t + \frac{\epsilon}{2}(\nabla\phi)^2 + \frac{\epsilon}{2}\phi_z^2 + \eta = 0, \quad z = 1 + \epsilon\eta. \tag{2.7}$$

Expanding the two surface boundary conditions about $z = 1$ and eliminating η leads to a single boundary condition in ϕ at $z = 1$, correct to $\mathrm{O}(\epsilon^2)$:

$$\phi_{tt} + \phi_z + \epsilon\mathcal{Q}_1(\phi,\phi) + \epsilon^2\mathcal{Q}_2(\phi,\phi,\phi) = 0, \tag{2.8}$$

where the quadratic terms are

$$\mathcal{Q}_1(\phi,\phi) = \left[\frac{1}{2}((\nabla\phi)^2 + \phi_z^2) - \phi_t\phi_{tz}\right]_t + \nabla\cdot(\phi_t\nabla\phi) \tag{2.9}$$

and the cubic terms are

$$\begin{aligned}\mathcal{Q}_2(\phi,\phi,\phi) = &\left[-\frac{1}{2}\phi_t((\nabla\phi)^2 + \phi_z^2)_z + \phi_t\phi_{tz}^2 + \frac{1}{2}\phi_{tzz}\phi_t^2\right]_t \\ &+ \nabla\cdot\left[\frac{1}{2}(\nabla\phi)((\nabla\phi)^2 + \phi_z^2) - (\nabla\phi)\phi_t\phi_{tz} - \frac{1}{2}(\nabla\phi_z)\phi_t^2\right].\end{aligned} \tag{2.10}$$

Next, we solve Laplace's equation with the bottom boundary condition, obtaining

$$\phi(x,y,z,t) = \cosh[z(-\Delta)^{\frac{1}{2}}]\Phi(x,y,t), \tag{2.11}$$

with

$$u(x,y,t) = \phi(x,y,1,t) = \cosh[(-\Delta)^{\frac{1}{2}}]\Phi(x,y,t) \tag{2.12}$$

being the velocity potential at $z = 1$. With this notation, it follows that $\phi_z(x,y,1,t) = \mathcal{L}u$ and $\phi_{zz}(x,y,1,t) = -\Delta u$, where $\mathcal{L}$ is defined as $\mathcal{L} = (-\Delta)^{\frac{1}{2}}\tanh[(-\Delta)^{\frac{1}{2}}]$ and has the symbol $\hat{\mathcal{L}}(\mathbf{k}) = |\mathbf{k}|\tanh(|\mathbf{k}|)$. Thus if $\hat{u}(\mathbf{k},t)$ is the Fourier transform of $u(\mathbf{x},t)$, then

$$\mathcal{L}u = \frac{1}{2\pi}\int_{-\infty}^{\infty} |\mathbf{k}|\tanh(|\mathbf{k}|)e^{i\mathbf{k}\cdot\mathbf{x}}\hat{u}(\mathbf{k},t)d\mathbf{k}. \tag{2.13}$$

Substitution into the boundary condition (2.8) yields, after some simplification, the fBL equation (2.1).

We note that since ϵ is the ratio of the amplitude of the free surface displacement to depth, the wave slope appears to be arbitrary. However, note that for $|\mathbf{k}|$ large the wave slope $\hat{\eta}_x = O(|\mathbf{k}|^{3/2}\hat{u})$ and that in (2.1), $\mathcal{N}_j = O(|\mathbf{k}|^{(1+3/2j)}\hat{u}^j)$, thus implying that solutions of the fBL are relevant only if the wave slope is also small as waves get short compared to depth. For the shallow limit, $|\mathbf{k}|$ small, $\hat{\eta} = O(|\mathbf{k}|\hat{u})$, $\hat{\eta}_x = O(|\mathbf{k}|^2\hat{u})$, and $\mathcal{N}_j = O(|\mathbf{k}|^{(2+j)}\hat{u}^j)$, requiring only that η be small.

A similar equation applies for gravity waves in water of infinite depth, now with ϵ being the wave slope (ratio of the amplitude of the surface displacement to a characteristic length scale), and $\hat{\mathcal{L}} = |k|$. Therefore, $\mathcal{L} = (-\Delta)^{\frac{1}{2}}$, which, in the case of one horizontal dimension, is $\mathcal{L} = -\partial_x \mathcal{H}$, where $\mathcal{H}$ is the Hilbert transform. For the deep-water limit, the derivation must be modified slightly. The origin of the vertical axis is shifted to the undisturbed fluid level, and the bottom boundary condition becomes $|\nabla\phi| \to 0, z \to -\infty$. Expanding the two surface boundary conditions about $z = 0$ and eliminating η again leads to (2.8). Solving Laplace's equation with the new bottom boundary condition modifies the depth dependence of the velocity potential:

$$\phi(x, y, z, t) = e^{z(-\Delta)^{\frac{1}{2}}} \Phi(x, y, t). \tag{2.14}$$

Correspondingly, the velocity potential at $z = 0$ is just

$$u(x, y, t) = \phi(x, y, 0, t) = \Phi(x, y, t), \tag{2.15}$$

and $\mathcal{L}$ is now defined as $\mathcal{L} = (-\Delta)^{\frac{1}{2}}$ and has the symbol $\hat{\mathcal{L}}(\mathbf{k}) = |\mathbf{k}|$.

In the remainder of this paper we assume that the free surface is one-dimensional.

3. Nonlinear modulation of gravity waves. We consider the slow modulation of one-dimensional gravity waves in water of finite depth using the fBL equation, obtaining an NLS equation, in agreement with earlier results. This equation predicts instability for $kH > 1.363$. Of critical importance in the derivation of this NLS equation is a wave-induced mean flow, which vanishes in the deep water limit.

3.1. Derivation of an NLS equation. In what follows, we employ the method of multiple scales, introducing the slow space and time scales $X = \epsilon x$, $T = \epsilon t$, and $\tau = \epsilon^2 t$. The NLS equation governs the evolution of wave packets or, alternatively, of a narrowly peaked Fourier spectrum centered at k_c. Thus one expands the governing equations with $k = k_c + \epsilon\Delta k$. The equation in physical space is then recovered with the duality $\partial_X \leftrightarrow i\epsilon\Delta k$. Thus, in (2.1) we make the substitutions $\partial_t \to \partial_t + \epsilon\partial_T + \epsilon^2\partial_\tau$ and $\partial_x \to \partial_x + \epsilon\partial_X$ and, for $\mathcal{L}$,

$$\mathcal{L} \to \mathcal{L} - \epsilon i \frac{\partial \hat{\mathcal{L}}}{\partial k}\partial_X - \frac{1}{2}\epsilon^2 \frac{\partial^2 \hat{\mathcal{L}}}{\partial k^2}\partial_{XX}, \tag{3.1}$$

where k_c is denoted k. The dispersion relation is

$$\omega^2(k) = \hat{\mathcal{L}} = |k| \tanh(|k|), \tag{3.2}$$

and

$$\frac{\partial \hat{\mathcal{L}}}{\partial k} = 2\omega c_g(k), \tag{3.3}$$

$$\frac{\partial^2 \hat{\mathcal{L}}}{\partial k^2} = 2c_g^2 + 2\omega \frac{\partial c_g}{\partial k}, \tag{3.4}$$

where $c_g(k)$ is the group velocity.

After substitution, we have the following equation for $u(x, t, X, T, \tau)$:

$$\begin{aligned} &u_{tt} + \mathcal{L}u + \epsilon\left(2u_{tT} - i\frac{\partial \hat{\mathcal{L}}}{\partial k}u_X + \mathcal{N}_1(u, u)\right) \\ &\quad + \epsilon^2\left(2u_{t\tau} + u_{TT} - \frac{1}{2}\epsilon^2 \frac{\partial^2 \hat{\mathcal{L}}}{\partial k^2}u_{XX} + \mathcal{N}_2(u, u, u) + \mathcal{M}(u, u)\right) = 0, \end{aligned} \tag{3.5}$$

where

$$\mathcal{M}(u,u) = u_T(u_{xx} - \mathcal{L}u_{tt}) + 2u_t(u_{xX} - \mathcal{L}u_{tT}) + iu_t\frac{\partial\hat{\mathcal{L}}}{\partial k}u_{ttX} + 2u_x(u_{xT} + u_{Xt})$$
$$+ 2u_Xu_{xt} - 2i\mathcal{L}u\frac{\partial\hat{\mathcal{L}}}{\partial k}u_{tX} + 2\mathcal{L}u\mathcal{L}u_T - 2i\frac{\partial\hat{\mathcal{L}}}{\partial k}u_X\mathcal{L}u_t. \tag{3.6}$$

Next, we expand u in the small parameter ϵ as $u = u_0 + \epsilon u_1 + \epsilon^2 u_2 + \cdots$ and look for a single plane wave of slowly varying amplitude and wavelength $\frac{2\pi}{k}$:

$$u_0(x,t,X,T,\tau) = A(X,T,\tau)e^{i\theta} + * + B(X,T,\tau), \tag{3.7}$$

where $\theta = kx - \omega t$, B is the "mean-flow" component, and the $*$ denotes the complex conjugate of the preceding terms. We note that although the waves are $O(\epsilon)$, the mean flow B_x is $O(\epsilon^2)$. Substitution into (3.5) leads to a series of equations at various orders of ϵ. The O(ϵ) equation is

$$u_{1tt} + \mathcal{L}u_1 = -\left(2u_{0tT} - i\frac{\partial\hat{\mathcal{L}}}{\partial k}u_{0X}\right) - \mathcal{N}_1(u_0,u_0). \tag{3.8}$$

The first terms on the right of the above equation are secular and impose that A is moving at the group velocity. Thus, with $\xi = X - c_gT$, the right-hand side becomes $3i\omega|k|^2(\sigma^2-1)A^2e^{2i\theta} + *$, where $\sigma = \tanh(|k|)$, $A = A(\xi,\tau)$, and

$$u_1 = \frac{3i|k|^2(1-\sigma^4)}{4\sigma^2\omega}A(\xi,\tau)^2e^{2i\theta} + *. \tag{3.9}$$

Proceeding to O(ϵ^2) terms, the equation is

$$u_{2tt} + \mathcal{L}u_2 = \left(2u_{1tT} - i\frac{\partial\hat{\mathcal{L}}}{\partial k}u_{1X}\right) - \left(2u_{0t\tau} + u_{0TT} - \frac{1}{2}\frac{\partial^2\hat{\mathcal{L}}}{\partial k^2}u_{0XX}\right)$$
$$- \left(\mathcal{N}_1(u_0,u_1) + \mathcal{N}_1(u_1,u_0) + \mathcal{M}(u_0,u_0) + \mathcal{N}_2(u_0,u_0,u_0)\right). \tag{3.10}$$

In this equation, eliminating the secular terms in $e^{i\theta}$ and mean flows (e^{i0}) leads to the two equations

$$B_{\xi\xi} = \frac{\gamma}{c_g^2 - 1}(AA^*)_\xi \tag{3.11}$$

and

$$iA_\tau + \alpha A_{\xi\xi} = \bar{\beta}|A|^2A + \frac{\gamma}{2\omega}B_\xi A, \tag{3.12}$$

where

$$\gamma(k) = 2k\omega + c_g|k|^2(1-\sigma^2),$$
$$\alpha(k) = \frac{1}{2}\frac{\partial c_g}{\partial k},$$
$$\bar{\beta}(k) = \frac{9 - 12\sigma^2 + 13\sigma^4 - 2\sigma^6}{4\omega\sigma^2}|k|^4.$$

Integrating (3.11) to obtain the induced horizontal mean flow $B_\xi = \frac{\gamma}{c_g^2-1}|A|^2$ (we ignore the constant of integration, which would correspond to an imposed weak flow)

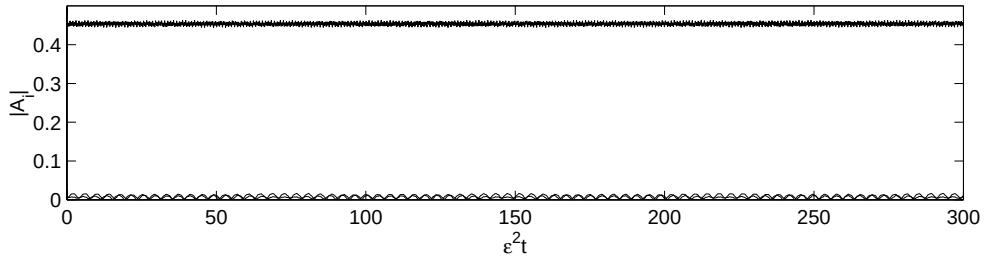


FIG. 3.1. *Evolution of a single plane wave and two small side-bands using the one-dimensional fBL equation. Here* $kH = 1.344 < 1.363$, *and stability is expected.*

and substituting into (3.12) yields a nonlinear Schrödinger equation for the complex amplitude $A(\xi, \eta)$:

$$iA_\tau + \alpha A_{\xi\xi} = \beta |A|^2 A, \tag{3.13}$$

with

$$\beta(k) = \bar{\beta} + \frac{\gamma^2}{2\omega(c_g^2 - 1)}. \tag{3.14}$$

The well-known fact that the mean flow vanishes in the deep water limit can be obtained by writing the mean flow in dimensional variables and taking $H \to \infty$.

3.2. Benjamin–Feir instability. The plane wave solution of the NLS equation $A = A_0 e^{-i\beta|A_0|^2\tau}$ for constant A_0 corresponds to the Stokes wave train to $O(\epsilon^2)$ (see [13]). Moreover, linear stability analysis (see [10] and [13]) shows that a plane wave solution to (3.13) will be unstable if the product $\alpha\beta < 0$. Given the finite-depth dispersion relation, we find $\alpha(k) < 0$ for all k, and $\beta(k)$ changes sign at $k \approx 1.363$, becoming positive for k larger than this value. This is the well-known Benjamin–Feir instability criterion. Note that the induced mean-flow plays an important result in this derivation, and in the deep-water limit this flow is not present.

To verify the stability predictions of this NLS equation, we numerically solve the fBL equation with initial condition $u(x, 0) = Ae^{ikx} + a(e^{i(k+\Delta k)x} + e^{i(k-\Delta k)x}) + *$ corresponding to a primary plane wave of wavenumber k and two side-bands of the next adjacent wavenumbers. We use a relative amplitude of $a = 0.01A$ with $\Delta k = \frac{1}{32}$. Dimensionally, our wavenumber k corresponds to kH, and we take two values on either side of the $kH = 1.363$ limit. Figures 3.1 and 3.2 show the results for $kH = 1.344$ and $kH = 1.438$, respectively, on the long time scale $\tau = \epsilon^2 t$. Note the instability of the primary mode and the side-bands in the second figure. The calculations do not show cyclic modulation and demodulation (or recurrence, present for some limits of the Benjamin–Feir instability) due to the relatively large amplitude of the carrier wave.

The extension to two dimensions (two-dimensional instabilities of plane waves) is straightforward, and the fBL equation is an appropriate starting point for an asymptotic study (such as that of Davey and Stewartson and of Benney and Roskes) or numerical experiments.

We note that in our calculations of the unstable Benjamin–Feir regime it is the lower Fourier side-bands that dominate the spectrum. This "frequency downshift"

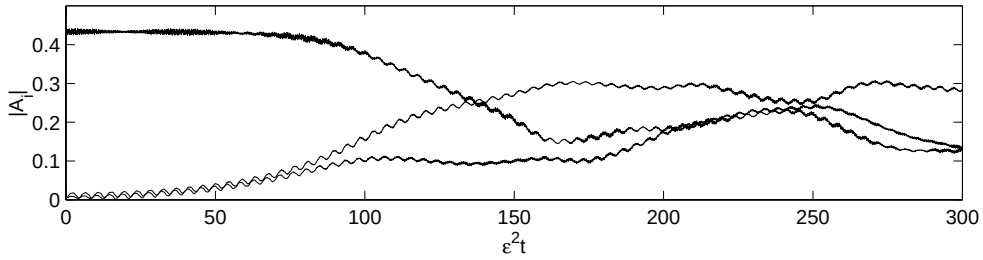


FIG. 3.2. *Evolution of a single plane wave and two small side-bands using the one-dimensional fBL equation. Here* $kH = 1.438 > 1.363$, *and growth of the side-bands is observed.*

has been observed experimentally [18] and is thought to be a three-dimensional phenomenon requiring a combination of nonlinear wave modulation and dissipation [25]. We do not perform here detailed calculations of this phenomenon; however, we believe that the equations used here (at least for two-dimensional free surface waves) could be used for this purpose.

4. Resonant interaction of gravity waves. Nonlinear resonance, an important mechanism for the transfer of energy among periodic wave trains, was pioneered by Phillips, Benney, Longuet-Higgins, and others in the 1960s (see below). The basic idea is that two or more distinct wave trains can combine to produce a perturbation with a frequency that corresponds to the natural frequency of a free wave with the same wavenumber. When this occurs, we have resonance, and the amplitude of the response grows linearly. Resonance with three waves, known collectively as a *triad*, is only possible when the dispersion curve has an inflection point (such as in capillary-gravity waves). For pure gravity waves, resonance is possible only among sets of four waves, known as *quartets*.

The idea of resonance for dispersive waves was first suggested by Phillips [24], who showed that three gravity surface waves could resonantly force a fourth wave, forming a quartet. Using the method of multiple scales, Benney [3] derived a coupled set of ordinary differential equations describing the amplitude evolution of a quartet of deep water gravity waves. Bretherton [8] showed that these types of coupled ordinary differential equations could be solved exactly using Jacobi elliptic functions. Experimental confirmation of the existence and importance of resonant water wave interactions was provided by Longuet-Higgins and Smith [17] and McGoldrick et al. [20]. Hammack and Henderson [12] provide a review of experimental results concerning resonant interaction theory for water waves, while the book by Craik [10] gives a comprehensive treatment of wave interactions in general, including triads and quartets in surface waves.

Here we derive a set of equations describing the resonant interaction of four gravity waves in water of finite depth using the fBL equation. The derivation of these "quartet equations" will closely parallel that of the NLS equation in the previous section, except that we will consider the amplitudes of *four* surface waves as well as the induced mean flow. A similar derivation in infinite depth leads to a set of equations whose primary interaction coefficients are zero, indicating that there is *no* quartet interaction in the one-dimensional "deep water" case. This is predicted by the analysis of Longuet-Higgins [16] and shown analytically by Zahkarov [27]. This is not true of two-dimensional infinite-depth gravity waves, nor of the one-dimensional *finite-depth* waves, which we consider here (although the quartet coefficients vanish for $|k| \to \infty$).

4.1. Derivation of quartet/mean-flow equations. Beginning with the one-dimensional fBL equation (2.1), we proceed with the method of multiple scales as before, obtaining (3.5). Again, we expand u in the small parameter ϵ as

$$u = u_0 + \epsilon u_1 + \epsilon^2 u_2 + \cdots \tag{4.1}$$

but now consider a set of plane waves of slowly varying amplitude:

$$u_0(x,t,X,T,\tau) = B(X,T,\tau) + \sum_{j=1}^{4} A_j(X,T,\tau) e^{i(k_j x - \omega_j t)} + *. \tag{4.2}$$

Furthermore, we assume that the four plane waves form a resonant quartet satisfying (1.1)–(1.2). Using the notation $\theta = kx - \omega(k)t$, the resonance condition is $\theta_1 + \theta_2 = \theta_3 + \theta_4$. Substitution into (3.5) again leads to a series of equations at various orders of ϵ. The $O(\epsilon)$ equation is (3.8), and we introduce four frames $\xi_j = X - c_g(k_j)T$ moving at the four group velocities $c_g(k_j)$ and assume that $A_j = A_j(\xi_j, \tau)$.

The quadratic term $\mathcal{N}_1$ is the product of two sums of eight terms each (four plane waves and their conjugates). We need to keep track of only a subset of the sixty-four possible quadratic terms, since we are interested in only those terms that can combine to form quartets at the next order. We will ignore the creation of the conjugate modes, since they are derivable from the main result for the primary modes. For example, at this order we need to account for the $\theta_1 + \theta_2$ term, since $\theta_4 = \theta_1 + \theta_2 - \theta_3$, but can ignore the $-\theta_1 - \theta_2$ term, since this is used only in forming the conjugate of the $e^{i\theta_4}$ wave. With this in mind, the $O(\epsilon)$ problem can be written

$$u_{1tt} + \mathcal{L} u_1 = -i \sum_{a,b} G(a,b) A_a A_b e^{i(\theta_a + \theta_b)}, \tag{4.3}$$

in which

$$G(a,b) = \omega_a (k_b^2 - \omega_b^2 \hat{\mathcal{L}}(k_b)) + 2\omega_b (k_a k_b - \hat{\mathcal{L}}(k_a)\hat{\mathcal{L}}(k_b)) \tag{4.4}$$

and $a, b \in \{1, -1, 2, -2, 3, -3, 4, -4\}$, with the notation $k_{-1} = -k_1$, $\omega_{-1} = -\omega_1$, $A_{-1} = A_1^*$, etc. Here the sum is over those combinations of (a,b) that are relevant to forming quartets, and it varies for which of the four waves is being created.

A particular solution to the $O(\epsilon)$ equation (4.3) is

$$u_1 = -i \sum_{a+b \neq 0} \frac{G(a,b)}{\hat{\mathcal{L}}(k_a + k_b) - (\omega_a + \omega_b)^2} A_a A_b e^{i(\theta_a + \theta_b)}, \tag{4.5}$$

where, again, the details of the sum (discussed below) depend on the primary wave being formed. The restriction on the sum ($a + b \neq 0$) is added because both numerator and denominator vanish (there is no mean flow generated at this order). Proceeding to $O(\epsilon^2)$ terms, the equation is

$$\begin{aligned} u_{2tt} + \mathcal{L} u_2 = &\left(2u_{1tT} - i \frac{\partial \hat{\mathcal{L}}}{\partial k} u_{1X} \right) - \left(2u_{0t\tau} + u_{0TT} - \frac{1}{2} \frac{\partial^2 \hat{\mathcal{L}}}{\partial k^2} u_{0XX} \right) \\ &- \left(\mathcal{N}_1(u_0, u_1) + \mathcal{N}_1(u_1, u_0) + \mathcal{M}(u_0, u_0) + \mathcal{N}_2(u_0, u_0, u_0) \right). \end{aligned} \tag{4.6}$$

In this equation, we seek to eliminate the secular terms in $e^{i\theta_j}, j = 1, \ldots, 4$, and the zero-mode terms (e^{i0}). Upon transforming to the moving frame $\xi_j = X - c_g(k_j)T$, the linear terms on the right-hand side reduce to

$$\left(-2i\omega_j A_{j\tau} - \omega_j \frac{\partial c_g}{\partial k}(k_j) A_{j\xi_j\xi_j}\right) e^{i\theta_j} + B_{TT} - B_{XX}, \tag{4.7}$$

where $j = 1, \ldots, 4$.

On the right-hand side, the terms $\mathcal{N}_1(u_0, u_1) + \mathcal{N}_1(u_1, u_0)$ will yield cubic terms, since u_1 contains quadratic terms (4.5) and u_0 has the original plane waves. We are interested in only combinations that yield a member of the quartet, i.e., terms in $e^{i\theta_j}$. No e^{i0} terms are created here. The relevant contributions of the terms $\mathcal{N}_1(u_0, u_1) + \mathcal{N}_1(u_1, u_0)$ can be written

$$\sum_{a,b,c} \frac{G(a,b)\,[G(a+b,c) + G(c,a+b)]}{\hat{\mathcal{L}}(k_a + k_b) - (\omega_a + \omega_b)^2} A_a A_b A_c e^{i(\theta_a+\theta_b+\theta_c)} + *, \tag{4.8}$$

in which we keep the notation $a, b, c \in \{1, -1, 2, -2, 3, -3, 4, -4\}$. Note that c always comes from the u_0 term, and a and b come from the u_1 term that was solved at $\mathrm{O}(\epsilon)$. Also, we keep the restriction that $a + b \neq 0$. The notation $G(a + b, c)$ means to use $k_a + k_b$ and $\omega_a + \omega_b$ for k_a and ω_a in the expression (4.4).

The term $e^{i(\theta_a+\theta_b+\theta_c)}$ will be equal to $e^{i\theta_j}$ for one of the original θ_j when a, b, and c are chosen appropriately. For example, to form terms in $e^{i\theta_1}$, we are interested in the six permutations of the set $\{-2, 3, 4\}$ since $\theta_1 = -\theta_2 + \theta_3 + \theta_4$. Evaluating the sum in (4.8) with these six sets of a, b, c yields the term $q_1 A_2^* A_3 A_4 e^{i\theta_1}$, where q_1 is a coefficient. However, we must also consider permutations of the sets $\{1, 1, -1\}$, $\{1, 2, -2\}$, $\{1, 3, -3\}$, and $\{1, 4, -4\}$, since they also lead to terms in $e^{i\theta_1}$. (In using these values of a, b, c, however, care must be taken to avoid duplicates and the cases when $a + b = 0$.) Thus, the contribution of $\mathcal{N}_1(u_0, u_1) + \mathcal{N}_1(u_1, u_0)$ to the quartet equations will be

$$\left(q_j A_l^* A_m A_n + \sum_{k=1}^{4} p_{jk} |A_k|^2 A_j\right) e^{i\theta_j} \tag{4.9}$$

for $j = 1, \ldots, 4$, where l, m, and n depend on j and satisfy $\theta_j + \theta_l = \theta_m + \theta_n$.

The contributions of the cubic terms $\mathcal{N}_2(u_0, u_0, u_0)$ are very similar to those of the quadratic terms $\mathcal{N}_1(u_0, u_1) + \mathcal{N}_1(u_1, u_0)$, except that they come from u_0 directly. We again sum over five distinct groups of six permutations and obtain the same twenty terms as in (4.9), but with different coefficients. We add these coefficients to those obtained from the quadratic terms. The contribution of $\mathcal{N}_2(u_0, u_0, u_0)$ is

$$\sum_{a,b,c} H(a,b,c) A_a A_b A_c e^{i(\theta_a+\theta_b+\theta_c)} + *, \tag{4.10}$$

in which

$$\begin{aligned} H(a,b,c) = {} & \omega_a\omega_b\hat{\mathcal{L}}(k_b)(k_c^2 + \hat{\mathcal{L}}^2(k_c)) + 2\omega_a\omega_b\hat{\mathcal{L}}(k_c)(k_b\omega_b - k_c\omega_b - k_c\omega_c) \\ & + \frac{1}{2}k_a^2(3k_bk_c + \hat{\mathcal{L}}(k_b)\hat{\mathcal{L}}(k_c)) + \frac{1}{2}\hat{\mathcal{L}}^2(k_a)(\hat{\mathcal{L}}(k_b)\hat{\mathcal{L}}(k_c - k_bk_c) - 2k_ak_c\hat{\mathcal{L}}(k_b)\hat{\mathcal{L}}(k_c)). \end{aligned}$$

Note that, like the quadratics, the cubics do not contribute any terms in e^{i0}.

Finally, $\mathcal{M}(u_0, u_0)$ gives both terms in $e^{i\theta_j}$ and "mean-flow" terms in e^{i0}:

$$\begin{aligned}&\sum_{j=1}^{4}(-2k_j\omega_j - c_g(k_j)(k_j^2 - \hat{\mathcal{L}}^2(k_j)))(A_jA_j^*)_{\xi_j}\\ &+\sum_{j=1}^{4}(2k_j\omega_j B_X + (\hat{\mathcal{L}}^2(k_j) - k_j^2)B_T)A_j e^{i\theta_j}.\end{aligned} \tag{4.11}$$

Equating terms from (4.6) in like powers of $e^{i\theta}$ leads to a coupled set of *five* partial differential equations governing the evolution of the resonant quartet and the "mean-flow" term B:

$$B_{TT} - B_{XX} = \sum_{j=1}^{4}(2k_j\omega_j + c_g(k_j)(k_j^2 - \hat{\mathcal{L}}^2(k_j)))(A_jA_j^*)_{\xi_j}, \tag{4.12}$$

$$\begin{aligned}iA_{j\tau} + \frac{1}{2}\frac{\partial c_g}{\partial k}(k_j)A_{j\xi_j\xi_j} &= \frac{\alpha_j}{2\omega_j}A_l^*A_mA_n + \frac{1}{2\omega_j}\left(\sum_{k=1}^{4}\beta_{jk}|A_k|^2A_j\right)\\ &+ \frac{1}{2\omega_j}(2k_j\omega_jB_X + (\hat{\mathcal{L}}^2(k_j) - k_j^2)B_T)A_j\end{aligned} \tag{4.13}$$

for $j = 1, \ldots, 4$, where $\theta_j + \theta_l = \theta_m + \theta_n$. Note that we have intentionally glossed over some notational inconsistency by using X, T, and ξ_j as independent variables in (4.12). Furthermore, the induced mean flows on any particular member of the quartet from the other three members are rapidly varying on the time scale of (4.13), unless group velocities are close. However, we will not be concerned with initial conditions for (4.12) and (4.13) that involve spatial modulation of the four primary waves and will thus treat (4.13) as a set of ordinary differential equations (see below).

Historically, the derivation of quartet equations was done for deep water, for which *spatial* modulation effects are ignored since the mean-flow is known to be zero. Thus (4.12) would not be present, $B = 0$ in (4.13), and these equations become ordinary differential equations. Of course the dispersion relation $\omega^2 = \hat{\mathcal{L}}$ also changes for deep water. As noted by Bretherton [8] for the two-dimensional deep-water case, the primary interaction coefficients α_i turn out to be *equal*. This is also true for finite-depth quartet equations, which we have confirmed using the fBL model. For deep water we find $\alpha_i = 0$, as expected. Since the α_i are primarily responsible for the exchange of energy among the four waves (the β_{jk} modify the period and amplitude), there is no interaction in the one-dimensional deep-water case.

4.2. Quartet simulations. To verify that (4.12) and (4.13) correctly capture the finite-depth quartet interactions, we compare the solutions to these equations to a simulation using the fBL equation. Since computing with these five coupled partial differential equations is computationally intensive, we seek a simpler special case. By choosing an initial condition in which the amplitude of the four primary waves is not spatially modulated, we ignore the second term in (4.13) and (4.12) altogether, leaving a set of four coupled ordinary differential equations. Although an exact solution involving Jacobi elliptic functions is known for a set of ordinary differential equations of this form (see [8]), we solve them numerically using a fourth-order Runge–Kutta method, while we evolve the fBL equation in time using a pseudospectral method. The

initial condition for both computations is the same. We pick four plane waves among a discrete set of wavenumbers that satisfy the resonance conditions $k_1 + k_2 = k_3 + k_4$ and $\omega_1 + \omega_2 = \omega_3 + \omega_4$, and give them each an initial amplitude. The quartet equations govern the four amplitudes as a function of $\tau = \epsilon^2 t$, while the simulation of the fBL equation computes the evolution of *all* the Fourier modes.

The pseudospectral method used here was developed by Milewski and Tabak [23] and involves the factoring of the fBL equation. Equation (2.1) can be also written as

$$(\partial_{tt} + L^2)u = \mathcal{G}(u), \tag{4.14}$$

in which $L^2 = \mathcal{L} = (-\partial_{xx})^{\frac{1}{2}} \tanh[(-\partial_{xx})^{\frac{1}{2}}]$ and $\mathcal{G}(u) = -\epsilon\mathcal{N}_1(u,u) - \epsilon^2\mathcal{N}_2(u,u,u)$. We factor the left-hand side by introducing $U(x,t) = (\partial_t - iL)u(x,t)$ and recast the equation in terms of U as

$$U_t + iLU = \mathcal{G}(U). \tag{4.15}$$

Thus the free surface is, to leading order, $\eta = -u_t = -Re(U)$. To solve (4.15), we transform it to Fourier space, introduce an integrating factor, and numerically integrate using a Runge–Kutta scheme. Since we compute with U directly and not u, we choose to initially set the quartet amplitudes to have equal values *in terms of* U. The conversion to the amplitudes of u is straightforward. If $u(x,t) = \sum_{j=1}^{4} A_j e^{i\theta_j} + *$, then $U(x,t) = \sum_{j=1}^{4} -2i\omega_j A_j e^{i\theta_j}$ and $\eta(x,t) = \sum_{j=1}^{4} i\omega_j A_j e^{i\theta_j} + *$. In the figures below, we graph the absolute value of the relevant Fourier modes of $U(x,t)$, i.e., $|U_j| = 2\omega_j |A_j|$.

A slight modification to the fBL equation (2.1) must be made before using our pseudospectral method. The problem lies with the $O(\epsilon)$ quadratic terms which have the term $-u_t \mathcal{L} u_{tt}$. Because of this term, we cannot integrate (2.1) in the form given. With the substitution $u_{tt} = -\mathcal{L}u - \epsilon\mathcal{N}_1(u,u) + O(\epsilon^2)$, the quadratic term becomes

$$\bar{\mathcal{N}}_1(u,u) = 2u_x u_{xt} + 2\mathcal{L}u\mathcal{L}u_t + u_t u_{xx} + u_t \mathcal{L}^2 u, \tag{4.16}$$

and there is an additional cubic term in the equation which becomes

$$u_{tt} + \mathcal{L}u + \epsilon\bar{\mathcal{N}}_1(u,u) + \epsilon^2 \left(\mathcal{N}_2(u,u,u) + u_t \mathcal{L}[\bar{\mathcal{N}}_1(u,u)]\right) = 0. \tag{4.17}$$

These modifications are similar to the formal manipulations that one uses to "regularize" the Korteweg–de Vries (KdV) equation and obtain the Benjamin–Bona–Mahoney (BBM) equation. There are also corresponding changes to the details of the quartet equations, in particular to the definitions of the functions $G(a,b)$ and $H(a,b,c)$ in (4.4) and (4.11), respectively. The new cubic term adds

$$\omega_a\omega_b(k_c^2 - \hat{\mathcal{L}}(k_c))\hat{\mathcal{L}}(k_b + k_c) + 2\omega_a\omega_c(k_b k_c - \hat{\mathcal{L}}(k_b)\hat{\mathcal{L}}(k_c))\hat{\mathcal{L}}(k_b + k_c) \tag{4.18}$$

to the function H.

Figure 4.1 shows the numerical solution of the four coupled quartet equations using a fourth-order Runge–Kutta scheme. We use the quartet wavenumbers $(k_1, k_2, k_3, k_4) = (81, 46, 142, -15)\Delta k$, where $\Delta k = \frac{1}{64}$, and the corresponding frequencies ω_j to precompute the twenty quartet coefficients. The initial amplitude of each mode U_j is 0.2. For this quartet, energy is periodically exchanged between the four waves which is inherently on the $\tau = \epsilon^2 t$ time scale. The total energy, given by $\sum_{j=1}^{4} \frac{1}{\alpha_j}|A_j|^2$, remains constant.

Figure 4.2 shows the pseudospectral simulation of the fBL equation initialized with energy only in the same four wavenumbers considered above. We use a total of

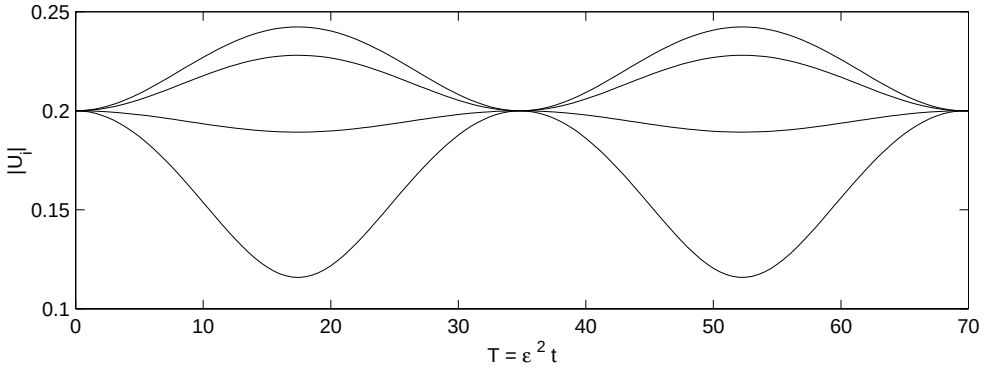


FIG. 4.1. *Numerical solution of the quartet equations using a fourth-order Runge–Kutta scheme. The wavenumbers are* $(k_1, k_2, k_3, k_4) = (81, 46, 142, -15)\Delta k$, *where* $\Delta k = \frac{1}{64}$. *The initial amplitude is* $U_j = 0.2$.

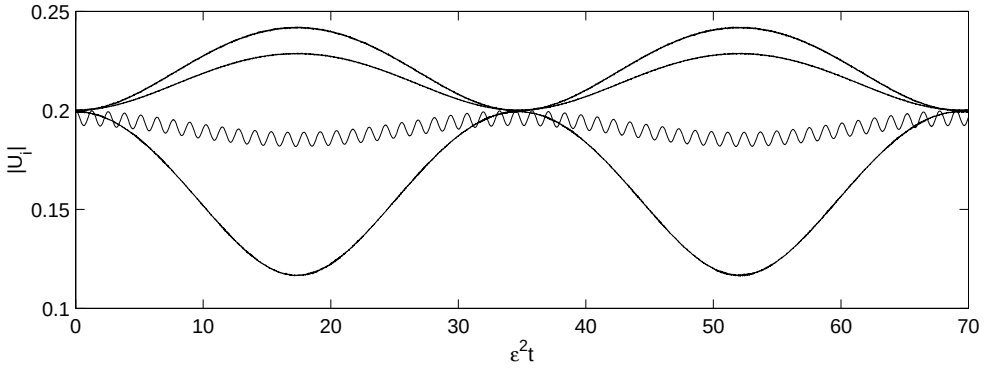


FIG. 4.2. *Simulation of the fBL equation initialized with four waves of the same initial amplitude. The wavenumbers are* $(k_1, k_2, k_3, k_4) = (81, 46, 142, -15)\Delta k$, *where* $\Delta k = \frac{1}{64}$. *The initial amplitude is* $U_j = 0.2$, $\epsilon = .05$, *and* $\Delta t = 0.1$. *We use* 1024 *wavenumbers in this computation.*

1024 wavenumbers with the initial amplitude of each member of the quartet being $U_j = 0.2$. Note that energy is periodically exchanged between the four waves on the long time scale $\tau = \epsilon^2 t$, as predicted (here $\epsilon = .05$). On a shorter time scale, the longest wave in the quartet $k_4 = -15/64$ periodically exchanges energy with its near-resonant second harmonic, the $k = -30/64$ mode. This accounts for the smaller oscillations in the amplitude of this mode. In the next section, we derive the equations governing this interaction and show that they can be combined with the quartet equations to correctly predict the simulation results.

Quartets containing wavenumbers closer to the shallow water regime ($kH < 1$) will exhibit prominent second harmonic interaction, as the dispersion curve is nearly linear in this range. This draws energy from the primary quartet and may account for the slight variation in period between the two quartet graphs. Quartets without this second-harmonic interaction do not exist for this one-dimensional model because quartets containing many larger wavenumbers ($kH > 1$) are very weakly coupled since the $\alpha_j \to 0$ as $H \to \infty$, and other mechanisms such as nonresonant interactions and Benjamin–Feir instability are relatively more significant.

4.3. Derivation of second-harmonic interaction. Beginning with the one-dimensional fBL equation (2.1), we proceed with the method of multiple scales. We restrict our attention to the slow time scale $T = \epsilon t$ since we expect the interaction to occur on this scale. We also ignore slow spatial variation, consistent with our integration of the quartet equations above. With the substitution $\partial_t \to \partial_t + \epsilon\partial_T$ we have the following equation for $u(x,t,T)$:

$$u_{tt} + \mathcal{L}u + \epsilon\left(2u_{tT} + \mathcal{N}_1(u,u)\right) = 0, \tag{4.19}$$

where higher-order terms are unnecessary. Next, we expand u as $u = u_0 + \epsilon u_1 + \cdots$ with

$$u_0(x,t,T) = A_1(T)e^{i(kx-\omega t)} + A_2(T)e^{i(2kx-\omega(2k)t)} + *. \tag{4.20}$$

With the notation $\omega_1 = \omega(k)$, $\omega_2 = \omega(2k)$, $\theta_1 = kx - \omega_1 t$, and $\theta_2 = 2kx - \omega_2 t$, the balance of terms at $\mathrm{O}(\epsilon)$ in (4.19)

$$2\omega_1 A_{1T}e^{i\theta_1} + 2\omega_2 A_{2T}e^{i\theta_2} + * = \sum_{a,b} G(a,b)A_a A_b e^{i(\theta_a+\theta_b)}, \tag{4.21}$$

in which the function $G(a,b)$ is the same as that derived before for the quadratic terms (4.4). We can only get terms in e^{2ik} with $a = b = 1$, for which the right-hand side becomes $G(1,1)A_1^2 e^{i(2kx-2\omega_1 t)} = G(1,1)A_1^2 e^{i\theta_2}e^{-i\Delta t}$, where the frequency mismatch is $\Delta = 2\omega_1 - \omega_2$. In a similar way, we can create terms in e^{ik} with $(a,b) = (-1,2)$ or $(2,-1)$, giving a right-hand side of $(G(-1,2) + G(2,-1))A_1^* A_2 e^{i\theta_1}e^{i\Delta t}$. Thus the wave-second-harmonic interaction equations are

$$\frac{dA_1}{dT} = \delta_1 A_1^* A_2 e^{i\Delta t}, \qquad \delta_1 = \frac{G(-1,2)+G(2,-1)}{2\omega_1} < 0, \tag{4.22}$$

$$\frac{dA_2}{dT} = \delta_2 A_1^2 e^{-i\Delta t}, \qquad \delta_2 = \frac{G(1,1)}{2\omega_2} > 0. \tag{4.23}$$

For $k \ll 1$, $\Delta = k^3$, $\delta_1 = -3k^2$, and $\delta_2 = (3/2)k^2$. The transformations $A_1 \to A_1 e^{i\Delta t}$, $A_2 \to A_2 e^{i\Delta t}$ remove the periodic coefficient (detuning term), yielding

$$\frac{dA_1}{dT} = -i\Delta A_1 + \delta_1 A_1^* A_2 e^{i\Delta t}, \qquad \frac{dA_2}{dT} = -i\Delta A_2 + \delta_2 A_1^2 e^{-i\Delta t}. \tag{4.24}$$

These equations for $\epsilon = O(\Delta)$ can be solved analytically. Writing $A_1 = \rho_1 e^{i\phi_1}$, $A_2 = \rho_2 e^{i\phi_2}$, the equations (4.22), (4.23) conserve

$$E = -\frac{1}{\delta_1}\rho_1^2 + \frac{1}{\delta_2}\rho_2^2, \tag{4.25}$$

$$H = \rho_1^2\rho_2\sin(\phi_2 - 2\phi_1) - \frac{\Delta}{4}\left(\frac{1}{\delta_1}\rho_1^2 + \frac{1}{\delta_2}\rho_2^2\right), \tag{4.26}$$

where H is the Hamiltonian in appropriate coordinates. From these, one can conclude $|A_1|^2 = \delta_1 I(t) + c_1$, $|A_2|^2 = \delta_2 I(t) + c_2$, where

$$\left(\frac{dI}{dT}\right)^2 = 4(\delta_1 I + c_1)^2(\delta_2 I + c_2) - 4\left(H + \frac{\Delta}{2\delta_2}(\delta_2 I + c_2)\right)^2. \tag{4.27}$$

The solution can be written in terms of elliptic functions. For the results described below, $I(0) = 0$, $c_1 = -\delta_1 E$, $c_2 = 0$, and $H + (\Delta/4)E = 0$.

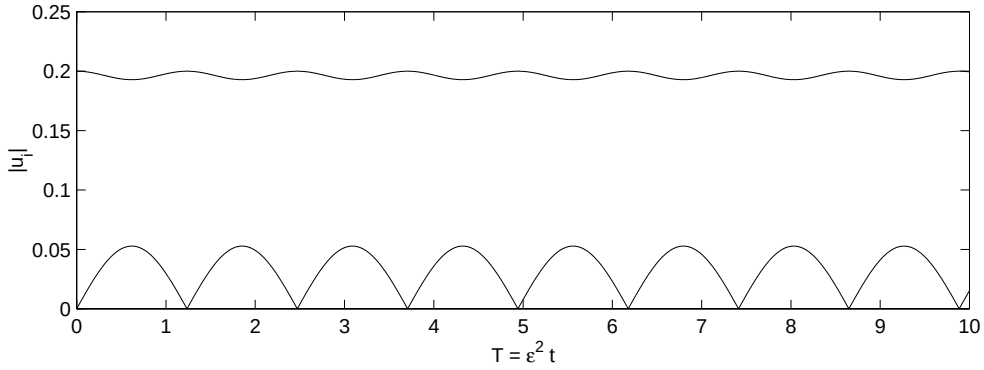


FIG. 4.3. *Solution of the second-harmonic interaction equations with* $k = -15/64$. *The initial amplitude is* $U_1 = 0.2$ *for the primary mode and* $U_2 = 0$ *for the second harmonic.* ($\epsilon = .05$.)

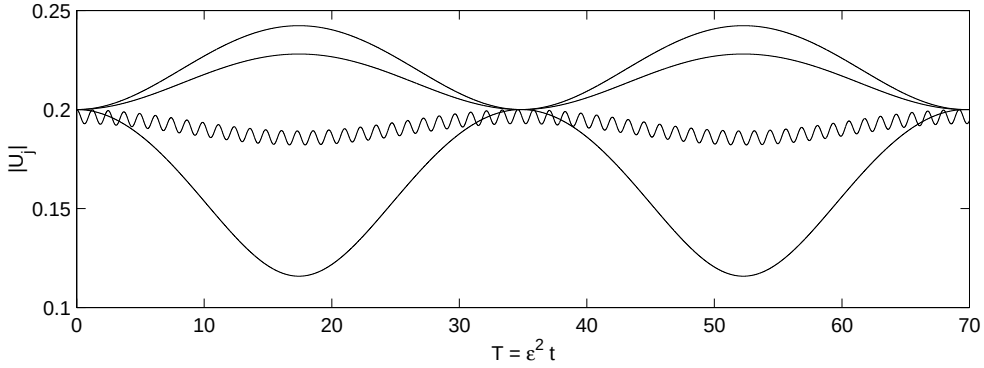


FIG. 4.4. *Numerical solution of the quartet equations with second-harmonic interaction using a fourth-order Runge–Kutta scheme. The wavenumbers are* $(k_1, k_2, k_3, k_4) = (81, 46, 142, -15)\Delta k$, *where* $\Delta k = \frac{1}{64}$. *The initial amplitude is* $U_j = 0.2$.

For the quartet of waves that we consider here, only the mode $k_4 = -\frac{15}{64}$ will generate significant second harmonic energy. Figure 4.3 shows the solution of (4.22), (4.23). The primary mode is $k = -\frac{15}{64}$ with initial amplitude $U_1 = 0.2$, as in the quartet simulation. The $2k$ mode has zero initial amplitude. (Note that, for consistency, we show the results in terms of our computational variable $U = (\partial_t - iL)u$, as discussed above.)

A pseudospectral simulation of the fBL equation initialized with energy in only the single mode $k = -\frac{15}{64}$ yields a virtually identical result.

Finally we augment the quartet equations (4.13) with the second-harmonic interaction term (for the $k_4 = -\frac{15}{64}$ mode only) and the second harmonic equation (4.23). Although these equations combine two time scales and thus are not formally correct, they give results virtually identical (see Figure 4.4) to those of the fBL simulation of Figure 4.2.

5. Gravity wave turbulence simulations. Since we have shown that the one-dimensional fBL equation captures the deterministic dynamics of the water wave problem, we turn our attention to the simulation of dispersive wave turbulence using

this equation. Statistical dispersive wave-turbulence theory relies on a closure (see [14], [7], [5], [29]) that, in essence, restricts the dynamics to the resonant set of waves satisfying (1.1), (1.2). Briefly, the closure is based on writing (4.15) in Fourier space,

$$\hat{U}_t + i\omega\hat{U} = \int Q(k_1,k_2,k_3,k)\hat{U}_1\hat{U}_2\hat{U}_3\delta(k_1+k_2+k_3-k)dk_1dk_2dk_3, \tag{5.1}$$

where $\hat{U}_1 = \hat{U}(k_1,t)$. The expression for the "collision" kernel Q is essentially the quartet coefficients computed previously. From (5.1) one obtains the equation for the second-order moment $n_k = \langle\hat{U}(k,t)\hat{U}^*(k,t)\rangle$:

$$\frac{dn_k}{dt} = \int 2\ \mathrm{Re}\ Q\langle\hat{U}_1\hat{U}_2\hat{U}_3\hat{U}^*\rangle\delta(k_1+k_2+k_3-k)dk_1dk_2dk_3, \tag{5.2}$$

where $\langle\cdot\rangle$ denotes the ensemble average. Next, one writes the equation for the evolution of the fourth-order moments appearing in the integrand of (5.2) in terms of sixth-order moments. The closure consists in reducing these sixth-order moments to products of second-order moments (a quasi-Gaussianity assumption). This leads to a relation of the form

$$\langle\hat{U}_1\hat{U}_2\hat{U}_3\hat{U}^*\rangle \sim Q\frac{n_2n_3n_k + n_1n_3n_k + n_1n_2n_k - n_1n_2n_3}{i(\omega_1+\omega_2+\omega_3-\omega_k)}. \tag{5.3}$$

Now, substituting (5.3) into (5.2) and replacing the reciprocal of the sum of frequencies by $\delta(\omega_1+\omega_2+\omega_3-\omega_k)$, one obtains a closed equation that concentrates the dynamics on the resonant set. (The difference in the sign in front of k_3 and ω_3 in these delta functions and in (1.1), (1.2) is just a matter of convention.)

The steady state ($\frac{dn_k}{dt} = 0$) of this resulting equation has two types of solutions: solutions in statistical equilibrium and solutions with finite fluxes (cascades) of energy. The latter have been of particular interest in attempts to describe the ocean's wave spectrum.

Since these cascades require that the governing equation (4.15) be forced and dissipated, we augment the equation by adding forcing and dissipation terms (which are meant to model physical processes such as wind forcing, viscous damping, etc.) at various ranges of wavenumbers. Then, from long time computations, we observe the evolution of the energy spectrum until a statistical steady state is reached. Since both energy and wave action are conserved in this system, we must dissipate at both ends of the Fourier spectrum and force at some intermediate scale. Thus, the factored form of the fBL equation (4.15) in Fourier space, with forcing and dissipation, becomes

$$\hat{U}_t + i\hat{L}\hat{U} = \hat{G}(\hat{U}) + \hat{F}, \tag{5.4}$$

in which we define the forcing-dissipation function $\hat{F}$ as

$$\hat{F}(k) = \begin{cases} f_r\hat{U} & \text{for } k_{fl}\Delta k \le |k| \le k_{fh}\Delta k, \\ d_{r1}|k|^{-2}\hat{U} & \text{for } k_{dl}\Delta k \le |k| \le k_{dh}\Delta k, \\ d_{r2}|k|^2\hat{U} & \text{for } |k| \ge K_d\Delta k, \\ 0 & \text{otherwise.} \end{cases} \tag{5.5}$$

Here k_{fl}, k_{fh}, k_{dl}, k_{dh}, and K_d are integers which define the range of forcing and the two dissipation ranges, the latter being $|k| > K_d\Delta k$. The forcing rate f_r is positive, while the dissipation rates d_{r1} and d_{r2} should be negative. The closure theory

described above appears insensitive on the particular form of the forcing, and this is confirmed by the numerical simulations. Various forms of forcing (both deterministic and random) and dissipation ("standard" viscosity and "hyper" viscosity) were experimented with, and the results did not change appreciably. Our approach is similar to that of Majda and coworkers (see [19] and [9]), who perform computations with a simpler NLS-like model equation.

6. Direct cascades. To generate a direct (or forward) cascade over a significant range, we force at low wavenumbers and dissipate at both the lowest and highest wavenumbers. We construct the experiment such that the *finite-depth* regime lies in the inertial range (the range of wavenumbers that are neither forced nor dissipated) for reasons mentioned in the introduction. Figure 6.1 compares the dispersion relations of the shallow water ($\omega = k$), infinite-depth ($\omega = |k|^{\frac{1}{2}}$), and arbitrary-depth ($\omega = (|k| \tanh |k|)^{\frac{1}{2}}$) problems. We will arbitrarily denote the range $0.5 < k < 2.5$ as the finite-depth regime and indicate this range in our numerical result.

We compute the correlation function

$$p(k) = \overline{\hat{u}(k,t)\hat{u}^*(k,t)}, \tag{6.1}$$

where the overbar denotes time average (after a statistical steady state is reached). It can be shown that

$$p(k) \sim \frac{1}{2\pi} \int_{-\infty}^{\infty} \overline{u(x,t)u(x+r,t)} e^{-ikr} dr, \tag{6.2}$$

when correlations are spatially independent.

Figure 6.2 shows a typical weak turbulence spectrum that we obtain using the one-dimensional fBL model. The computation uses 2048 dealiased modes with $\Delta k = \frac{1}{100}$. The parameters chosen are $\epsilon = 0.05$, $f_r = 0.00001$, $d_{r1} = -0.0009$, and $d_{r2} = -0.01$. Initially, the system has significant energy only in the lower wavenumbers $\hat{U}(4\Delta k \leq k \leq 12\Delta k) = 0.5$. All other modes are initialized to $\hat{U}(k) = 0.00001$. We show the average spectrum from $t = 150,000$ to $t = 200,000$, computed with the methods described earlier with $\Delta t = 0.1$.

There are six regions of the spectrum divided by five vertical dotted lines. From left to right, these are (1) the low wavenumber dissipation range, (2) the forcing range, (3) the "shallow" inertial range, (4) the "finite depth" inertial range, (5) the "infinite depth" inertial range, (6) the high wavenumber dissipation range.

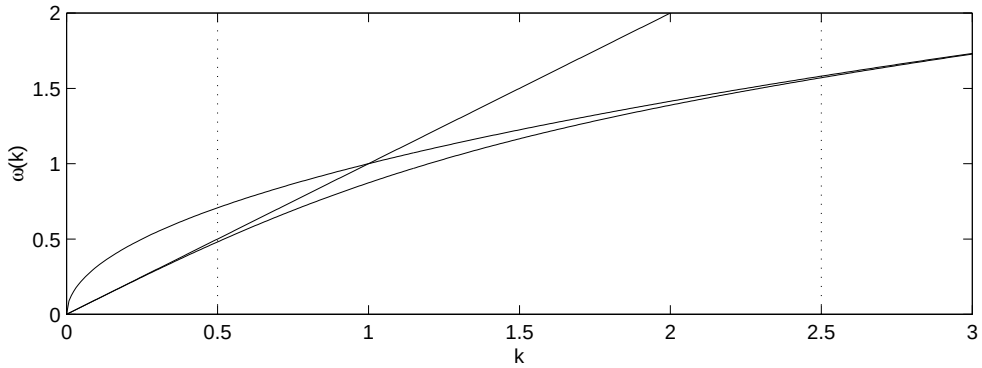


FIG. 6.1. *Finite-depth dispersion relation* $\omega(k) = (|k| \tanh(|k|))^{\frac{1}{2}}$. *As indicated in the figure,* $\omega \sim |k|$ *as* $k \to 0$, *and* $\omega \sim (|k|)^{1/2}$ *for* k *large. The vertical lines in the figure reflect an arbitrary choice for the transition region* $(0.5 < k < 2.5)$ *between these two power laws.*

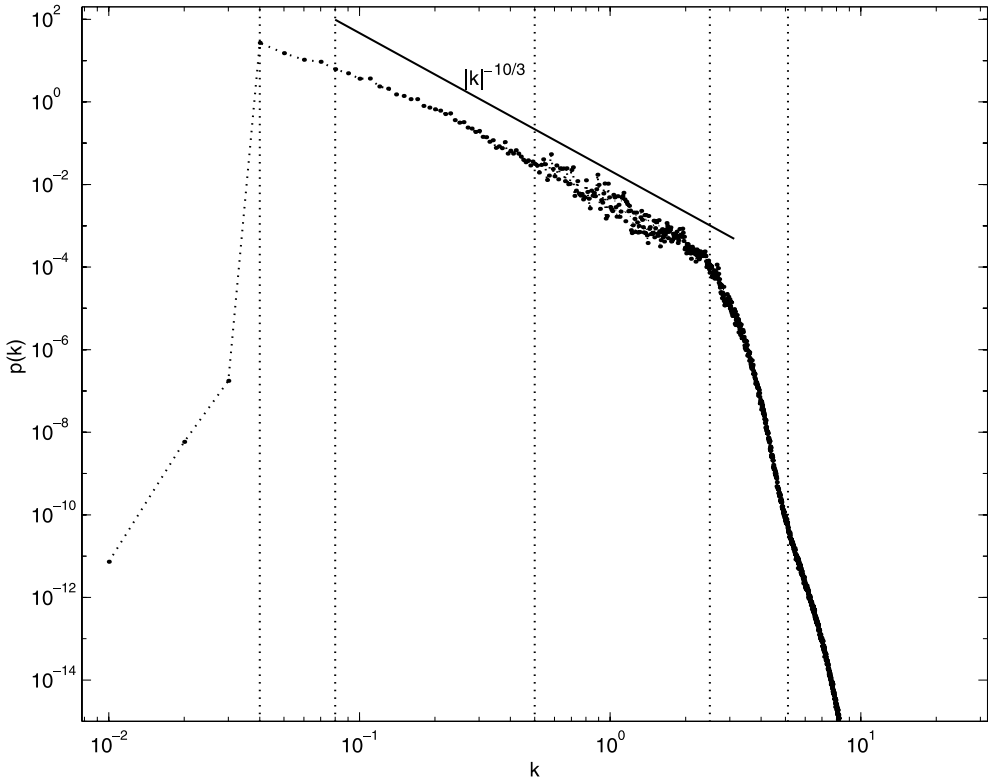


FIG. 6.2. Experiment 1. *Direct cascade using the fBL model with forcing between* $k = 4\Delta k$ *and* $k = 8\Delta k$ *and dissipation from* $k = \Delta k$ *to* $k = 3\Delta k$ *and for* $k > 512\Delta k$. *Here* $\Delta k = \frac{1}{100}$ *and* $\epsilon = 0.05$. *We estimate* $p(k)$ *by the time average of* $\hat{u}(k,t)\hat{u}^*(k,t)$ *for* $t = 150,000$ *to* $t = 200,000$ *with data every* $t = 500$.

We note that some features of the spectrum seem to change in correlation with the shape of the dispersion curve. In the shallow and finite-depth regimes, there is a good agreement with the weak turbulence theory of Zakharov [27] described above. He predicts a direct cascade of $p(k) \sim |k|^{-10/3}$ (in the present variables), subject to some strict conditions on the wave amplitudes (which are not strictly satisfied in the computations). Using least-square interpolation of the data yields $p(k) \sim |k|^{\alpha}$ with $-3 > \alpha > -3.4$, depending on where the endpoints of the inertial range are chosen. The Zakharov slope is shown in Figure 6.2 for comparison. We also note that in the finite-depth region (region 4) the data is more spread. This is probably because in this regime the discrete quartets are sparser, whereas in shallower water, wave interaction is denser, and nondispersive wave steepening plays a more important role.

6.1. Inverse cascade. To obtain an inverse cascade (from high to low wavenumbers) we modified the forcing and dissipation parameters from the previous experiment. We force near the deep water regime, between wavenumbers 2.25 and 2.50. Again, we use 2048 dealiased modes, now with $\Delta k = \frac{1}{200}$, $\epsilon = 0.05$, $f_r = 0.0006$, $d_{r1} = -0.75$, and $d_{r2} = -0.50$. Initially, all modes are initialized to $\hat{U}(k) = 0$. We show the average spectrum from $t = 150,000$ to $t = 200,000$, computed with the methods described earlier with $\Delta t = 0.1$.

The results are shown in Figure 6.3. The results here are less clear. In the shallow water regime there appears to be a region with $p(k) \sim |k|^{\alpha}$, with $-2.2 > \alpha > -2.4$.

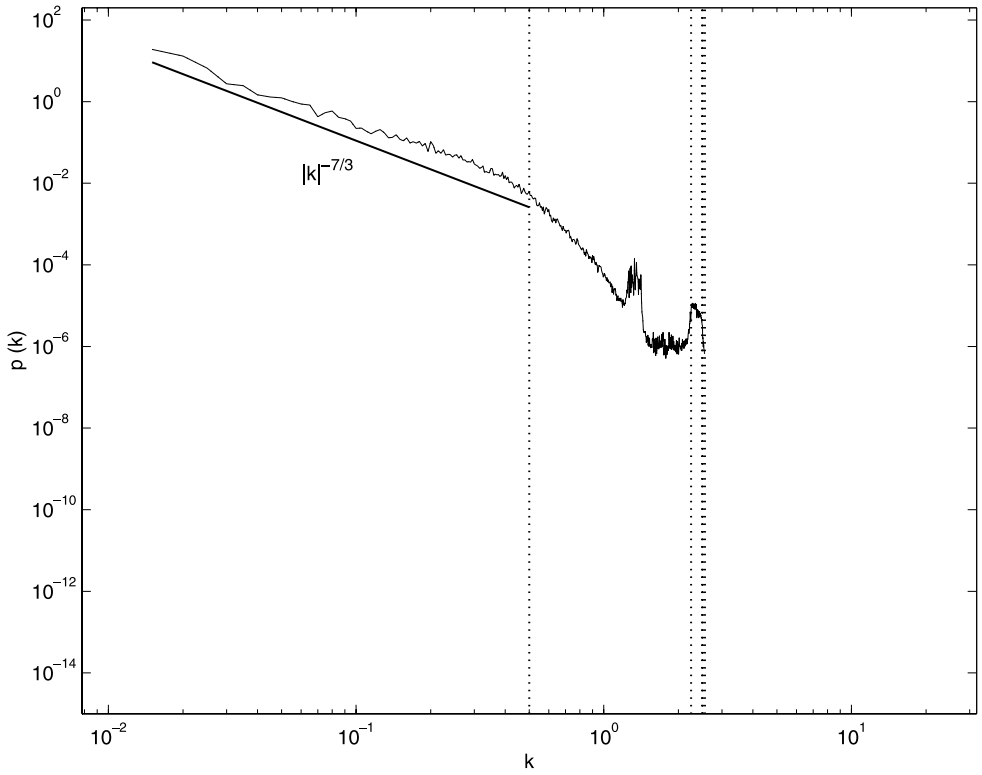


FIG. 6.3. *Inverse cascade using the fBL model with forcing between* $k = 450\Delta k$ *and* $k = 500\Delta k$ *and dissipation from* $k = \Delta k$ *to* $k = 2\Delta k$ *and for* $k > 512\Delta k$. *Here* $\Delta k = \frac{1}{200}$ *and* $\epsilon = 0.05$. *We estimate* $p(k)$ *by the time average of* $\hat{u}(k,t)\hat{u}^*(k,t)$ *for* $t = 250,000$ *to* $t = 300,000$ *with data every* $t = 1000$.

(We show a slope of $\alpha = -7/3$ for reference.) Zakharov's [27] prediction for the shallow water inverse cascade is $|k|^{-3.0}$. The reasons for this difference may be related to the generation of coherent structures (solitons) which are excluded from the theory (by assuming sufficiently small amplitudes compared to dispersive effects). In fact, the more recent work on NLS [9], [30] explores the role of coherent structures in the various spectra observed.

At finite depth, our computed spectrum drops much more steeply. The two visible peaks in the spectrum are due to the forcing: the peak at higher wavenumbers is over the forcing region, and the second peak is a direct subharmonic generation from the forced modes.

7. Conclusion. We have derived a Benney–Luke model for waves in arbitrary depth and verified its utility by demonstrating its accuracy in important deterministic water wave phenomena: Benjamin–Feir wave packet instability, resonant quartet interactions, and harmonic generation in shallow water. We have then used the model, together with forcing and dissipation, to simulate wave turbulence. The numerical spectra that we obtain agree with Zakharov's prediction for the direct cascades but not for inverse cascades. Possible reasons for the departure from Zakharov's prediction include the narrow range of applicability of his theory in this regime to avoid solitons. The present work validates the use of the fBL equation for the more interesting problem of two-dimensional turbulent simulations.

REFERENCES

[1] T. B. Benjamin, *Instability of periodic wave trains in nonlinear dispersive systems*, Proc. Roy. Soc. London A, 299 (1967), pp. 59–75.

[2] T. B. Benjamin and J. E. Feir, *The disintegration of wave trains on deep water, Part* 1, J. Fluid Mech., 27 (1967), pp. 417–430.

[3] D. J. Benney, *Nonlinear gravity wave interactions*, J. Fluid Mech., 14 (1962), pp. 577–584.

[4] D. J. Benney and J. C. Luke, *Interactions of permanent waves of finite amplitude*, J. Math. Phys., 43 (1964), pp. 309–313.

[5] D. J. Benney and A. C. Newell, *Random wave closures*, Stud. Appl. Math., 1 (1969), pp. 39–53.

[6] D. J. Benney and G. J. Roskes, *Wave instabilities*, Studies Appl. Math., 48 (1969), pp. 377–385.

[7] D. J. Benney and P. G. Saffman, *Nonlinear interaction of random waves in a dispersive medium*, Proc. Roy. Soc. A, 289 (1965), pp. 301.

[8] F. P. Bretherton, *Resonant interactions between waves. The case of discrete oscillations*, J. Fluid Mech., 20 (1964), pp. 457–479.

[9] D. Cai, A. J. Madja, D. W. McLaughlin, and E. G. Tabak, *Spectral bifurcations in dispersive wave turbulence*, Proc. Natl. Acad. Sci. USA, 96 (1999), pp. 14216–14221.

[10] A. D. D. Craik, *Wave Interactions in Fluid Flows*. Cambridge University Press, Cambridge, UK, 1985.

[11] A. Davey and K. Stewartson, *On three-dimensional packets of surface waves*, Proc. R. Soc. London A, 338 (1974), pp. 101–110.

[12] J. L. Hammack and D. M. Henderson, *Resonant interactions among surface water waves*, Ann. Rev. Fluid Mech., 25 (1993), pp. 55–97.

[13] H. Hasimoto and H. Ono, *Nonlinear modulation of gravity waves*, J. Phys. Soc. Japan, 33 (1972), pp. 805–811.

[14] K. Hasselmann, *On the nonlinear energy transfer in a gravity wave spectrum*, J. Fluid Mech., 12 (1962), pp. 481–500.

[15] M. J. Lighthill, *Contributions to the theory of waves in nonlinear dispersive systems*, J. Inst. Math. Appl., 1 (1965), pp. 269–306.

[16] M. S. Longuet-Higgins, *Resonant interactions between two trains of gravity waves*, J. Fluid Mech., 12 (1962), pp. 321–332.

[17] M. S. Longuet-Higgins and N. D. Smith, *An experiment on third-order wave interactions*, J. Fluid Mech., 25 (1966), pp. 417–435.

[18] B. M. Lake, H. C. Yuen, H. Rungaldier, and W. E. Ferguson, *Nonlinear deep-water waves: Theory and experiment. Part* 2: *Evolution of a continuous wave train*, J. Fluid Mech., 83 (1977), pp. 49–74.

[19] A. J. Majda, D. W. McLaughlin, and E. G. Tabak, *A one-dimensional model for dispersive wave turbulence*, J. Nonlinear Sci., 7 (1997), pp. 9–44.

[20] L. F. McGoldrick, O. M. Philips, N. E. Huang, and T. H. Hodgson, *Measurements of third-order resonant wave interactions*, J. Fluid Mech., 25 (1966), pp. 437–456.

[21] P. A. Milewski, *A formulation for water waves over topography*, Stud. Appl. Math., 100 (1998), pp. 95–106.

[22] P. A. Milewski and J. B. Keller, *Three-dimensional water waves*, Stud. Appl. Math., 97 (1996), pp. 149–166.

[23] P. A. Milewski and E. G. Tabak, *A pseudospectral procedure for the solution of nonlinear wave equations with examples from free-surface flows*, SIAM J. Sci. Comput., 21 (1999), pp. 1102–1114.

[24] O. M. Philips, *On the dynamics of unsteady gravity waves of finite amplitude. Part* 1. *The elementary interactions*, J. Fluid Mech., 9 (1960), pp. 193–217.

[25] K. Trulsen and K. B. Dysthe, *Frequency downshift in three-dimensional wave trains in a deep basin*, J. Fluid Mech., 352 (1997), pp. 359-373.

[26] G. B. Whitham, *Nonlinear dispersion of water waves*, J. Fluid Mech., 27 (1967), pp. 399–412.

[27] V. E. Zakharov, *Statistical theory of gravity and capillary waves on the surface of a finite-depth fluid*, Eur. J. Mech. B Fluids, 18 (1999), pp. 327–344.

[28] V. E. Zakharov, *Stability of periodic waves of finite amplitude on the surface of a deep fluid*, J. Appl. Mech. Tech. Phys., 2 (1968), pp. 190–194.

[29] V. E. Zakharov, *Kolmogorov spectra in weak turbulence problems*, in Handbook of Plasma Physics, Vol. 2, Elsevier, New York, 1984, pp. 1–36.

[30] V. E. Zakharov, P. Guyenne, A. N. Pushkarev, and F. Dias, *Wave turbulence in one-dimensional models*, Phys. D, 152–153 (2001), pp. 572–619.

SIAM J. APPL. MATH.
Vol. 63, No. 4, pp. 1141–1183

OPTIMIZATION OF ACOUSTIC SOURCE STRENGTH IN THE PROBLEMS OF ACTIVE NOISE CONTROL*

J. LONČARIĆ† AND S. V. TSYNKOV‡

Abstract. We consider a problem of eliminating the unwanted time-harmonic noise on a predetermined region of interest. The desired objective is achieved by active means, i.e., by introducing additional sources of sound called control sources, which generate the appropriate annihilating acoustic signal (antisound). A general solution for the control sources has been obtained previously in both continuous and discrete formulation of the problem. In the current paper, we focus on optimizing the overall absolute acoustic source strength of the control sources. Mathematically, this amounts to the minimization of multivariable complex-valued functions in the sense of L_1 with conical constraints, which are only "marginally" convex. The corresponding numerical optimization problem appears very challenging even for the most sophisticated state-of-the-art methodologies, and even when the dimension of the grid is small and the waves are long.

Our central result is that the global L_1-optimal solution can, in fact, be obtained without solving the numerical optimization problem. This solution is given by a special layer of monopole sources on the perimeter of the protected region. We provide a rigorous proof of global L_1 minimality for both continuous and discrete optimization problems in the one-dimensional case. We also provide numerical evidence that corroborates our result in the two-dimensional case, when the protected domain is a cylinder. Even though we cannot fully justify it, we believe that the same result holds in the general case, i.e., for multidimensional settings and domains of arbitrary shape. We formulate this notion as a conjecture at the end of the paper.

Key words. noise cancellation, control sources, minimization of amplitude, volume velocity, surface monopoles

AMS subject classifications. 35J05, 35C15, 35B37, 49J20, 65N06, 76Q05, 90C25, 90C30

PII. S0036139902404220

1. Introduction.

The area of active control of sound has a rich history of development, both as a chapter of theoretical acoustics and in the perspective of many different applications. Any attempt to adequately overview this extensive area in the framework of a focused research publication would obviously be deficient. Therefore, we simply refer the reader to monographs [3, 5, 11] that, among other things, contain a detailed survey of the literature.

The formulation of the problem that we use in the current paper has been introduced and studied in our previous work [7]; here, we analyze this formulation from the standpoint of optimization. Let Ω be a given domain, $\Omega \subset \mathbb{R}^n$, where the space dimension $n = 2$ or $n = 3$. (These two cases are most interesting for applications.) The domain Ω can be either bounded or unbounded; for reasons of simplicity we will further assume that Ω is bounded. Let Γ be the boundary of Ω: $\Gamma = \partial\Omega$. Both on Ω and on its (unbounded) complement $\Omega_1 = \mathbb{R}^n \backslash \Omega$ we consider the time-harmonic acoustic

*Received by the editors March 18, 2002; accepted for publication (in revised form) September 16, 2002; published electronically March 26, 2003. This research was supported by the National Aeronautics and Space Administration under NASA contract NAS1–97046, and in the framework of the Creativity and Innovation Program (C&I) while the authors were in residence at ICASE, NASA Langley Research Center.

http://www.siam.org/journals/siap/63-4/40422.html

†ICASE, MS 132C, NASA Langley Research Center, Hampton, VA 23681–2199. Current address: National Institute of Aerospace, 144 Research Drive, Hampton, VA 23666 (josip@nianet.org).

‡Department of Mathematics, North Carolina State University, Box 8205, Raleigh, NC 27695 and School of Mathematical Sciences, Tel Aviv University, Ramat Aviv, Tel Aviv 69978, Israel (tsynkov@math.ncsu.edu, http://www.math.ncsu.edu/∼stsynkov/).

field $u = u(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^n$, governed by the nonhomogeneous Helmholtz equation:

$$\boldsymbol{L}u \equiv \Delta u + k^2 u = f. \tag{1.1}$$

Equation (1.1) is subject to the Sommerfeld radiation boundary conditions at infinity, which for $n = 2$ are formulated as

$$u(\boldsymbol{x}) = O(|\boldsymbol{x}|^{-1/2}), \quad \frac{\partial u(\boldsymbol{x})}{\partial |\boldsymbol{x}|} + iku(\boldsymbol{x}) = o(|\boldsymbol{x}|^{-1/2}) \quad \text{as} \quad |x| \longrightarrow \infty, \tag{1.2a}$$

and for $n = 3$ as

$$u(\boldsymbol{x}) = O\left(|\boldsymbol{x}|^{-1}\right), \quad \frac{\partial u(\boldsymbol{x})}{\partial |\boldsymbol{x}|} + iku(\boldsymbol{x}) = o\left(|\boldsymbol{x}|^{-1}\right) \quad \text{as} \quad |\boldsymbol{x}| \longrightarrow \infty. \tag{1.2b}$$

The Sommerfeld boundary conditions specify the direction of wave propagation and distinguish between the incoming and outgoing waves at infinity by prescribing the outgoing direction only; they guarantee the unique solvability of the Helmholtz equation (1.1) for any compactly supported right-hand side $f = f(\boldsymbol{x})$. We define $\operatorname{supp} f = \{\boldsymbol{x} | f(\boldsymbol{x}) \neq 0\}$.

The source terms $f = f(\boldsymbol{x})$ in (1.1) can be located on both Ω and its complement $\Omega_1 = \mathbb{R}^n \backslash \Omega$; to emphasize the distinction, we define

$$f = f^+ + f^-, \tag{1.3}$$

where the sources f^+ are interior, $\operatorname{supp} f^+ \subset \Omega$, and the sources f^- are exterior, $\operatorname{supp} f^- \subset \Omega_1$, with respect to Ω. Accordingly, the overall acoustic field $u = u(\boldsymbol{x})$ can be represented as the sum of two components:

$$u = u^+ + u^-, \tag{1.4}$$

where

$$\boldsymbol{L}u^+ = f^+, \tag{1.5a}$$

$$\boldsymbol{L}u^- = f^-. \tag{1.5b}$$

Note that both $u^+ = u^+(\boldsymbol{x})$ and $u^- = u^-(\boldsymbol{x})$ are defined on the entire $\mathbb{R}^n$; the superscripts "+" and "−" refer to the sources that drive each of the field components, rather than to the domains of these components. The setup described above is schematically shown in Figure 1.1.

Hereafter, we will call the component u^+ of (1.4), (1.5a) *sound*, or the "friendly" part of the total acoustic field; the component u^- of (1.4), (1.5b) will accordingly be called *noise*, or the "adverse" part of the total acoustic field. In the formulation that we are presenting, Ω will be a (predetermined) region of space to be shielded. This means that we would like to eliminate the noise inside Ω while leaving the sound component there unaltered. In the mathematical framework that we have adopted, the component u^- of the total acoustic field, i.e., the response to the adverse sources f^- (see (1.3), (1.4), (1.5)), will have to be cancelled on Ω, whereas the component u^+, i.e., the response to the friendly sources f^+, will have to be left unaffected on Ω. A physically more involved but conceptually easy-to-understand example can be given to illustrate the foregoing idea of shielding: inside the passenger compartment of an aircraft we would like to eliminate the noise coming from the propulsion system

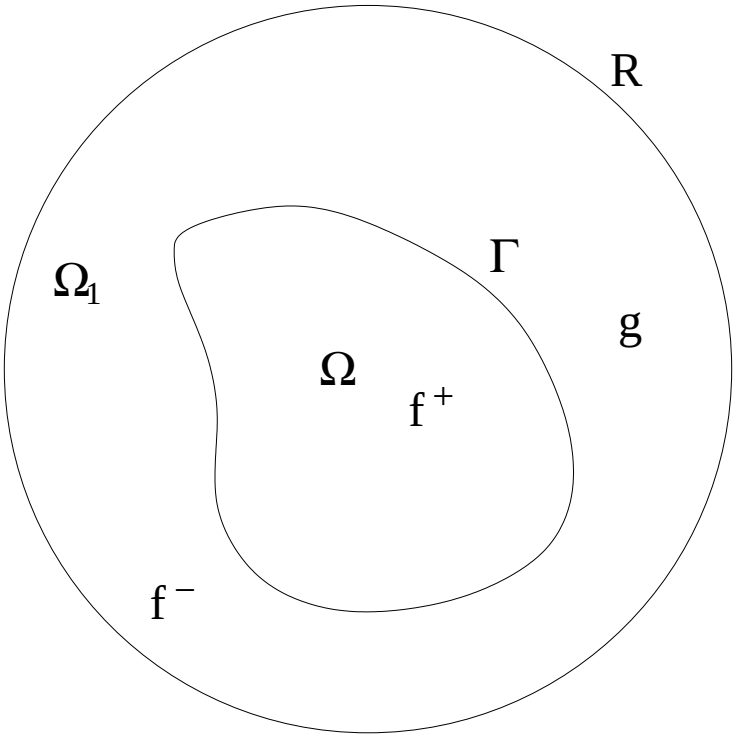


FIG. 1.1. *Geometric setup.*

located outside the fuselage, while not interfering with the ability of the passengers to listen to the in-flight entertainment programs or to converse.

The concept of *active noise control* that we will be discussing implies that the component u^- is to be suppressed on Ω by introducing additional sources of sound $g = g(\boldsymbol{x})$ exterior with respect to Ω, $\operatorname{supp} g \subset \Omega_1$, so that the total acoustic field $\tilde{u} = \tilde{u}(\boldsymbol{x})$ can now be governed by the equation (cf. formulae (1.1), (1.3))

$$\boldsymbol{L}\tilde{u} = f^+ + f^- + g \tag{1.6}$$

and coincide with only the friendly component u^+ on the domain Ω:

$$\tilde{u}|_{\boldsymbol{x}\in\Omega} = u^+|_{\boldsymbol{x}\in\Omega}. \tag{1.7}$$

The new sources $g = g(\boldsymbol{x})$ of (1.6)—see Figure 1.1—will hereafter be referred to as the *control sources* or simply *controls.* An obvious solution for these control sources is $g = -f^-$. This solution, however, is excessively expensive. One one hand, the excessiveness comes from the information-type considerations, as the solution $g = -f^-$ requires explicit and detailed knowledge of the structure and location of the sources f^-. As shown in [7], this knowledge is, in fact, superfluous. On the other hand, the implementation of the solution $g = -f^-$ may encounter most serious technical difficulties. In the example above, it is obviously not feasible to directly counter the actual noise sources, which are aircraft propellers or turbofan jet engines located on, or underneath, the wings. Therefore, solutions of this noise control problem

other than the most obvious one may be preferable from both the theoretical and practical standpoints. The general solution for the control sources g was obtained in our previous work [7], and we describe it in section 2.

Before proceeding, let us note only that in the current paper we focus on the case of the standard constant-coefficient Helmholtz equation (1.1), which governs the acoustic field and is valid throughout the entire space $\mathbb{R}^n$. This allows us to make subsequent analysis most straightforward. As a matter of fact, other, more complex, cases that involve variable coefficients and possibly nonlinearities in the governing equations over some regions, as well as different types of far-field behavior, discontinuities in the material properties, etc., can be considered as well. Approaches to obtaining solutions for active controls in these cases are outlined in our previous paper [7] for the continuous formulation of the problem, and in the monograph by Ryaben'kii [15, Part VIII] for the discrete formulation of the problem.

The material in the rest of the paper is organized as follows. In section 2, we introduce the control sources for the continuous formulation of the problem. In section 3, we obtain the control sources in the discrete formulation of the problem, i.e., on the grid. In section 4, we discuss minimization of the overall acoustic source strength of active controls that we have constructed, which mathematically amounts to the optimization in the sense of L_1. We present convincing two-dimensional numerical evidence, as well as a rigorous one-dimensional proof, of the global L_1-optimality of a particular layer of monopole sources concentrated on the perimeter of the protected region. We believe that the combination of computations in two space dimensions and general proof in one space dimension cannot be a mere coincidence. As such, even though we cannot fully justify it, we formulate the corresponding general result on global L_1-optimality of surface monopoles as a conjecture in the concluding section 5. This conjecture basically implies that the aforementioned L_1-optimization problem can be solved without using any numerical optimization techniques.

2. Continuous control sources.

2.1. General solution. As demonstrated in [7], the general solution for the control sources $g = g(\boldsymbol{x})$ is given by the following formula ($\Omega_1 = \mathbb{R}^n \backslash \Omega$):

$$g(\boldsymbol{x}) = -\boldsymbol{L}w\Big|_{\boldsymbol{x}\in\Omega_1}, \tag{2.1}$$

where $w = w(\boldsymbol{x})$, $\boldsymbol{x} \in \Omega_1$, is a special auxiliary function-parameter that parameterizes the family of controls (2.1). The requirements that the function $w(\boldsymbol{x})$ must meet are, in fact, relatively "loose." At infinity, it has to satisfy the Sommerfeld boundary conditions (1.2a) or (1.2b). At the interface Γ, the function w and its normal derivative have to coincide with the corresponding quantities that pertain to the total acoustic field u given by formula (1.4):[1]

$$w\Big|_\Gamma = u\Big|_\Gamma, \qquad \frac{\partial w}{\partial \boldsymbol{n}}\Big|_\Gamma = \frac{\partial u}{\partial \boldsymbol{n}}\Big|_\Gamma. \tag{2.2}$$

Other than that, the function $w(\boldsymbol{x})$ used in (2.1) is arbitrary, and consequently formula (2.1) defines a large family of control sources, which, as will be seen in section 4, provides ample room for optimization. To make the discussion in the current paper

[1] In practice, the quantities u and $\frac{\partial u}{\partial \boldsymbol{n}}$ on Γ can be measured and supplied to the control system as the input data.

self-contained, we briefly outline below the justification for formula (2.1) as a general solution for controls, while referring the reader to [7] for further detail.

Introducing the fundamental solution $G = G(\boldsymbol{x})$ of the Helmholtz operator $\boldsymbol{L}$ of (1.1) for $n = 2$,

$$G(\boldsymbol{x}) = -\frac{1}{4i} H_0^{(2)}(k|\boldsymbol{x}|), \tag{2.3a}$$

where $H_0^{(2)}(z) = J_0(z) - iY_0(z)$ is the Hankel function of the second kind, and for $n = 3$,

$$G(\boldsymbol{x}) = -\frac{e^{-ik|\boldsymbol{x}|}}{4\pi|\boldsymbol{x}|}, \tag{2.3b}$$

we can obviously represent the sound portion $u^+ = u^+(\boldsymbol{x})$ of the overall acoustic field that satisfies (1.5a) everywhere on $\mathbb{R}^n$ as follows:

$$u^+(\boldsymbol{x}) = \int_\Omega G(\boldsymbol{x} - \boldsymbol{y}) f^+(\boldsymbol{y}) d\boldsymbol{y} = \int_\Omega G(\boldsymbol{x} - \boldsymbol{y}) \boldsymbol{L} u^+(\boldsymbol{y}) d\boldsymbol{y}, \quad \boldsymbol{x} \in \mathbb{R}^n. \tag{2.4}$$

Consequently, applying the classical Green's formula (see, e.g., [18] or [24]) to the function $u^+ = u^+(\boldsymbol{x})$ on Ω, we have

$$\int_\Gamma \left(u^+ \frac{\partial G}{\partial \boldsymbol{n}} - \frac{\partial u^+}{\partial \boldsymbol{n}} G \right) ds_{\boldsymbol{y}} = 0, \quad \boldsymbol{x} \in \Omega, \tag{2.5}$$

where integrals in (2.5) are, again, convolutions. Similarly, applying the same Green's formula on Ω to $u^- = u^-(\boldsymbol{x})$ and using (2.5), we obtain

$$u^-(\boldsymbol{x}) = \int_\Gamma \left(u^- \frac{\partial G}{\partial \boldsymbol{n}} - \frac{\partial u^-}{\partial \boldsymbol{n}} G \right) ds_{\boldsymbol{y}} = \int_\Gamma \left(u \frac{\partial G}{\partial \boldsymbol{n}} - \frac{\partial u}{\partial \boldsymbol{n}} G \right) ds_{\boldsymbol{y}}, \quad \boldsymbol{x} \in \Omega. \tag{2.6}$$

Therefore, from (2.6) we can conclude that the desired annihilating acoustic signal $v = v(\boldsymbol{x})$ that cancels out the unwanted noise on Ω can be obtained as

$$v(\boldsymbol{x}) = -\int_\Gamma \left(u \frac{\partial G}{\partial \boldsymbol{n}} - \frac{\partial u}{\partial \boldsymbol{n}} G \right) ds_{\boldsymbol{y}}, \tag{2.7}$$

so that for $\boldsymbol{x} \in \Omega$ we indeed have

$$[u(\boldsymbol{x}) + v(\boldsymbol{x})]_{\boldsymbol{x} \in \Omega} = u^+(\boldsymbol{x}).$$

To actually implement the annihilating signal v of (2.7), we introduce the auxiliary function $w = w(\boldsymbol{x})$ on $\mathbb{R}^n$ that satisfies the aforementioned conditions at the interface Γ and at infinity, and apply the Green's formula to w, which yields

$$w(\boldsymbol{x}) - \int_\Omega G \boldsymbol{L} w d\boldsymbol{y} = \int_\Gamma \left(w \frac{\partial G}{\partial \boldsymbol{n}} - \frac{\partial w}{\partial \boldsymbol{n}} G \right) ds_{\boldsymbol{y}}, \quad \boldsymbol{x} \in \Omega. \tag{2.8}$$

As $w(\boldsymbol{x})$ satisfies the Sommerfeld conditions (1.2a) or (1.2b), we obviously have $w(\boldsymbol{x}) = \int_{\mathbb{R}^n} G \boldsymbol{L} w d\boldsymbol{y}$, $\boldsymbol{x} \in \mathbb{R}^n$, which, along with formulae (2.2) and (2.7), allows us to transform equality (2.8) to

$$v(\boldsymbol{x}) = -\int_{\Omega_1} G \boldsymbol{L} w d\boldsymbol{y}. \tag{2.9}$$

Equation (2.9) implies that for any $w = w(\boldsymbol{x})$ chosen as described above, formula (2.1) describes an appropriate control function.

Conversely, assume that $g = g(x)$, $\operatorname{supp} g \subset \Omega_1$, is a control field such that the solution $\tilde{u} = \tilde{u}(\boldsymbol{x})$ of (1.6) subject to the Sommerfeld conditions (1.2a) or (1.2b) satisfies equality (1.7). Then, by choosing $w = w(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^n$, as the solution to the nonhomogeneous equation

$$-\boldsymbol{L}w = g - f^+,$$

subject to the corresponding Sommerfeld condition, (1.2a) or (1.2b), one can represent g in the form (2.1); see [7]. Altogether, we obtain that formula (2.1) describes the general solution for controls. In other words, for any $w(\boldsymbol{x})$, formula (2.1) provides an appropriate control field $g(\boldsymbol{x})$, and any appropriate control field $g(\boldsymbol{x})$ can be represented in the form (2.1) with some particular choice of $w(\boldsymbol{x})$.

Let us emphasize several important properties of controls (2.1). First of all, from the foregoing derivation we can see that to obtain these controls one needs no knowledge of the actual exterior sources of noise f^-. In other words, neither their location, nor structure, nor strength are required. All one needs to know is u and $\frac{\partial u}{\partial \boldsymbol{n}}$ on the perimeter Γ of the protected region Ω. As has been mentioned, in a practical setting $u|_\Gamma$ and $\frac{\partial u}{\partial \boldsymbol{n}}|_\Gamma$ can be interpreted as measurable quantities that are supplied to the control system as the input data. Moreover, these measurable quantities refer to the overall acoustic field u, rather than only to its unwanted component u^-. In other words, the methodology can automatically distinguish between the signals coming from the exterior and interior sources, and can tune the controls so that they cancel only the unwanted exterior signal. This capability is extremely important, as in many applications the overall acoustic field always contains a component that needs to be suppressed along with the part that needs to be left intact. Alternatively, one can say that the control sources (2.1) are insensitive to the interior sound $u^+(\boldsymbol{x})$. Indeed, given a function $w(\boldsymbol{x})$ that satisfies interface conditions (2.2) and the radiation boundary conditions at infinity, we can take instead $\tilde{w}(\boldsymbol{x}) = w(\boldsymbol{x}) - u^+(\boldsymbol{x})$; this new function will satisfy the interface conditions (2.2) with u replaced by u^- and the same Sommerfeld conditions at infinity. Most important, the control sources generated by $\tilde{w}(\boldsymbol{x})$ will be the exact same control sources as those generated by $w(\boldsymbol{x})$ because $\boldsymbol{L}\tilde{w}(\boldsymbol{x}) = \boldsymbol{L}[w(\boldsymbol{x}) - u^+(\boldsymbol{x})] = \boldsymbol{L}w(\boldsymbol{x})$ for $\boldsymbol{x} \in \Omega_1$.

Let us also note that in a more general framework, formula (2.9) (with (2.2) taken into account) can be interpreted as a particular case of the generalized potential of Calderon's type with the vector density $-\big(u, \frac{\partial u}{\partial \boldsymbol{n}}\big)|_\Gamma$. This more general framework allows us to analyze more complex formulations of the active noise control problem (see [7]) such as those that involve variations in material properties and alternative types of far-field behavior of the solution. We refer the reader to the original work by Calderon [2] and Seeley [16], as well as to the monograph by Ryaben'kii [15], for general concepts related to Calderon's potentials and associated pseudodifferential boundary projection operators. As concerns the aforementioned more advanced formulations of the noise control problem, the key result is basically the same as above. The general solution for control sources is still given by formula (2.1), where the auxiliary function $w(\boldsymbol{x})$ should still satisfy the interface conditions (2.2) and the problem-specific far-field boundary conditions (in case they differ from the previously mentioned Sommerfeld conditions). The operator $\boldsymbol{L}$ in formula (2.1) will, however, no longer be the constant-coefficient Helmholtz operator of (1.1); it will rather be the problem-specific variable-coefficients operator that accounts for the particular variations in material properties, etc. However, since formula (2.1) for controls does not

change (see [7]), we immediately conclude that we, in fact, do not need to know the operator $\boldsymbol{L}$ on Ω. In other words, for obtaining the control sources we do not need to know the material properties of the sound-conducting medium inside the protected region. This result (see [7]), which at first seems counterintuitive, has, in fact, a straightforward physical explanation. We only need to realize that the noise we want to suppress, and the output of controls that is supposed to annihilate this noise, propagate across one and the same medium, and we do not need to know what this medium is (under some relatively nonrestrictive limitations; see [7]). It is also interesting to mention that the aforementioned Calderon boundary projection operators essentially render the decomposition of the wave field $u(\boldsymbol{x})$ on the boundary Γ into its incoming and outgoing component with respect to the domain Ω; see [7]. Subsequently, the controls (2.1) can be interpreted as either sources cancelling the incoming wave field for the domain to be shielded, i.e., Ω, or alternatively, as sources cancelling the outgoing wave field for the domain complementary to the one to be shielded, i.e., $\Omega_1 = \mathbb{R}^n \backslash \Omega$. The latter interpretation is often more versatile; see [7].

Another important thing to notice is that the control sources $g(\boldsymbol{x})$ of (2.1) are defined, generally speaking, on the entire complementary domain $\Omega_1 = \mathbb{R}^n \backslash \Omega$. For the analysis of specific problems, especially when the protected region Ω is bounded and the complementary region Ω_1 is unbounded, as in Figure 1.1, it may be convenient to consider compactly supported control sources, i.e., the control sources concentrated in the vicinity of the interface Γ. To obtain such controls, one needs to narrow down the class of functions $w(\boldsymbol{x})$ used in formula (2.1). Namely, instead of considering arbitrary $w(\boldsymbol{x})$ subject only to constraints (2.2) and Sommerfeld boundary conditions at infinity, one needs to consider $w(\boldsymbol{x})$ that become a solution to the homogeneous Helmholtz equation everywhere outside some larger domain that fully contains the protected region Ω. In so doing, the area outside Ω that supports the controls $g(\boldsymbol{x})$ of (2.1) will stretch from Γ to the outer boundary of the aforementioned larger domain, and may basically look like a curvilinear strip adjacent to Γ from the exterior side. This strip may, in principle, be made as narrow as desired and may eventually shrink completely, thus reducing to only the interface Γ itself. As will be seen, the sources $g(\boldsymbol{x})$ supported only on the perimeter Γ represent an important class of active controls. Such distributions $g(\boldsymbol{x})$ include monopole and dipole layers as special cases, which are idealizations of pulsating or vibrating membranes.

2.2. Artificial boundary conditions and compactly supported controls. To actually obtain compactly supported controls $g(\boldsymbol{x})$ in a particular setting, it is often convenient to use the methodology known as artificial boundary conditions (see the review paper [19]) for the selection of the appropriate function $w(\boldsymbol{x})$ in formula (2.1). Assume that $w(\boldsymbol{x})$ satisfies the homogeneous Helmholtz equation $\boldsymbol{L}w = 0$ outside some outer artificial boundary, which is a closed surface (curve) with an interior that fully contains Ω. In Figure 1.1, we schematically represent this outer boundary as a sphere (circle) of radius R. It turns out that one can equivalently replace the homogeneous equation $\boldsymbol{L}w = 0$ along with the Sommerfeld boundary conditions at infinity by the special artificial boundary conditions (ABCs) at the outer boundary. For the outer boundary of a general shape, this can be done most efficiently using the same apparatus of Calderon's pseudodifferential boundary projection operators (see [15, 19]) that has been mentioned before. For the particular case of a regular spherical or circular outer boundary (see Figure 1.1), which is convenient due to the simplicity of the analysis, the construction of the ABCs is described below. We emphasize that the forthcoming boundary conditions are exact. In other words, they

are set at a finite artificial boundary and are fully equivalent to the Sommerfeld boundary conditions set at infinity. They also appear to be nonlocal. High accuracy (exactness) and nonlocality of the ABCs that we introduce and use hereafter present a notable distinction compared to many approximate methods, which are typically local but not as accurate; see, e.g., [19].

We will use spherical coordinates (ρ, θ, ϕ) in $\mathbb{R}^n$, $n = 3$, and assume that $\boldsymbol{L}w = 0$ for $|\boldsymbol{x}| \equiv \rho \geq R$; see Figure 1.1. In addition, we will assume that $w(\boldsymbol{x})$ satisfies the Sommerfeld boundary condition (1.2b). Expanding $w(\boldsymbol{x})$ with respect to spherical functions Y_l^m, $l = 0, 1, 2, \ldots$, $m = 0, \pm 1, \ldots, \pm l$, and separating the variables in the differential operator $\boldsymbol{L}$, we arrive at the following collection of second-order ordinary differential equations:

$$\begin{aligned} &\frac{d^2\hat{w}_{lm}}{d\rho^2} + \frac{2}{\rho}\frac{d\hat{w}_{lm}}{d\rho} + \left[k^2 - \frac{l(l+1)}{\rho^2}\right]\hat{w}_{lm} = 0, \\ &\rho \geq R, \quad l = 0, 1, 2, \ldots, \quad m = 0, \pm 1, \ldots, \pm l, \end{aligned} \tag{2.10}$$

for the unknown radial modes $\hat{w}_{lm}$. These modes are also supposed to satisfy boundary conditions at infinity

$$\hat{w}_{lm}(\rho) = O\left(\rho^{-1}\right), \quad \frac{d\hat{w}_{lm}(\rho)}{d\rho} + ik\hat{w}_{lm}(\rho) = o\left(\rho^{-1}\right) \quad \text{as} \quad \rho \longrightarrow \infty, \tag{2.11}$$

which immediately follow from the Sommerfeld condition (1.2b). For any given pair (l, m) the general solution of (2.10) is given by

$$\hat{w}_{lm} = \frac{c_1}{\sqrt{\rho}} H^{(1)}_{l+1/2}(k\rho) + \frac{c_2}{\sqrt{\rho}} H^{(2)}_{l+1/2}(k\rho), \tag{2.12}$$

where c_1 and c_2 are arbitrary constants and $H^{(1)}_{l+1/2}(k\rho)$ and $H^{(2)}_{l+1/2}(k\rho)$ are Hankel functions of the first and second kind, respectively. Taking into account the asymptotic expressions for the Hankel functions for a fixed order ν and large values of ρ (see, e.g., [24]),

$$\begin{aligned} H^{(1)}_\nu(\rho) &= \sqrt{\frac{2}{\pi\rho}} \exp\left[i\left(\rho - \frac{\pi}{2}\nu - \frac{\pi}{4}\right)\right] + O\left(\rho^{-3/2}\right), \\ H^{(2)}_\nu(\rho) &= \sqrt{\frac{2}{\pi\rho}} \exp\left[-i\left(\rho - \frac{\pi}{2}\nu - \frac{\pi}{4}\right)\right] + O\left(\rho^{-3/2}\right), \end{aligned}$$

we conclude that only $\rho^{-1/2}H^{(2)}_{l+1/2}(k\rho)$ satisfies boundary conditions (2.11), and consequently the constant c_1 (see (2.12)) in any particular solution that satisfies the Sommerfeld conditions at infinity has to be equal to zero. As the two functions $\rho^{-1/2}H^{(1)}_{l+1/2}(k\rho)$ and $\rho^{-1/2}H^{(2)}_{l+1/2}(k\rho)$ form a fundamental system of solutions for the linear homogeneous second-order ODE (2.10), requiring that only one of them, $\rho^{-1/2}H^{(2)}_{l+1/2}(k\rho)$, be present in the actual solution $\hat{w}_{lm}$ is equivalent to requiring that $\hat{w}_{lm}(\rho)$ be parallel to $\rho^{-1/2}H^{(2)}_{l+1/2}(k\rho)$ for $\rho \geq R$ in the sense of the corresponding Wronskian vanishing at $\rho = R$:

$$\det \begin{bmatrix} \hat{w}_{lm} & \rho^{-1/2}H^{(2)}_{l+1/2}(k\rho) \\ \frac{d}{d\rho}\hat{w}_{lm} & \frac{d}{d\rho}\left(\rho^{-1/2}H^{(2)}_{l+1/2}(k\rho)\right) \end{bmatrix}\Bigg|_{\rho=R} = 0. \tag{2.13}$$

Obviously, equality (2.13) enforced at $\rho = R$ implies that it will hold for all $\rho \geq R$ as well. Equality (2.13) is a linear homogeneous relation between a given Fourier component $\hat{w}_{lm}$ of the solution $w(\boldsymbol{x})$ and the corresponding Fourier component $\frac{d}{d\rho}\hat{w}_{lm}$ of its normal derivative $\frac{d}{d\rho}w(\boldsymbol{x})$ on the spherical surface $\rho = R$. The entire family of such relations for all $l = 0, 1, 2, \ldots$ and $m = 0, \pm 1, \ldots, \pm l$ set at $\rho = R$ is equivalent to saying that the function $w = w(\boldsymbol{x})$ originally defined inside the sphere, i.e., for $\rho \leq R$, can be smoothly extended to the region $\rho \geq R$ so that the extension will solve the homogeneous equation $\boldsymbol{L}w = 0$ for $\rho \geq R$ and have a proper far-field behavior, i.e, satisfy the Sommerfeld condition (1.2b). Hereafter, we will refer to relations (2.13) for $l = 0, 1, 2, \ldots$ and $m = 0, \pm 1, \ldots, \pm l$ as *the artificial boundary conditions* for the three-dimensional Helmholtz equation on the spherical surface $\rho = R$.

The ABCs for the two-dimensional case on the circular external artificial boundary $\rho = R$ can be obtained similarly. We introduce polar coordinates (ρ, θ) in $\mathbb{R}^n$, $n = 2$, use standard Fourier expansion in the circumferential direction with respect to the complex exponents $e^{-il\theta}$, $l = 0, \pm 1, \pm 2, \ldots$, and arrive at the following collection of second-order ordinary differential equations:

$$\frac{d^2\hat{w}_l}{d\rho^2} + \frac{1}{\rho}\frac{d\hat{w}_l}{d\rho} + \left[k^2 - \frac{l^2}{\rho^2}\right]\hat{w}_l = 0, \quad \rho \geq R, \quad l = 0, \pm 1, \pm 2, \ldots, \tag{2.14}$$

for the unknown radial modes $\hat{w}_l$. These modes are also supposed to satisfy boundary conditions at infinity

$$\hat{w}_l(\rho) = O(\rho^{-1/2}), \quad \frac{d\hat{w}_l(\rho)}{d\rho} + ik\hat{w}_l(\rho) = o(\rho^{-1/2}) \quad \text{as} \quad \rho \longrightarrow \infty, \tag{2.15}$$

which immediately follow from the Sommerfeld condition (1.2a). For every given l, (2.14) is the Bessel equation and has general solution

$$\hat{w}_l = c_1 H_l^{(1)}(k\rho) + c_2 H_l^{(2)}(k\rho), \tag{2.16}$$

where c_1 and c_2 are, again, arbitrary constants. The asymptotics of the Hankel functions for large ρ's indicates that to satisfy (2.15) one must have $c_1 = 0$ in any particular solution that satisfies the radiation conditions at infinity. This requirement leads to the following ABCs for the two-dimensional Helmholtz equation:

$$\det \begin{bmatrix} \hat{w}_l & H_l^{(2)}(k\rho) \\ \frac{d}{d\rho}\hat{w}_l & \frac{d}{d\rho}H_l^{(2)}(k\rho) \end{bmatrix}\Bigg|_{\rho=R} = 0, \tag{2.17}$$

where relations (2.17) should be considered for all $l = 0, \pm 1, \pm 2, \ldots$. We refer the reader to the review article [19] for further detail on the construction of ABCs for different equations in different settings. Let us also reiterate that once the ABCs (2.13) or (2.17) are satisfied for all radial modes, then we can consider $\boldsymbol{L}w = 0$ for $\rho \geq R$, and as such, the resulting control sources $g(\boldsymbol{x})$ given by (2.1) will be compactly supported between the interface Γ and the external artificial boundary $\rho = R$.

2.3. Types of control sources. Let us now analyze the continuous control sources from the standpoint of their geometric location and the type of acoustic excitation that they provide. To put this analysis into a mathematical perspective, it will be convenient to use the apparatus of distributions; see, e.g., [24].

In our original derivation of formula (2.1), we have implicitly assumed that the function $w(\boldsymbol{x})$ was sufficiently smooth so that the operator $\boldsymbol{L}$ could be applied in the classical sense everywhere on $\Omega_1 = \mathbb{R}^n \backslash \Omega$. In this case, the function $g(\boldsymbol{x})$ is locally absolutely integrable, $g \in \boldsymbol{L}_1^{(\mathrm{loc})}(\Omega_1)$, and can be interpreted as a regular distribution. As for any other distribution, it can be represented as a convolution of its own self with the δ-function: $g = \delta * g$. This means that from the mathematical standpoint the control field $g(\boldsymbol{x})$ can be viewed as an expansion in terms of the elementary point monopoles $g(\boldsymbol{y})\delta(\boldsymbol{x}-\boldsymbol{y})$ with regular density g:

$$g(\boldsymbol{x}) = \delta * g = \int_{\mathbb{R}^n} g(\boldsymbol{y})\delta(\boldsymbol{x}-\boldsymbol{y})d\boldsymbol{y}. \tag{2.18}$$

Next, we recall that, by definition of the fundamental solution $\boldsymbol{L}G = \delta(\boldsymbol{x})$, the response to every such elementary monopole $g(\boldsymbol{y})\delta(\boldsymbol{x}-\boldsymbol{y})$ will be given by $g(\boldsymbol{y})G(\boldsymbol{x}-\boldsymbol{y})$, and consequently, the overall control output

$$G * g = \int_{\mathbb{R}^n} g(\boldsymbol{y})G(\boldsymbol{x}-\boldsymbol{y})d\boldsymbol{y} \tag{2.19}$$

will be interpreted as superposition of the foregoing elementary responses, i.e., solutions generated by the aforementioned point monopoles. Altogether we see that the original formula (2.1) provides the general solution for controls in the class of regular distributions $\boldsymbol{L}_1^{(\mathrm{loc})}(\Omega_1)$, which corresponds to the volumetric control sources of monopole type on the complementary domain $\Omega_1 = \mathbb{R}^n \backslash \Omega$. Compactly supported controls discussed in section 2.2 obviously fall into this category. Later on we will see that this class of functions contains, in fact, all meaningful volumetric excitations.

In many cases it may also be desirable to consider surface controls, i.e., the control sources that are concentrated only on the interface Γ. Let us first assume that there are no interior sources, which means that $u^+(\boldsymbol{x}) = 0$, and the acoustic field we want to control consists only of its adverse component, $u(\boldsymbol{x}) \equiv u^-(\boldsymbol{x})$. After the control, the overall acoustic field has to be equal to zero on the domain Ω. As shown in [20], the general solution for surface controls is given by

$$g^{(\mathrm{surf})} = -\left[\frac{\partial w}{\partial \boldsymbol{n}} - \frac{\partial u}{\partial \boldsymbol{n}}\right]_\Gamma \delta(\Gamma) - \frac{\partial}{\partial \boldsymbol{n}}\big([w-u]_\Gamma \delta(\Gamma)\big), \tag{2.20}$$

where $w = w(\boldsymbol{x})$, as before, denotes the auxiliary function-parameter that in this case has to satisfy the homogeneous Helmholtz equation on the complementary domain, $\boldsymbol{L}w = 0$ for $\boldsymbol{x} \in \Omega_1$, and the Sommerfeld boundary condition (1.2a) or (1.2b) at infinity. Expressions in rectangular brackets in formula (2.20) denote discontinuities of the corresponding quantities across the interface Γ. The first term on the right-hand side of (2.20) represents the density of a single-layer potential, which is a layer of monopoles on the interface Γ, and the second term on the right-hand side of (2.20) represents the density of a double-layer potential, which is a layer of dipoles on the interface Γ.

A detailed justification of formula (2.20) as general solution for surface controls can be found in [20]. Here we mention only that it basically amounts to proving that a given solution $u(\boldsymbol{x})$ of the homogeneous equation $\boldsymbol{L}u = 0$ on Ω can be represented as a combination of a single-layer potential and a double-layer potential if and only if the densities of the aforementioned potentials are defined as $\left[\frac{\partial w}{\partial \boldsymbol{n}} - \frac{\partial u}{\partial \boldsymbol{n}}\right]_\Gamma \delta(\Gamma)$ and

$\frac{\partial}{\partial \boldsymbol{n}}\left([w-u]_\Gamma \delta(\Gamma)\right)$, respectively (cf. formula (2.20)). The direct implication is easy to establish by applying the operator $\boldsymbol{L}$ to the discontinuous function

$$v(\boldsymbol{x}) = \begin{cases} u(\boldsymbol{x}) & \text{for } \boldsymbol{x} \in \Omega, \\ w(\boldsymbol{x}) & \text{for } \boldsymbol{x} \in \Omega_1, \end{cases} \tag{2.21}$$

in the sense of distributions (see [24]), which yields

$$\boldsymbol{L}v = \left[\frac{\partial w}{\partial \boldsymbol{n}} - \frac{\partial u}{\partial \boldsymbol{n}}\right]_\Gamma \delta(\Gamma) + \frac{\partial}{\partial \boldsymbol{n}}\big([w-u]_\Gamma \delta(\Gamma)\big), \tag{2.22}$$

and subsequently reconstructing $u(\boldsymbol{x})|_{\boldsymbol{x}\in\Omega} = v(\boldsymbol{x})|_{\boldsymbol{x}\in\Omega}$ as a convolution with the fundamental solution. The inverse implication requires explicitly obtaining $w(\boldsymbol{x})$ for a given $u(\boldsymbol{x})$ and given surface densities, which is done in [20], again, in the form of a special surface integral. Then, the control sources (2.20) are obtained by simply taking (2.22) with the opposite sign, which guarantees the cancellation of $u(\boldsymbol{x})$ on Ω by $-v(\boldsymbol{x})$.

In the family of surface controls (2.20) we identify two important particular cases. First, the cancellation of $u(\boldsymbol{x})$, $\boldsymbol{x} \in \Omega$, can be achieved by using only the surface monopoles, i.e., by employing only a single-layer potential as the annihilating signal. To do that, we need to find $w(\boldsymbol{x})$, $\boldsymbol{x} \in \Omega_1$, such that the overall function $v(\boldsymbol{x})$ of (2.21) would have the discontinuity on Γ only in its normal derivative and not in the function itself. This $w(\boldsymbol{x})$ will then be a solution of the following external Dirichlet problem:

$$\begin{aligned} \boldsymbol{L}w = 0, \quad & \boldsymbol{x} \in \Omega_1, \\ w\big|_\Gamma = u\big|_\Gamma, \quad & \end{aligned} \tag{2.23}$$

subject to the appropriate Sommerfeld boundary condition (1.2a) or (1.2b). Problem (2.23) is always uniquely solvable on $\Omega_1 = \mathbb{R}^n \backslash \Omega$. Second, one can employ only the double-layer potential to cancel out $u(\boldsymbol{x})$, $\boldsymbol{x} \in \Omega$, i.e., use only surface dipoles as the control sources. In this case, the function $w(\boldsymbol{x})$, $\boldsymbol{x} \in \Omega_1$, has to be such that $v(\boldsymbol{x})$ given by (2.21) would have discontinuity on Γ only in the function itself and not in its normal derivative. This $w(\boldsymbol{x})$ should then solve the following external Neumann problem:

$$\begin{aligned} \boldsymbol{L}w = 0, \quad & \boldsymbol{x} \in \Omega_1, \\ \frac{\partial w}{\partial \boldsymbol{n}}\Big|_\Gamma = \frac{\partial u}{\partial \boldsymbol{n}}\Big|_\Gamma, \quad & \end{aligned} \tag{2.24}$$

again, subject to the appropriate Sommerfeld condition at infinity, (1.2a) or (1.2b); the latter guarantees the solvability of (2.24).

In a more general case, when interior sources are present, the results, in fact, do not change. Assume, as before, that the overall acoustic field is the sum of its friendly and adverse components (see (1.4)), $u(\boldsymbol{x}) = u^+(\boldsymbol{x}) + u^-(\boldsymbol{x})$, and we want to cancel out $u^-(\boldsymbol{x})$ on Ω. Then, formula (2.20), where u shall now be interpreted as in (1.4) and w is a function-parameter, will still provide the general solution for surface controls. Indeed, since $\boldsymbol{L}u^+ = 0$ on Ω_1 and, moreover, $u^+(x)$ and $\frac{\partial u^+}{\partial \boldsymbol{n}}(\boldsymbol{x})$ are continuous across the interface Γ, then (2.20) simply reduces to

$$g^{(\mathrm{surf})} = -\left[\frac{\partial \tilde{w}}{\partial \boldsymbol{n}} - \frac{\partial u^-}{\partial \boldsymbol{n}}\right]_\Gamma \delta(\Gamma) - \frac{\partial}{\partial \boldsymbol{n}}\big([\tilde{w}-u^-]_\Gamma \delta(\Gamma)\big), \tag{2.25}$$

where the new function-parameter $\tilde{w}$ is given by $\tilde{w}(\boldsymbol{x}) = w(\boldsymbol{x}) - u^+(\boldsymbol{x})$ and, as such, satisfies the aforementioned general requirements of w's. This means that the control sources $g^{(\mathrm{surf})}(\boldsymbol{x})$, $\boldsymbol{x} \in \Gamma$, defined by (2.20), or equivalently (2.25), appear insensitive to the friendly component $u^+(\boldsymbol{x})$ of the acoustic field, and will precisely annihilate $u^-(\boldsymbol{x})$ on the domain Ω. Furthermore, the entire family of surface controls (2.25) is obviously the exact same family as we would have obtained if there were no interior sources and the overall acoustic field consisted of only $u^-(\boldsymbol{x})$. It is also clear that the same reasoning will apply to the particular cases of purely monopole and purely dipole controls. Namely, if in problems (2.23) and (2.24) we interpreted the boundary data u and $\frac{\partial u}{\partial \boldsymbol{n}}$ (respectively) in the sense of (1.4), then the solution w of either problem would obviously be $w(\boldsymbol{x}) = \tilde{w}(\boldsymbol{x}) + u^+(\boldsymbol{x})$, where $\tilde{w}(\boldsymbol{x})$ is the solution that corresponds to $u^+(\boldsymbol{x}) \equiv 0$. This means that both the monopole and the dipole layers constructed using the respective solution $w(\boldsymbol{x})$ would be the exact same monopole or dipole layer that suppresses $u^-(\boldsymbol{x})$ on Ω in the case of no interior sources and no interior sound.

Altogether we conclude that, as indicated by formula (2.20), surface control sources are combinations of monopole and dipole layers, with the two "extreme" cases corresponding to either only monopoles (see (2.23)) or only dipoles (see (2.24)). From the standpoint of physics, the monopole and dipole sources provide different types of excitation to the surrounding sound-conducting medium. A point monopole source can be interpreted as a vanishingly small pulsating sphere that radiates acoustic waves symmetrically in all directions, whereas a point dipole source resembles a small oscillating membrane that has a particular directivity of radiation (see, e.g., [11] for a more detailed discussion on the properties of different sources). This distinction basically warrants a separate consideration of the monopole- and dipole-type sources as far as the pointwise or surface excitation may be concerned. However, in the context of volumetric excitation, a separate consideration of dipole fields appears, in effect, superfluous.

Indeed, a point dipole is characterized by its (complex) magnitude b and direction $\boldsymbol{m}$ (cf. formula (2.20)):

$$-b\frac{\partial \delta}{\partial \boldsymbol{m}} = -b\langle \boldsymbol{m}, \nabla\delta\rangle = -\langle \boldsymbol{b}, \nabla\delta\rangle, \tag{2.26}$$

where the definition of $\nabla\delta$ is standard, i.e., for any test function $\phi \in \mathcal{D}$ we have $(\nabla\delta, \phi) = -(\delta, \nabla\phi) = -\nabla\phi(0)$, $\langle\cdot,\cdot\rangle$ denotes a conventional real scalar product, and the complex n-dimensional vector $\boldsymbol{b} \equiv b\boldsymbol{m}$ is called the dipole moment.[2] Let us now consider a volumetric distribution of dipoles, i.e., a right-hand side to the Helmholtz equation given in the following convolution form (cf. formula (2.18)):

$$b(\boldsymbol{x}) = -\int_{\mathbb{R}^n} \langle \boldsymbol{b}(\boldsymbol{y}), \nabla\delta(\boldsymbol{x}-\boldsymbol{y})\rangle d\boldsymbol{y} = -\sum_{i=1}^{n} b_i * \nabla_i\delta, \tag{2.27}$$

where $\boldsymbol{b}$ is a regular vector field. Since the convolution of any distribution with the δ-function always exists [24], we can rewrite (2.27) as

$$b(\boldsymbol{x}) = -\sum_{i=1}^{n} \nabla_i b_i * \delta = -\int_{\mathbb{R}^n} \mathrm{div}\boldsymbol{b}(\boldsymbol{y})\delta(\boldsymbol{x}-\boldsymbol{y})d\boldsymbol{y} = -\mathrm{div}\boldsymbol{b} * \delta = -\mathrm{div}\boldsymbol{b}. \tag{2.28}$$

[2]We emphasize that $\boldsymbol{b}$ is not a most general complex vector, but rather a product of a single complex quantity b and a real n-dimensional vector $\boldsymbol{m}$.

Formula (2.28) implies that if we additionally require that the vector field $\boldsymbol{b}(\boldsymbol{y})$ be somewhat more regular than simply $\boldsymbol{b} \in L_1^{(\mathrm{loc})}(\Omega_1)$, namely, $\mathrm{div}\boldsymbol{b} \in L_1^{(\mathrm{loc})}(\Omega_1)$, then the volumetric distribution of dipoles can be reduced to an equivalent volumetric distribution of monopoles as in (2.18).

Next, by differentiating the relation $\boldsymbol{L}G = \delta$, we obtain that the response to the point dipole excitation (2.26) is given by $-\langle \boldsymbol{b}, \nabla G\rangle$. Accordingly, the solution that corresponds to the distribution of sources (2.27) or (2.28) is given by (cf. formula (2.19))

$$
\begin{aligned}
u(\boldsymbol{x}) &= -\int_{\mathbb{R}^n} \langle \boldsymbol{b}(\boldsymbol{y}), \nabla G(\boldsymbol{x}-\boldsymbol{y})\rangle d\boldsymbol{y} = -\sum_{i=1}^{n} b_i * \nabla_i G \\
&= -\sum_{i=1}^{n} \nabla_i b_i * G = -\int_{\mathbb{R}^n} \mathrm{div}\boldsymbol{b}(\boldsymbol{y}) G(\boldsymbol{x}-\boldsymbol{y}) d\boldsymbol{y}.
\end{aligned} \tag{2.29}
$$

Note that, for the representation via the divergence operator to hold, we need to require that the convolutions $b_i * G$ exist in $\mathcal{D}'$; see [24]. A convenient sufficient condition for that may be $\boldsymbol{b}$ having compact support, which, as we have seen in section 2.2, does not present a significant restriction of generality from the standpoint of active control of sound. To make sure that the function $u(\boldsymbol{x})$ of (2.29) does solve the equation $\boldsymbol{L}u = b$, we apply the operator $\boldsymbol{L}$ of (1.1) to it:

$$
\begin{aligned}
\boldsymbol{L}u(\boldsymbol{x}) &= -\boldsymbol{L}\left[\sum_{i=1}^{n} b_i * \nabla_i G\right] = -\sum_{i=1}^{n} \boldsymbol{L}\left[b_i * \nabla_i G\right] \\
&= -\sum_{i=1}^{n} b_i * \boldsymbol{L}\nabla_i G = -\sum_{i=1}^{n} b_i * \nabla_i \boldsymbol{L}G = -\sum_{i=1}^{n} b_i * \nabla_i \delta = b(\boldsymbol{x}).
\end{aligned}
$$

Again, for the foregoing chain of equalities to hold, we need to require the existence of some convolutions in $\mathcal{D}'$, this time $b_i * \nabla_i G$, $i = 1, \ldots, n$. This is guaranteed by the same sufficient condition of $\mathrm{supp}\boldsymbol{b}$ being compact.

From the previous discussion we see that mathematically we can describe volumetric sources of time-harmonic sound only in terms of monopoles. In a real-life acoustic setting, however, both the actual monopole and the dipole sources may be present. It is instrumental to see how those sources that have different physical interpretation enter the right-hand side of the Helmholtz operator. We postpone the corresponding discussion till section 4.1, in which we provide the motivation for selecting a particular optimization criterion. In the meantime, let us emphasize that in the rest of the paper we are going to analyze only one type of the control sources, namely, the monopoles. This class includes all of the volumetric controls (2.1), i.e., monopoles distributed in space, and their limiting case given by the monopole layer on the surface Γ (cf. formula (2.20)):

$$
g^{(\mathrm{surf})}_{\mathrm{monopole}} = -\left[\frac{\partial w}{\partial \boldsymbol{n}} - \frac{\partial u}{\partial \boldsymbol{n}}\right]_{\Gamma} \delta(\Gamma), \tag{2.30}
$$

where $w = w(\boldsymbol{x})$ in (2.30) solves the Dirichlet problem (2.23).[3] In other words, we have narrowed down the class of admissible controls by excluding the surface

[3] As has been mentioned (see section 4.1), volumetric monopoles considered in the mathematical perspective may include contributions from both actual physical monopoles and dipoles.

dipoles. Clearly, having only one and the same type of sources greatly facilitates their description and comparison using the same quantitative characteristics, e.g., acoustic strength (see [11] for the definition and section 4 for more detail). In particular, it allows us to formulate (see section 4.1) and solve (see sections 4.2 and 4.3) the optimization problem for the control sources with the overall absolute acoustic source strength being the cost function. Moreover, the solution of this optimization problem, i.e., the global minimum (more precisely, largest lower bound) in the class of all volumetric controls that guarantee the exact cancellation of noise, happens to be the uniquely defined single layer (2.30), (2.23) on the interface Γ. In this framework, surface dipoles naturally fall out of consideration.

There are, however, optimization criteria that employ physically meaningful quantities other than the total source strength, which naturally justify using combinations of both monopole and dipole control sources. These criteria would typically compare outputs of controls rather than the controls themselves. For example, in the forthcoming paper [8] we employ one such criterion, which is based on the power required by the control system. It turns out that the corresponding analysis necessarily involves interaction between the sources of sound and the surrounding acoustic field. Even though it may seem counterintuitive at a first glance, one can build a control system (a particular combination of monopoles and dipoles) that would require no power input for operation and would even produce a net power gain while providing exact noise cancellation. This, of course, comes at the expense of having the original sources of noise produce even more energy; see [8].

3. Discrete control sources. Similarly to the continuous constructions of the previous section, one can discretize the problem on the grid and obtain the control sources for the discrete formulation. From the standpoint of applications this is, of course, preferred, because any practical design of a noise control system can contain only a finite number of elements or devices (acoustic sensors and actuators) that will be associated with the grid nodes in the discrete case. Details regarding the discrete formulation of the noise control problem can be found in the monograph by Ryaben'kii [15, Part VIII], as well as in the papers [22, 23]; here we provide only a brief account of the corresponding work. The analysis hereafter will not be limited to any specific type of the grid. In particular, no adaptation or grid fitting to either the shape of the protected region Ω (i.e., interface Γ) or that of the external artificial boundary will generally be required. In some cases, though, it may simply be convenient and inexpensive to use a regular grid of an appropriate geometry. For example, as we discuss later, having a polar or spherical grid may greatly simplify setting the ABCs on the circle or sphere, respectively, of radius R in the discrete framework.

3.1. Grids and discretization. Let us now introduce a finite-difference grid $\mathbb{N}$ that would span both Ω and Ω_1. In the discrete formulation, the grid never stretches all the way to infinity; it is always truncated by the external artificial boundary, which implies that the discrete control sources that we obtain will always be compactly supported. Later in section 3.3, we will discuss how to set the appropriate ABCs for the discrete formulation. Now let $u^{(h)}$ be a representation of the acoustic field on the grid, and $\boldsymbol{L}^{(h)}$ be a finite-difference approximation of the differential operator $\boldsymbol{L}$ of (1.1). To accurately define the approximation, we will need to introduce another grid $\mathbb{M}$ along with the previously defined $\mathbb{N}$. On the grid $\mathbb{M}$, we will consider the residuals of the operator $\boldsymbol{L}^{(h)}$, and subsequently the right-hand sides to the corresponding inhomogeneous finite-difference equation. We will use the notations n and m for the

individual nodes of the grids $\mathbb{N}$ and $\mathbb{M}$, respectively, and the notation $\mathbb{N}_m$ for the stencil of the discrete operator $\boldsymbol{L}^{(h)}$ centered at a given $m \in \mathbb{M}$, so that

$$\boldsymbol{L}^{(h)} u^{(h)}\big|_m = \sum_{n \in \mathbb{N}_m} a_{mn} u_n^{(h)}, \tag{3.1}$$

where a_{nm} are the coefficients associated with particular nodes of the stencil. There are no limitations to the type of discrete operators that one may use. We only require that the difference operator $\boldsymbol{L}^{(h)}$ of (3.1) approximate the differential operator $\boldsymbol{L}$ of (1.1) with the accuracy sufficient for a particular application.

Next, we introduce the following subsets of the grids $\mathbb{M}$ and $\mathbb{N}$, which will allow us to accurately distinguish between the interior and exterior domains, interior and exterior sources, and interior and exterior solutions on the discrete level:

$$\begin{gathered} \mathbb{M}^+ = \mathbb{M} \cap \Omega, \quad \mathbb{M}^- = \mathbb{M} \backslash \mathbb{M}^+ = \mathbb{M} \cap \Omega_1, \\ \mathbb{N}^+ = \bigcup_{m \in \mathbb{M}^+}, \quad \mathbb{N}^- = \bigcup_{m \in \mathbb{M}^-} \mathbb{N}_m, \\ \gamma = \mathbb{N}^+ \cap \mathbb{N}^-, \quad \gamma^+ = \mathbb{N}^- \cap \Omega, \quad \gamma^- = \mathbb{N}^+ \cap \Omega_1. \end{gathered} \tag{3.2}$$

We emphasize that the grid $\mathbb{M}$ that pertains to the residuals of the finite-difference operator $\boldsymbol{L}^{(h)}$ is partitioned into $\mathbb{M}^+$ and $\mathbb{M}^-$ directly, i.e., following the geometry of Ω and Ω_1. In contradistinction to that, the grid $\mathbb{N}$ is not partitioned directly; we rather consider the collection of all nodes of $\mathbb{N}$ swept by the stencil $\mathbb{N}_m$ when its center belongs to $\mathbb{M}^+$, and call this subgrid $\mathbb{N}^+$; see (3.2). Obviously, some of the nodes of $\mathbb{N}^+$ obtained by this approach happen to be outside Ω, i.e., in Ω_1, and these nodes are called γ^-. The sets $\mathbb{N}^-$ and γ^+ are defined similarly starting from $\mathbb{M}^-$. The key idea is that whereas the grids $\mathbb{M}^+$ and $\mathbb{M}^-$ do not overlap, the grids $\mathbb{N}^+$ and $\mathbb{N}^-$ do overlap, and their overlap is denoted γ; obviously, $\gamma = \gamma^+ \cup \gamma^-$. The subset of grid nodes γ is called *the grid boundary*; it is a fringe of nodes that is located near the continuous boundary Γ and in some sense straddles it. The specific structure of γ obviously depends on the construction of the operator $\boldsymbol{L}^{(h)}$ of (3.1) and the stencil $\mathbb{N}_m$. For example, for the conventional second-order central-difference Laplacians on rectangular grids, γ will be a two-layer fringe of grid nodes located near Γ, as shown schematically in Figure 3.1. Further specifics on the construction of grid boundaries can be found in the monograph [15].

3.2. Discrete noise control problem and its general solution. Having introduced the discretization (3.1) and grid subsets (3.2), we can formulate and solve the noise control problem on the grid. We will reproduce below the key results of [15, Part VIII]; see also [22, 23] for detail.

The discrete noise control problem is formulated similarly to the continuous one; see section 1. Let $f_m^{(h)+}$, $m \in \mathbb{M}^+$, and $f_m^{(h)-}$, $m \in \mathbb{M}^-$, be the interior and exterior discrete acoustic sources, respectively. Let $u_n^{(h)+}$, $n \in \mathbb{N}$, and $u_n^{(h)-}$, $n \in \mathbb{N}$, be the corresponding solutions, i.e., $\boldsymbol{L}^{(h)} u^{(h)+} = f^{(h)+}$ and $\boldsymbol{L}^{(h)} u^{(h)-} = f^{(h)-}$. Using the same terminology as before, we will call $u^{(h)+}$ the discrete sound and $u^{(h)-}$ the discrete noise. The overall discrete acoustic field $u^{(h)}$ is the sum of its sound and noise components, $u^{(h)} = u^{(h)+} + u^{(h)-}$ on $\mathbb{N}$, and obviously satisfies the equation $\boldsymbol{L}^{(h)} u^{(h)} = f^{(h)} \equiv f^{(h)+} + f^{(h)-}$. The goal is to obtain the discrete control sources $g^{(h)} = g_m^{(h)}$ so that the solution $\tilde{u}^{(h)}$ of the equation $\boldsymbol{L}^{(h)} \tilde{u}^{(h)} = f^{(h)+} + f^{(h)-} + g^{(h)}$ will be equal to only the sound component $u^{(h)+}$ on the subgrid $\mathbb{N}^+$.

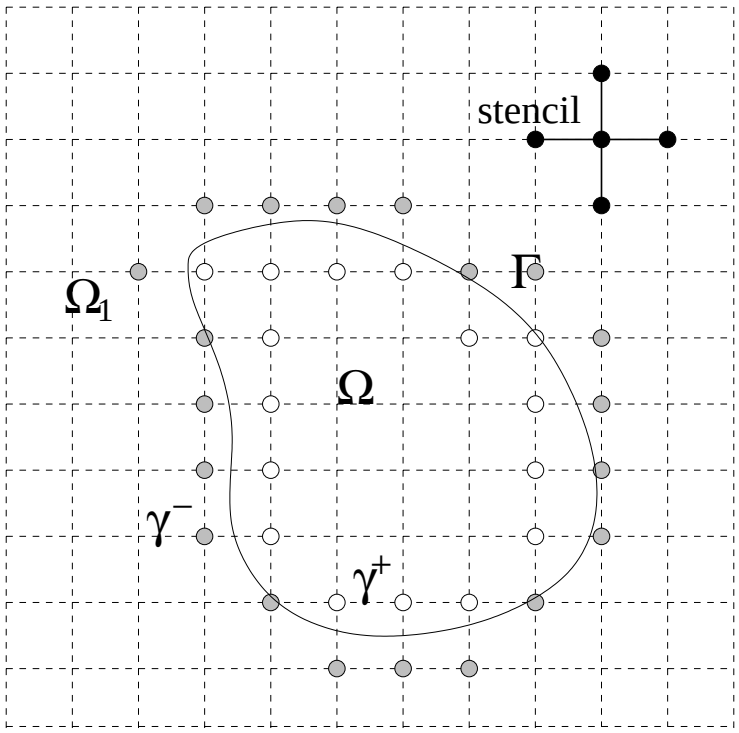


FIG. 3.1. *Schematic geometry of the domains, the stencil, and the grid boundary* $\gamma = \gamma^+ \cup \gamma^-$*: Hollow circles denote* γ^+*, filled circles denote* γ^-*.*

Let us now recall that, in the continuous case, the unique solvability of the governing differential equation (inhomogeneous Helmholtz' equation) was guaranteed by the Sommerfeld radiation conditions (1.2a) or (1.2b) at infinity. In the discrete case, we also need to guarantee the unique solvability of the foregoing finite-difference equations, but we obviously cannot directly set the boundary conditions at infinity. Therefore, the Sommerfeld radiation conditions have to be replaced by some other boundary conditions set at a finite location. To preserve the physics of the model that involves the propagation of waves toward infinity, one may choose to set the appropriate ABCs at the external artificial boundary that truncates our domain.[4] We emphasize that previously, i.e., in the continuous case, we have introduced and used the ABCs (see section 2.2) only for the purpose of obtaining compactly supported controls. Those ABCs were applied to the auxiliary function-parameter $w(\boldsymbol{x})$ (see (2.13) or (2.17)). In the discrete case, the ABCs should apply to the actual solutions $u^{(h)}$, $u^{(h)+}$, $u^{(h)-}$, and $\tilde{u}^{(h)}$, which represent the acoustic fields on the grid. For the purpose of constructing the controls, however, we will never need to implement the ABCs for the acoustic solutions on the grid explicitly. We will only need to know that these boundary conditions can be obtained (a variety of different approaches can be found, e.g., in [19]) and that they will guarantee the solvability of the difference equations involved.

[4]These ABCs would guarantee that the interior solution can be extended beyond the artificial boundary, so that the extension solves the Helmholtz equation and displays the correct far-field behavior.

There will, of course, be an explicit role for the ABCs in the discrete framework as well. We will employ these boundary conditions in the same capacity as we have used the original continuous ABCs (see section 2.2). Namely, when obtaining compactly supported controls, the ABCs will truncate the corresponding discrete function-parameter $w^{(h)}$. A specific approach to constructing the discrete ABCs that we use in this work is based on discretization of the continuous ABCs of section 2.2, and we discuss it in section 3.3. Altogether, both the discrete acoustic fields and the discrete-function parameter that is used for constricting the control sources (see formula (3.3) below) are going to satisfy the same ABCs. This is similar to the continuous case, when both the solution itself (i.e., acoustic field) and the function-parameter w satisfied the same Sommerfeld radiation conditions at infinity.

The general solution for the discrete control sources $g^{(h)} = g_m^{(h)}$ that eliminate the unwanted noise $u^{(h)-}$ on $\mathbb{N}^+$ is given by the following formula (cf. formula (2.1)):

$$g_m^{(h)} = -\boldsymbol{L}^{(h)} w^{(h)}\big|_{m\in\mathbb{M}^-}, \tag{3.3}$$

where $w^{(h)} = w_n^{(h)}$, $n \in \mathbb{N}^-$, is a special auxiliary grid function-parameter that parameterizes the family of controls (3.3). The requirements that this function $w^{(h)}$ must satisfy are, again, rather "loose," and can be considered natural discrete counterparts of the corresponding requirements of the continuous function-parameter $w(\boldsymbol{x})$; see the discussion around formula (2.2) in section 2.1. Namely, at the grid boundary γ the function $w^{(h)}$ has to coincide with the overall acoustic field $u^{(h)}$ to be controlled:

$$w_n^{(h)}\big|_{n\in\gamma} = u_n^{(h)}\big|_{n\in\gamma}. \tag{3.4}$$

We note that since, e.g., for the second-order discretizations the grid boundary γ contains two layers of nodes, γ^+ and γ^- (see Figure 3.1), then specifying the corresponding nodal values on γ is in some sense equivalent to specifying the function and its normal derivative on Γ in the continuous case; see (2.2). Of course, this is not a rigorous statement from the standpoint of approximation; we will address the approximation-related issues later on. We also note that when creating practical designs, the boundary data $u_n^{(h)}\big|_{n\in\gamma}$ shall be interpreted as measurable quantities that provide input for the control system. In other words, we can think of a microphone at every node of γ; these microphones measure the characteristics of the actual acoustic field and generate the input signal $u_n^{(h)}\big|_{n\in\gamma}$.

The other requirement of $w^{(h)}$, besides the interface boundary conditions (3.4), has already been mentioned. The function $w^{(h)}$ must satisfy the appropriate discrete ABCs at a finite external artificial boundary. The role of the discrete ABCs is the same as that of the continuous ABCs—to provide a replacement for the Sommerfeld radiation boundary conditions. This is done in the same approximate sense as the operator $\boldsymbol{L}^{(h)}$ approximates $\boldsymbol{L}$; see section 3.3. Other than the two aforementioned requirements, the function $w^{(h)}$ is arbitrary and, as such, parameterizes a substantial variety of discrete control sources; see (3.3). The latter will provide the search space for optimization in section 4.

The justification for formula (3.3) as the general solution for the discrete control sources is based on the theory of difference potentials; see [15]. In the framework of this theory one can show that the solution $v^{(h)} = v_n^{(h)}$ of the equation $\boldsymbol{L}^{(h)} v^{(h)} = g^{(h)}$ subject to the appropriate ABCs, where $g^{(h)}$ is defined according to (3.3) and (3.4), will be equal to exactly $-u^{(h)-}$ on the interior subgrid $\mathbb{N}^+$: $v_n^{(h)}|_{n\in\mathbb{N}^+} = -u^{(h)-}|_{n\in\mathbb{N}^+}$.

In other words, when the controls $g^{(h)}$ of (3.3) are added to the original source terms of the governing finite-difference equation, they annihilate the unwanted noise on the domain of interest in the discrete sense, i.e., on the grid. The aforementioned solution $v_n^{(h)}\big|_{n\in\mathbb{N}^+}$ is called the generalized difference potential with the density $-u_n^{(h)}\big|_{n\in\gamma}$ defined on the grid boundary γ. It is shown in the theory of difference potentials (see [15]) that the potential depends only on its density $-u_n^{(h)}|_{n\in\gamma}$ and not on the values of the function-parameter outside γ: $w_n^{(h)}|_{n\in\mathbb{N}^-\setminus\gamma}$. Consequently, all possible controls $g^{(h)}$ obtained according to (3.3), with different $w^{(h)}$'s subject only to (3.4) and the corresponding ABCs (see section 3.3), will produce identical output on $\mathbb{N}^+$ that will cancel out the unwanted noise $u_n^{(h)-}|_{n\in\mathbb{N}^+}$. This provides room for optimization of the discrete control sources; see section 4. It can also be shown that every discrete control source $g_m^{(h)}|_{m\in\mathbb{M}^-}$ that cancels out $u_n^{(h)-}|_{n\in\mathbb{N}^+}$ can be represented in the form (3.3) with some function $w^{(h)} = w_n^{(h)}$, $n \in \mathbb{N}^-$, that satisfies (3.4) and the external boundary conditions (ABCs). Similarly to the continuous case, this is done by explicitly constructing the appropriate $w^{(h)}$ for a given $g^{(h)}$ and $u_n^{(h)}|_{n\in\gamma}$; we refer the reader to [15, Part VIII] and [22, 23] for detail.

As has been mentioned, the cancellation of noise in the discrete framework is obtained on the grid $\mathbb{N}^+$. It is important to understand in what sense this discrete cancellation models the continuous cancellation described in section 2. This is basically the question of approximation of the continuous generalized potentials by the discrete ones. To that effect, the theory of difference potentials (see [15]) says that, under certain natural conditions, the difference potential $v^{(h)} = v_n^{(h)}$, $n \in \mathbb{N}^+$, i.e., the solution to $\boldsymbol{L}^{(h)}v^{(h)} = g^{(h)}$ with $g^{(h)}$ given by (3.3), approximates the continuous Calderon's potential $v = v(\boldsymbol{x})$, $\boldsymbol{x} \in \Omega$ (see (2.9)), i.e., the solution to $\boldsymbol{L}v = g$ with g given by (2.1). The aforementioned natural conditions include first the consistency and stability of the finite-difference scheme for the Helmholtz equation. Consistency and stability will guarantee convergence as the grid size vanishes. In addition, the discrete boundary data $u_n^{(h)}|_{n\in\gamma}$ of (3.4) have to approximate the continuous boundary data $(u, \frac{\partial u}{\partial \boldsymbol{n}})|_\Gamma$ of (2.2) in the following sense. Once the continuous function u and its first-order normal derivative $\frac{\partial u}{\partial \boldsymbol{n}}$ are known at the boundary Γ, normal derivatives of higher orders can be obtained via the differential equation itself, and the near-boundary values $u_n^{(h)}|_{n\in\gamma}$ can be calculated using Taylor's expansion; the order of accuracy of the latter calculation with respect to the grid size h has to be at least as high as the order of accuracy of the interior scheme. In this case, the quality of approximation, i.e., the rate of convergence of the discrete potential to the continuous one with respect to h, will be the same as prescribed by the finite-difference scheme itself. For the second-order central-difference schemes discussed in sections 3.3 and 4, this rate is $O(h^2)$. In other words, when designing an active control system following the finite-difference approach, one can expect to have the actual noise cancellation in the same approximate sense as the solution of the finite-difference equation approximates the corresponding solution of the original differential equation. Note that in any particular practical setting we will need to require sufficient wave resolution on the grid, i.e., the waves of length $\lambda = 2\pi/k$, where k is the wavenumber in (1.1), will have to be well resolved by the specific discretization.

3.3. Specific discretization and discrete artificial boundary conditions. As has been mentioned, one can use different approaches to construct discrete ABCs (see, e.g., [19]) that are needed to obtain compactly supported controls. The most

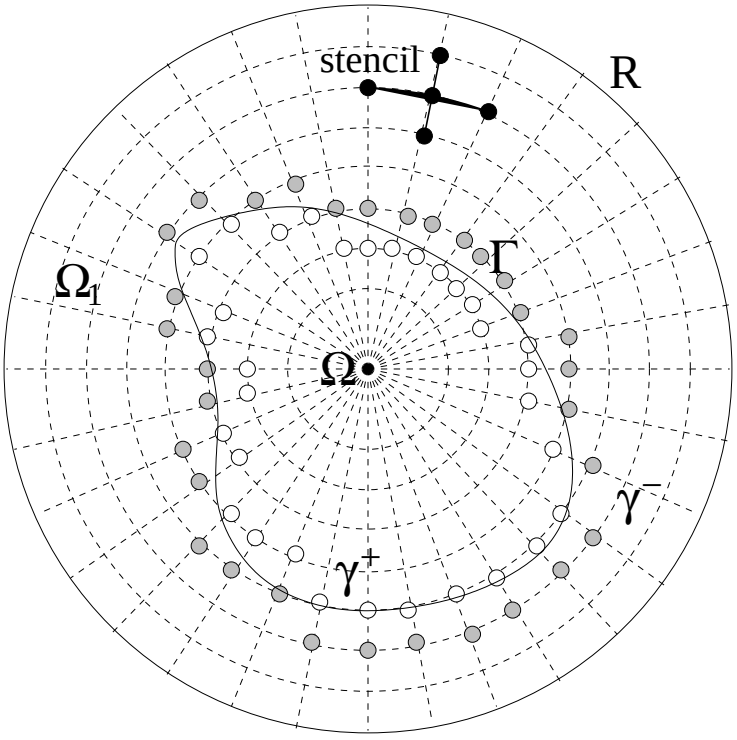


FIG. 3.2. *Schematic geometry of the domains, the stencil, and the grid boundary* $\gamma = \gamma^+ \cup \gamma^-$ *in polar coordinates: Hollow circles denote* γ^+*, filled circles denote* γ^-*.*

straightforward technique, which is adopted in the current study, although it is apparently not the most general one, is to directly approximate the continuous boundary conditions (2.13) or (2.17) with sufficient order of accuracy. For that, we will need a grid that would be fitted to the shape of the external artificial boundary, i.e., a polar or spherical grid. An example of the corresponding grid subsets γ^+ and γ^- for polar coordinates is schematically shown in Figure 3.2.

In all numerical experiments that follow in section 4, we use a two-dimensional setup. Accordingly, we introduce a polar grid that has J cells in the radial direction with the nodes $\rho_j = j\Delta\rho$, $j = 0, \dots, J$, so that $\rho_0 = 0$ and $\rho_J = R$, and L cells in the circumferential direction with the nodes $\theta_s = s\Delta\theta$, $s = 0, \dots, L$, so that $\theta_0 = 0$ and $\theta_L = 2\pi$. For simplicity, it is convenient to assume that the grid sizes $\Delta\rho = R/J$ and $\Delta\theta = 2\pi/L$ are constant; in applications, the grid in the radial direction may be stretched; see section 4.2.

The Helmholtz equation is discretized on this grid with second-order accuracy by central differences:

$$\begin{aligned} \boldsymbol{L}^{(h)} w^{(h)}\big|_{s,j} \equiv \frac{1}{\rho_j}\frac{1}{\Delta\rho}\left(\rho_{j+\frac{1}{2}}\frac{w^{(h)}_{s,j+1}-w^{(h)}_{s,j}}{\Delta\rho} - \rho_{j-\frac{1}{2}}\frac{w^{(h)}_{s,j}-w^{(h)}_{s,j-1}}{\Delta\rho}\right) \\ + \frac{1}{\rho_j^2}\frac{w^{(h)}_{s+1,j}-2w^{(h)}_{s,j}+w^{(h)}_{s-1,j}}{\Delta\theta^2} + k^2 w^{(h)}_{s,j} = 0. \end{aligned} \tag{3.5}$$

The left-hand side of (3.5) is a particular realization of the operator (3.1) that employs the five-node stencil shown in Figure 3.2. This operator will be used in section 4 for obtaining optimal discrete control sources.

To construct the finite-difference ABCs at $\rho = R$, we will also consider a semi-discrete form of the homogeneous equation in the far field:

$$\frac{1}{\rho}\frac{d}{d\rho}\left(\rho\frac{dw_s}{d\rho}\right) + \frac{1}{\rho^2}\frac{w_{s+1} - 2w_s + w_{s-1}}{\Delta\theta^2} + k^2 w_s = 0, \quad s = 0, \ldots, L-1. \tag{3.6}$$

Introducing the direct and inverse discrete Fourier transforms, $l = -L/2+1, \ldots, L/2$, $s = 0, \ldots, L-1$,

$$\hat{w}_l = \frac{1}{L}\sum_{s=0}^{L-1} w_s e^{-ils\Delta\theta}, \qquad w_s = \sum_{l=-L/2+1}^{L/2} \hat{w}_l e^{ils\Delta\theta}, \tag{3.7}$$

we reduce (3.6) to the following system of ordinary differential equations with respect to $\hat{w}_l = \hat{w}_l(\rho)$:

$$\frac{1}{\rho}\frac{d}{d\rho}\left(\rho\frac{d\hat{w}_l}{d\rho}\right) - \frac{\alpha_l^2}{\rho^2}\hat{w}_l + k^2\hat{w}_l = 0, \quad \alpha_l^2 = \frac{4}{\Delta\theta^2}\sin^2\frac{l\Delta\theta}{2}, \quad \rho \geq R, \tag{3.8}$$

where again $l = -L/2+1, \ldots, L/2$. Equations (3.8) are the same as (2.14), except that in the discrete case the range for l is finite, and l^2 in (2.14) has been replaced by α_l^2 in (3.8). Therefore, we can use the same boundary conditions (2.17) for $l = -L/2+1, \ldots, L/2$,

$$\frac{d}{d\rho}\hat{w}_l\bigg|_{\rho=R} = \hat{w}_l(R)\frac{\frac{d}{d\rho}H_{\alpha_l}^{(2)}(kR)}{H_{\alpha_l}^{(2)}(kR)}, \tag{3.9}$$

only with the Hankel functions of order l replaced by the Hankel functions of the order α_l. For implementation in the foregoing discrete framework, boundary conditions (3.9) for all $l = -L/2+1, \ldots, L/2$ have to be approximated with second-order accuracy, which can be easily done as follows:

$$\frac{\hat{w}_{l,J} - \hat{w}_{l,J-1}}{\Delta\rho} - \beta_l\frac{\hat{w}_{l,J} + \hat{w}_{l,J-1}}{2} = 0, \qquad \beta_l = \frac{\frac{d}{d\rho}H_{\alpha_l}^{(2)}(kR)}{H_{\alpha_l}^{(2)}(kR)}. \tag{3.10}$$

Finally, relations in (3.10) for all $l = -L/2+1, \ldots, L/2$ can be rewritten in the matrix form:

$$\boldsymbol{w}_{\cdot,J} = \boldsymbol{F}^{-1}\mathrm{diag}\left\{-\left(\frac{1}{\Delta\rho} + \beta_l\right)\left(\frac{1}{\Delta\rho} - \beta_l\right)^{-1}\right\}\boldsymbol{F}\,\boldsymbol{w}_{\cdot,J-1} \equiv \boldsymbol{T}\,\boldsymbol{w}_{\cdot,J-1}, \tag{3.11}$$

where $\boldsymbol{F}$ and $\boldsymbol{F}^{-1}$ are matrices of the direct and inverse discrete Fourier transforms of (3.7), and $\boldsymbol{w}_{\cdot,J}$ and $\boldsymbol{w}_{\cdot,J-1}$ are L-dimensional vectors of components $w_{s,J}^{(h)}$ and $w_{s,J-1}^{(h)}$, respectively, $s = 0, 1, \ldots, L-1$.

In the three-dimensional case, instead of the discrete Fourier transforms (3.7) one can use expansions with respect to the so-called finite-difference spherical functions (see [15, Part IV, Chapter 4]) that form a full orthogonal system of eigenvectors for the spherical part of the discrete Laplacian. Other than that, the construction of the

discrete three-dimensional ABCs will be similar to the foregoing two-dimensional construction. We do not expand on it here because we do not conduct three-dimensional computations in this paper. Let us also mention that for small l the difference between l^2 and the corresponding α_l^2 (see (3.8)) will obviously be small as well. Therefore, for smooth functions w, for which the short-wave part of the spectrum (large l's) is insignificant, one may not even have to replace l in (2.17) by α_l in (3.9). We have observed this type of behavior in the previous paper [14], in which we studied similar questions for the Poisson equation.

3.4. Types of discrete control sources. Similarly to the continuous case (see section 2.3) let us now identify some particular types of discrete control sources. First, we define another subset of the grid $\mathbb{M}$ (more precisely, of $\mathbb{M}^-$):

$$\mathbb{M}^-_{\text{int}} = \left\{ m \in \mathbb{M}^- \middle| \mathbb{N}_m \cap \gamma^+ = \emptyset \right\}.$$

Basically, $\mathbb{M}^-_{\text{int}}$ is the interior subset of $\mathbb{M}^-$ such that, when the center of the stencil sweeps this subset, the stencil itself does not touch γ^+; see Figures 3.1 and 3.2. In other words, we can say that $\mathbb{M}^-_{\text{int}}$ is a subset of $\mathbb{M}^-$ such that

$$\bigcup_{m \in \mathbb{M}^-_{\text{int}}} \mathbb{N}_m = \mathbb{N}^- \backslash \gamma^+.$$

Having defined this new subset $\mathbb{M}^-_{\text{int}}$, we now introduce the auxiliary function $w^{(h)} = w_n^{(h)}$, $n \in \mathbb{N}^-$, for (3.3) as follows:

$$w_n^{(h)}\big|_{n\in\gamma^+} = u_n^{(h)}\big|_{n\in\gamma^+}, \tag{3.12a}$$

and

$$\begin{aligned} &w_n^{(h)}\big|_{n\in\gamma^-} = u_n^{(h)}\big|_{n\in\gamma^-}, \\ &\boldsymbol{L}^{(h)} w^{(h)} = 0 \ \text{ on } \ \mathbb{M}^-_{\text{int}}. \end{aligned} \tag{3.12b}$$

As before, we also assume that $w^{(h)}$ satisfies the appropriate discrete ABCs; see, e.g., (3.11). Definition (3.12a) means that on the interior part of the grid boundary γ^+ we simply set $w^{(h)}$ equal to the given $u^{(h)}$: $w_n^{(h)}|_{n\in\gamma^+} = u_n^{(h)}|_{n\in\gamma^+}$. Definition (3.12b) is actually a discrete exterior boundary-value problem of the Dirichlet type. Indeed, everywhere on and "outside" the exterior part of the grid boundary γ^-, i.e., on $\mathbb{N}^-\backslash\gamma^+$, the grid function $w^{(h)}$ is obtained as a solution of the homogeneous equation $\boldsymbol{L}^{(h)} w^{(h)} = 0$ (enforced at the nodes $\mathbb{M}^-_{\text{int}}$) supplemented by the boundary data on γ^-: $w_n^{(h)}\big|_{n\in\gamma^-} = u_n^{(h)}\big|_{n\in\gamma^-}$, which is specified for the unknown function $w^{(h)}$ itself. Note, relation (3.12a) and the first relation (3.12b) together are obviously equivalent to (3.4). Therefore, the function $w^{(h)}$ defined via (3.12a), (3.12b) falls into the general class of $w^{(h)}$'s used for obtaining the discrete control sources; see section 3.2.

Problem (3.12b) can clearly be considered a finite-difference counterpart to the continuous Dirichlet problem (2.23). Therefore, it is natural to call the control sources $g^{(h)} \equiv g^{(\text{h,surf})}_{\text{monopole}}$ obtained by formulae (3.3), (3.12a), (3.12b) *the discrete surface monopoles.* Indeed, because of the definition of $w^{(h)}$ given by (3.12a) and (3.12b), these $g^{(\text{h,surf})}_{\text{monopole}}$ may, generally speaking, differ from zero only on the grid set $\mathbb{M}^-\backslash\mathbb{M}^-_{\text{int}}$, which is a single "curvilinear" layer of nodes of grid $\mathbb{M}$ that follows the geometry

of Γ. Accordingly, the output of these controls can be called the discrete single-layer potential; it was first introduced and analyzed in our recent paper [20]. Let us emphasize that unlike the continuous surface monopoles (2.30), which belong to a different class of functions rather than the volumetric sources (2.1) and (2.2) (singular δ-type distributions vs. regular locally integrable functions), the foregoing discrete surface monopoles belong to the same original class of discrete control sources (3.3) and (3.4). They can be considered as the ultimate reduction of the volumetric discrete controls (3.3) and (3.4) to the surface. In section 4, the discrete surface monopoles will play a fundamental role for the analysis of the optimization problems.

Besides the discrete surface monopoles and the corresponding single-layer potential, one can also define the discrete surface dipoles and, accordingly, the double-layer potential; see [20]. Grid dipoles are introduced for the pairs of neighboring nodes so that the nodes in the pair are assigned values equal in magnitude and opposite in sign. The control sources in the form of discrete surface dipoles can be obtained by solving a special Neumann-type discrete exterior boundary-value problem for the auxiliary function $w^{(h)}$, which would be analogous to the continuous problem (2.24). The construction of surface dipoles, however, is somewhat more elaborate than the foregoing definition of surface monopoles. And because in this paper we basically focus on the monopole-type sources only, we are not going to further elaborate here on the issue of discrete surface dipoles, but will rather refer the reader to our paper [20] for detail.

4. Optimization of control sources. Once the general solution for controls is available, in either continuous (2.1) or discrete (3.3) formulation, the next step is to decide what particular element of this large family of functions will be optimal for a specific setting. There is a multitude of possible criteria for optimality that one can use; we discuss some of them in the forthcoming papers [8, 9]. We should also emphasize that in many practical problems the cancellation of noise is only approximate, and, as such, the key criterion for optimization (or sometimes, the key constraint) is the quality of this cancellation, i.e., the extent of noise reduction. In contradistinction to that, in this paper we are considering ideal, or exact, cancellation; i.e., every particular control field from either the continuous (2.1) or discrete (3.3) family completely eliminates the unwanted noise on the domain of interest. Consequently, the criteria for optimality of the controls that we can employ will not include the level of the residual noise as a part of the corresponding function of merit, and should rather depend only on the control sources themselves. We realize, of course, that at a later stage of the work we will also need to look into the issues of approximate, rather than exact, noise cancellation, for the reason of further reducing the costs. In this case, optimal solutions found in the framework of the exact cancellation are likely to provide good initial guesses for subsequent optimization in the approximate framework. Moreover, it will probably be possible to use some results from the approximation theory to deal with the issues of approximate noise cancellation once we have solutions for the exact cancellation. We expect that this approach will be much faster than any algorithm of combinatorial type. A similar reduction in computational complexity of optimization was outlined in our earlier work [6] on the optimal distributed control of the exterior Stokes flow.

4.1. Optimization in the sense of L_1. To derive a meaningful criterion for optimization of the control sources, let us first discuss the physical meaning of the quantities involved in the formulation of the problem. The most natural way to interpret the field variable $u = u(\boldsymbol{x})$ (as well as its discrete counterpart $u_n^{(h)}$) is to call it acoustic pressure. Indeed, acoustic pressure is the quantity which is directly measured

by the sensing devices (microphones), and as such can be immediately supplied as the required boundary input data for the control system; see formulae (2.2) and (3.4). We now recall that in the current paper we only analyze a single-frequency formulation of the noise control problem. To better understand the nature of the source terms in the Helmholtz equation that governs the time-harmonic pressure $u = u(\boldsymbol{x})$, we will now examine the original unsteady acoustic formulation. Let $p = p(\boldsymbol{x}, t)$ be the actual acoustic pressure, and $\boldsymbol{v} = \boldsymbol{v}(\boldsymbol{x}, t)$ be the velocity of fluid particles. Then, the acoustics system can be written as

$$\begin{aligned} \frac{\partial \rho(\boldsymbol{x}, t)}{\partial t} + \rho_0 \operatorname{div} \boldsymbol{v}(\boldsymbol{x}, t) &= \rho_0 q_{\mathrm{vol}}(\boldsymbol{x}, t), \\ \rho_0 \frac{\partial \boldsymbol{v}(\boldsymbol{x}, t)}{\partial t} + \operatorname{grad} p(\boldsymbol{x}, t) &= \boldsymbol{b}_{\mathrm{vol}}(\boldsymbol{x}, t), \end{aligned} \tag{4.1}$$

where ρ_0 is the density of the ambient fluid. The quantity q_{vol} on the right-hand side of the continuity equation in (4.1) is known as the volume velocity per unit volume; see, e.g., [10, 11]. It is defined through the actual volume velocity $q = \int_S v_n d\sigma$, which is the integral of the normal component of the fluid particles' velocity v_n evaluated over a closed surface S; then $q_{\mathrm{vol}} = \lim_{V \to +0} \frac{1}{V} \int_S v_n d\sigma$, where V is the volume enclosed by S. The physical meaning of the sources $\rho_0 q_{\mathrm{vol}}$ is that they excite the medium by altering the balance of mass in the system, i.e., by injecting/draining certain amounts of fluid. Clearly, in the time-harmonic context the process of injecting/draining the fluid has to be periodic.

Similarly, the quantity $\boldsymbol{b}_{\mathrm{vol}}$ on the right-hand side of the second equation of (4.1) shall be interpreted as the force per unit volume. Obviously, it excites the medium by altering the balance of momentum in the system; and again, in the time-harmonic context the net force applied to the fluid particles has to be periodic. Altogether, we see that two types of sources can be introduced in unsteady acoustics that are distinctly different from the standpoint of physics. Next, we will see that they can be interpreted as monopoles and dipoles, respectively, as introduced in section 2.3.

We apply the adiabatic law $p = c^2 \rho$, where c is the speed of sound (constant), differentiate the first equation of (4.1) with respect to time, take the divergence of the second equation of (4.1), and substitute into the first one, which yields the inhomogeneous wave equation for the acoustic pressure:

$$-\frac{1}{c^2} \frac{\partial^2 p(\boldsymbol{x}, t)}{\partial t^2} + \Delta p(\boldsymbol{x}, t) = \operatorname{div} \boldsymbol{b}_{\mathrm{vol}}(\boldsymbol{x}, t) - \rho_0 \frac{\partial q_{\mathrm{vol}}(\boldsymbol{x}, t)}{\partial t}. \tag{4.2}$$

The second term on the right-hand side of (4.2) is the volume acceleration per unit volume multiplied by the ambient fluid density. The Helmholtz equation for the time-harmonic pressure is obtained by Fourier transforming (4.2) in time, which formally amounts to replacing the temporal derivatives $\frac{\partial}{\partial t}(\,\cdot\,)$ by $-i\omega(\,\cdot\,)$. For simplicity we will keep all the notations the same except that the time-harmonic problem quantities will depend only on $\boldsymbol{x}$ and not on t, and we will also use the previous notation $u(\boldsymbol{x})$ for the Fourier transformed pressure $p(\boldsymbol{x}, t)$:

$$\Delta u(\boldsymbol{x}) + k^2 u(\boldsymbol{x}) = \operatorname{div} \boldsymbol{b}_{\mathrm{vol}}(\boldsymbol{x}) + i\omega \rho_0 q_{\mathrm{vol}}(\boldsymbol{x}). \tag{4.3}$$

The wavenumber k in the Helmholtz equation (4.3) is given by $k = \omega/c$, where ω is the frequency of the original temporal oscillations.

Clearly, (4.3) is basically the same as (1.1), except that in (4.3) we provide a detailed description of the source terms based on their physical origin. Namely, we see that the scalar sources q_{vol} that alter the acoustic balance of mass (see (4.1)) enter the

right-hand side of the Helmholtz equation directly (up to a multiplicative constant), whereas the sources $\boldsymbol{b}_{\rm vol}$ that alter the acoustic balance of momentum (see (4.1)) enter the right-hand side of the Helmholtz equation through a divergence operator. Therefore, the analysis of section 2.3 allows us to interpret $q_{\rm vol}(\boldsymbol{x})$ as genuine volumetric monopoles and $\boldsymbol{b}_{\rm vol}(\boldsymbol{x})$ as volumetric dipoles that are rewritten as equivalent monopoles $\mathrm{div}\boldsymbol{b}_{\rm vol}(\boldsymbol{x})$ for mathematical convenience.

The latter operation, namely, recasting the physical dipoles into the monopole form, allows us to qualitatively study and optimize both types of sources in a uniform manner. In acoustics (see [10]), the overall right-hand side $f(\boldsymbol{x}) = \mathrm{div}\boldsymbol{b}_{\rm vol}(\boldsymbol{x}) + i\omega\rho_0 q_{\rm vol}(\boldsymbol{x})$ of (4.3) is often referred to as *the acoustic source density*. As we have seen, the meaning of this right-hand side is excitation per unit volume. Accordingly, the integral of this quantity over a given region, $\int_V f d\boldsymbol{x}$, is referred to as *the acoustic source strength* that pertains to the sources in this region, and the integral of its magnitude, $\int_V |f| d\boldsymbol{x}$, is known as *the absolute acoustic source strength.* If the actual sources involved in the consideration were only monopoles, then the acoustic source strength would obviously coincide with the overall volume velocity q, and the acoustic source density would coincide with the volume velocity per unit volume $q_{\rm vol}$.

For distributed sources, the acoustic source density is assumed to be finite; in particular, for distributed genuine monopoles the associated volume velocity per unit volume is assumed to be finite. In the case of an isolated point monopole, i.e., a δ-type source (see section 2.3), which can be represented as a vanishingly small oscillating sphere of radius ϵ with surface velocity v_ϵ, the volume velocity can be introduced as $q = \lim_{\epsilon\to+0} v_\epsilon 4\pi\epsilon^2$. When the strength q of this isolated monopole is finite, the associated source density is formally infinite (which is natural to expect for a δ-type source). In the case of a continuous distribution of sources, the relationship between the source density and source strength is standard (like that between the mass density and total mass, electric charge density and total charge, etc.) and basically says that the integral of the source density over a given region is equal to the overall source strength associated with this region. In other words, the acoustic source density is equal to the acoustic source strength per unit volume.

Obviously, the physical meaning of the quantity $g = g(\boldsymbol{x})$ of (2.1) that describes the control sources is the same as that of the original right-hand side $f = f(\boldsymbol{x})$ of (1.1) that we have recently specified according to (4.3). Namely, $g(\boldsymbol{x})$ shall be interpreted as *the acoustic source strength per unit volume* (up to a multiplicative constant) of the control sources. It is important to mention that as we are studying the time-harmonic traveling waves, all the quantities involved in the formulation of the problem are complex-valued. This is essential, as otherwise it would not have been possible to account for the key phenomenon of the variation of phase between different spatial locations.

Having identified the physical meaning of the variables involved in the noise control model that we have adopted, we would argue in the current paper for selecting the optimal control sources based on minimization of their *overall absolute acoustic source strength.* Mathematically, this translates into the minimization of the L_1 norm of the control sources:

$$\|g\|_1 \equiv \int_{\mathrm{supp}\, g} |g(\boldsymbol{x})| d\boldsymbol{x} \longrightarrow \min, \tag{4.4}$$

where the search space for minimization in (4.4) includes all the appropriate auxiliary functions $w(\boldsymbol{x})$, by means of which the controls $g(\boldsymbol{x})$ are defined (see formulae (2.1), (2.2), and the discussion in the beginning of section 2.1). The advantage of using

this criterion for optimization is that it has a clear physical interpretation, and the quantities involved, the volume velocity as well as the force applied to fluid particles, actually characterize the corresponding engineering devices (actuators in the active noise control system). For comparison, we note that the criterion based on the L_2 norm,

$$\|g\|_2 \equiv \sqrt{\int_{\operatorname{supp} g} |g(\boldsymbol{x})|^2 d\boldsymbol{x}} \longrightarrow \min,$$

does not have a similar clear physical interpretation, although as indicated below (see also our forthcoming paper [9]), the corresponding numerical optimization problem is much easier to solve. Another advantage of using the L_1 norm, or in other words, the overall absolute acoustic source strength, as the cost function for optimization (minimization) is that it characterizes only the control sources themselves. This is a convenient distinction compared, e.g., to the power-based criteria, which, as has been mentioned, would always involve interaction between the sources and the field they operate in. This interaction is often referred to as the "load" on the sources by the field (see [11]), and may lead to certain types of degeneration when solving the optimization problem; see [8].

In the discrete framework, the L_1 minimization problem that corresponds to (4.4) is formulated as follows:

$$\|g^{(h)}\|_1 \equiv \sum_{m \in \mathbb{M}^- \cap \{\rho < R\}} V_m |g_m^{(h)}| \longrightarrow \min, \tag{4.5}$$

where V_m accounts for the volume in three dimensions or area in two dimensions of a particular grid cell, and again, the search space includes all the appropriate auxiliary grid functions $w^{(h)}$ through which $g^{(h)}$ is defined; see formula (3.3). The function $w^{(h)}$ is supposed to satisfy boundary conditions (3.4) on the interface, and the selected ABCs at the external artificial boundary; for the two-dimensional examples analyzed in the following section 4.2 the latter will be boundary conditions (3.11).

Let us now return to the definition (3.3) of the discrete control sources $g^{(h)}$ and adopt the polar framework of section 3.3. The finite-difference operator $\boldsymbol{L}^{(h)}$ can obviously be interpreted as a matrix with N columns and M rows, where N is the number of nodes $n \equiv (s, j)$ of the grid $\mathbb{N}^-$ such that the corresponding radial coordinate $\rho_j \le R$, i.e., $j \le J$, and M is the number of nodes $m \equiv (s, j)$ of the grid $\mathbb{M}^-$ such that the corresponding radial coordinate $\rho_j < R$, i.e., $j \le J - 1$. Denote by $\boldsymbol{w}$ the vector of N components $w_n^{(h)} \equiv w_{s,j}^{(h)}$ such that $n \in \mathbb{N}^-$ and $j \le J$. The components of $\boldsymbol{w}$ can obviously be arranged in a particular way so that this vector can then be decomposed into four subvectors:

$$\boldsymbol{w} = [\boldsymbol{w}_\gamma, \boldsymbol{w}_0, \boldsymbol{w}_{\cdot,J-1}, \boldsymbol{w}_{\cdot,J}]^T, \tag{4.6}$$

where $\boldsymbol{w}_\gamma$ contains all those and only those $w_n^{(h)}$ for which $n \in \gamma$, $\boldsymbol{w}_{\cdot,J}$, and $\boldsymbol{w}_{\cdot,J-1}$ correspond to the outermost and second-to-last circles of the polar grid, respectively, as in formula (3.11), and $\boldsymbol{w}_0$ contains all the remaining components of $\boldsymbol{w}$. In accordance with (4.6), the matrix $\boldsymbol{L}^{(h)}$ can be decomposed into four submatrices:

$$\boldsymbol{L}^{(h)} = [\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}], \tag{4.7}$$

where the number of rows in all four is the same and equal to M, $\boldsymbol{A}$ has as many columns as there are nodes in γ (we denote this number $|\gamma|$), $\boldsymbol{C}$ and $\boldsymbol{D}$ each have L columns (see section 3.3), and the number of columns in $\boldsymbol{B}$ is obviously $N - |\gamma| - 2L$.

Using representations (4.6) and (4.7), one can rewrite the optimization problem (4.5) as follows:

$$\| \boldsymbol{V}(\boldsymbol{A}\boldsymbol{w}_\gamma + \boldsymbol{B}\boldsymbol{w}_0 + \boldsymbol{C}\boldsymbol{w}_{\cdot,J-1} + \boldsymbol{D}\boldsymbol{w}_{\cdot,J}) \|_1 \longrightarrow \min, \tag{4.8}$$

where the norm in (4.8) is a conventional L_1 norm on complex M-dimensional vectors, and $\boldsymbol{V}$ is an $M \times M$ diagonal matrix with the entries given by the corresponding cell areas V_m. Next, we recall that the search space for optimization (4.5) is composed of all the appropriate grid functions $w^{(h)}$, which means that the vector $\boldsymbol{w}$ in the optimization formulation (4.8) is, in fact, subject to a number of equality-type constraints that come from the interface conditions (3.4) and ABCs (3.11). More precisely, the first subvector $\boldsymbol{w}_\gamma$ in (4.6) is known and fixed because of (3.4), and we can rewrite (3.4) as $\boldsymbol{w}_\gamma = \boldsymbol{u}_\gamma$, where $\boldsymbol{u}_\gamma$ is given. The last subvector $\boldsymbol{w}_{\cdot,J}$ in (4.6) is a function of $\boldsymbol{w}_{\cdot,J-1}$ according to (3.11). Therefore, we can conclude that only $\boldsymbol{w}_0$ and $\boldsymbol{w}_{\cdot,J-1}$ contain free variables that provide the search space for optimization, and as such rewrite (4.8) as

$$\min_{\boldsymbol{w}_0, \boldsymbol{w}_{\cdot,J-1}} \| \boldsymbol{V}(\boldsymbol{B}\boldsymbol{w}_0 + (\boldsymbol{C} + \boldsymbol{D}\boldsymbol{T})\boldsymbol{w}_{\cdot,J-1} + \boldsymbol{A}\boldsymbol{w}_\gamma) \|_1 \equiv \min_{\boldsymbol{z}} \| \boldsymbol{E}\boldsymbol{z} - \boldsymbol{f} \|_1, \tag{4.9}$$

where $\boldsymbol{E} = \boldsymbol{V}[\boldsymbol{B}, \boldsymbol{C} + \boldsymbol{D}\boldsymbol{T}]$ is an $M \times (N - |\gamma| - L)$ given matrix, $\boldsymbol{z} = [\boldsymbol{w}_0, \boldsymbol{w}_{\cdot,J-1}]^T$ is an $(N - |\gamma| - L)$-dimensional vector of unknowns, and $\boldsymbol{f} = -\boldsymbol{V}\boldsymbol{A}\boldsymbol{w}_\gamma$ is an M-dimensional known vector of the right-hand side. Minimization problem (4.9) is, in fact, a problem of finding a weak solution in the sense of L_1 of an overdetermined complex linear system $\boldsymbol{E}\boldsymbol{z} = \boldsymbol{f}$.

Let us first note that the most conventional weak formulation for an overdetermined system $\boldsymbol{E}\boldsymbol{z} = \boldsymbol{f}$ would be that in the sense of L_2, rather than (4.9). The L_2 minimization problem $\| \boldsymbol{E}\boldsymbol{z} - \boldsymbol{f} \|_2 \longrightarrow \min$ has proven easy to solve numerically even for rather complex geometries. It does not require the Moore–Penrose-type arguments and can be conveniently solved by a standard QR algorithm; we report the corresponding results in our forthcoming paper [9]. As has been mentioned, though, this formulation lacks a convincing physical interpretation and therefore, hereafter we concentrate on solving the L_1 optimization problem (4.9).

By introducing M additional real variables $t_i \in \mathbb{R}$, $i = 1, \ldots, M$, one can reduce problem (4.9) to the following optimization problem with equality-type constraints,

$$\begin{gathered} \min \sum_i t_i, \\ \left| \sum_j e_{ij} z_j - f_i \right| - t_i = 0, \quad i = 1, \ldots, M, \end{gathered} \tag{4.10}$$

which is equivalent to the problem with inequality-type constraints:

$$\begin{gathered} \min \sum_i t_i, \\ \left| \sum_j e_{ij} z_j - f_i \right| - t_i \le 0, \quad i = 1, \ldots, M. \end{gathered} \tag{4.11}$$

If all the quantities involved in the formulation (4.9) were real, then problem (4.11) would, in turn, be equivalent to the linear programming problem (see [21, Chapter 12, section 4]):

$$\begin{aligned} &\min \sum_i t_i, \\ &-t_i \le \sum_j e_{ij} z_j - f_i \le t_i, \quad i = 1, \dots, M, \end{aligned} \tag{4.12}$$

which nowadays can be solved efficiently even for large dimensions. However, complex entries in $\boldsymbol{E}$, $\boldsymbol{z}$, and $\boldsymbol{f}$ (see (4.9)) are essential in order to account for traveling waves, so we will actually need to solve *a nonlinear problem* (4.11) rather than a linear problem (4.12).

The most obvious disadvantage of optimizing in the sense of L_1 is that the foregoing problem (4.11) appears very difficult to solve numerically. Besides being nonlinear, the constraints are obviously nonsmooth. Moreover, strictly speaking, those constraints are not convex either. Indeed, for every $i = 1, \dots, M$, the inequality $|\sum_j e_{ij} z_j - f_i| - t_i \le 0$ defines a cone in the space of variables t_i, $\Re(\sum_j e_{ij} z_j - f_i)$, and $\Im(\sum_j e_{ij} z_j - f_i)$. As we are only considering the upper half of the cone, this set is geometrically convex. However, algebraically the function $|\sum_j e_{ij} z_j - f_i|^2 - t_i^2$ of variables z_j, t_i obviously cannot be convex. And the algebraic convexity (i.e., positive semidefiniteness of the Hessian) is exactly what distinguishes between the convex and nonconvex programming problems, with the latter being substantially more difficult to treat in a numerical setting; see [21, Chapter 24]. Of course, problem (4.11) can be reformulated so that the constraints will become truly convex:

$$\begin{aligned} &\min \sum_i \sqrt{t_i}, \\ &\left| \sum_j e_{ij} z_j - f_i \right|^2 - t_i \le 0, \quad i = 1, \dots, M. \end{aligned} \tag{4.13}$$

However, in the formulation (4.13) the most "harmless" cost function that one can think of, i.e., the linear function $\sum_i t_i$, has been replaced by the function $\sum_i \sqrt{t_i}$ that has singular derivatives at the optimum. Experimentally, we have observed that this presents even more severe problems for a numerical optimizer.

Altogether, the combination of nonlinearity, nonsmoothness, and only "marginal" convexity (optimization over cones) makes problem (4.11) a serious challenge even for the most sophisticated state-of-the-art approaches to numerical optimization—the approaches that are typically based on interior point methods [12, 21]. The difficulties are further exacerbated by the large dimension of the grid on which the problem is formulated. Even for the aforementioned state-of-the art methods the maximum number of constraints that they can handle is typically on the order of hundreds. And in problem (4.11), the number of constraints is the same as the number of grid nodes M. As such, one can easily encounter an orders of magnitude difference between the number of constraints that the numerical optimizer will handle and the number of grid nodes that will make the formulation of an active noise control problem practically interesting. This is especially true for three-dimensional problems.

In spite of all difficulties, we have still managed to obtain numerical solutions in two space dimensions for some simple test cases. All numerical experiments that we have conducted indicate a very consistent behavior of the L_1-optimal solution for control sources; it happens to be the discrete layer of monopoles on the surface $g^{(\mathrm{h,\,surf})}_{\mathrm{monopole}}$ described in section 3.4. Recall, this solution is obtained by applying formula (3.3) to the auxiliary function $w^{(h)}$ defined by (3.12a), (3.12b). In the following section 4.2, we report the corresponding computational results, and in the subsequent section 4.3, we provide a general proof of the global L_1-optimality of this surface monopole solution in the case of one space dimension. Overall, the combination of the two-dimensional numerical evidence and the one-dimensional general proof prompts us to put forward a conjecture (see section 5) that the foregoing uniquely defined layer of surface monopoles always provides the control sources with minimal total absolute strength. This conjecture implies that the difficult procedure of numerical optimization in the sense of L_1 *can actually be bypassed* when building the L_1-optimal control sources; the latter can be obtained by simply solving the boundary-value problem (3.12b), which is an easy task. In other words, assuming that our arguments toward global minimality of surface monopoles are sufficiently convincing (see sections 4.2 and 4.3), we can claim that finding the L_1-optimal controls will now take only a most straightforward computation, apparently even easier than optimization in the sense of L_2; see [9]. Of course, comparison of the different optimization strategies in the framework of an approximate, rather than exact, noise cancellation will require a through future study.

4.2. Numerical solution of the L_1-optimization problem. For our numerical simulations, we have considered the simplest possible two-dimensional geometric setup, with the protected domain Ω in the form of a disk of radius $r = 1$ centered at the origin. The external artificial boundary was a circle of radius $R > r$, as in section 3.3. As such, the resulting discrete control sources were concentrated within the annular region $r \le \rho \le R$.

In contradistinction to section 3.3, here we have used a polar grid, which was stretched in the radial direction. This allowed us to keep the cell aspect ratio constant. The grid is first built in the coordinates $(\ln\rho, \theta)$; it has equal square cells $\frac{2\pi}{L} \times \frac{2\pi}{L}$ and is constructed on the rectangle $[-\frac{2\pi}{L}, \ln R] \times [0, 2\pi]$. Then, the conformal mapping $e^{\ln\rho + i\theta}$ maps it onto a polar grid with uniform angular spacing $\theta_s = s\Delta\theta$, where $\Delta\theta = \frac{2\pi}{L}$ and $s = 0, \ldots, L$, so that $\theta_0 = 0$ and $\theta_L = 2\pi$, and nonuniform radial spacing $\rho_j = \exp\big(\frac{2\pi}{L} \cdot j\big)$, $j = -1, 0, \ldots, J$, so that $\rho_{-1} = \exp\big(-\frac{2\pi}{L}\big)$, $\rho_0 = 1 = r$, and $\rho_J = R$. It is convenient to define the grid sizes in the radial direction as $\Delta\rho_j \equiv \rho_j - \rho_{j-1} = \exp\big(\frac{2\pi}{L} \cdot j\big) - \exp\big(\frac{2\pi}{L} \cdot (j-1)\big)$, $j = 0, \ldots, J$. The Helmholtz operator can be easily approximated on this new nonuniform grid with the second order of accuracy using the same five-node stencil as shown in Figure 3.2. This involves little change compared to the approximation (3.5), which works for uniform grids, and we refer the reader to our paper [14] for detail. The discrete ABCs (3.10) or (3.11) do not change, except that $\Delta\rho_J$ needs to be substituted instead of $\Delta\rho$.

As we are building our control sources outside of the protected region $\Omega = \{(\rho, \theta) \,|\, \rho < r = 1\}$, i.e., on $\Omega_1 = \mathbb{R}^2 \backslash \Omega$, we do not need to be concerned with the structure of the grid inside Ω. For our constructions, we will only need to use one grid circle inside Ω. This will be the innermost circle $j = -1$. The second to innermost circle $j = 0$ already represents the interface $\Gamma = \partial\Omega = \{(\rho, \theta) \,|\, \rho = r = 1\}$.

Adopting the definition (3.2) of the grid subsets introduced in section 3.1, we obtain

(4.14)
$$\mathbb{M}^+ = \{(\rho_j, \theta_s)\,|\,j = -1\}, \quad \mathbb{M}^- = \{(\rho_j, \theta_s)\,|\,0 \le j \le J-1\},$$
$$\mathbb{N}^+ = \{(\rho_j, \theta_s)\,|\,j = -1, 0\}, \quad \mathbb{N}^- = \{(\rho_j, \theta_s)\,|\,-1 \le j \le J\},$$
$$\gamma = \{(\rho_j, \theta_s)\,|\,j = -1, 0\}, \quad \gamma^+ = \{(\rho_j, \theta_s)\,|\,j = -1\}, \quad \gamma^- = \{(\rho_j, \theta_s)\,|\,j = 0\}.$$

For all definitions in (4.14), we assume $s = 0, \ldots, L-1$.

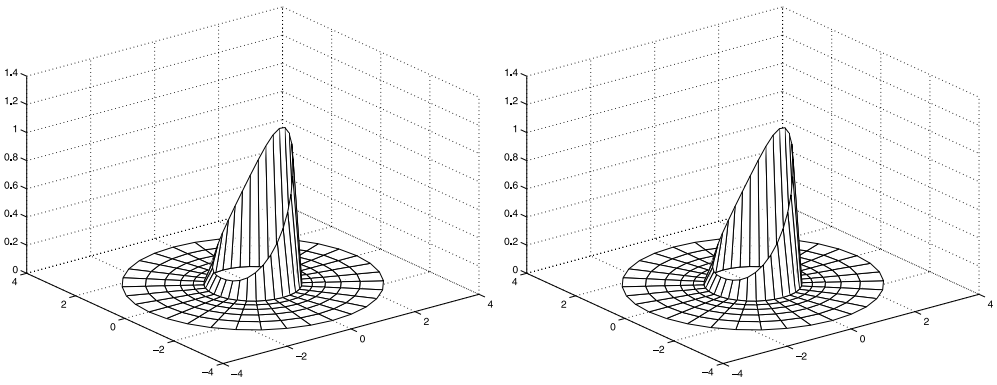
In the computational experiments, we have used grids with four times the number of cells in the circumferential direction compared to the radial direction. Specific grid dimensions were: $L = 32$ and $L = 48$, and accordingly, $J = 7$ and $J = 11$. (Note, as $j = -1, 0, \ldots, J$, the number of cells in the radial direction is $J+1$.) The wavenumber k in the Helmholtz equation (1.1) was chosen as $k = 0.5$. The excitation, i.e., the acoustic field $u^{(h)}$ that drives the control system, was taken in the analytic form of a shifted fundamental solution (see formula (2.3a)), as if it were generated by the point source $\delta(\boldsymbol{x} - \boldsymbol{x}_1)$, where $\boldsymbol{x}_1 = (\rho\cos\theta, \rho\sin\theta) = (5, 0)$. We reemphasize that our approach does not require an explicit knowledge of the exterior sources of noise. We only need this function $u^{(h)}$ as a sample field to be used as given data in formula (3.4).

We have also considered another case: $L = 48$, $J = 9$; for this case, we have selected $k = 0.9$. The excitation was produced by two point sources, $\delta(\boldsymbol{x} - \boldsymbol{x}_1) + \delta(\boldsymbol{x} - \boldsymbol{x}_2)$, where $\boldsymbol{x}_1 = (5, 0)$ and $\boldsymbol{x}_2 = (1, 2)$. Note, as for the wavenumber we have $k = \omega/c$, where ω is the temporal frequency and c is the speed of sound (see section 4.1). We also obtain the following relation between the wavelength λ and the wavenumber: $\lambda = 2\pi/k$. This means that in both cases, $k = 0.5$ and $k = 0.9$, we consider long waves relative to the diameter of the protected region Ω, which has been found advantageous from the standpoint of convergence of the numerical optimization algorithm.

The matrices and vectors involved in the formulation of the optimization problem (4.9) were constructed in accordance with the chosen geometric setup. Namely, the dimension of $\boldsymbol{L}^{(h)}$ (see (4.7)) was $M \times N \equiv (L \cdot J) \times (L \cdot (J+2))$; the dimension of $\boldsymbol{A}$, which corresponds to the variables on γ, was $M \times 2 \cdot L \equiv (L \cdot J) \times 2 \cdot L$; the dimension of $\boldsymbol{B}$ was $M \times (N - 4L) \equiv (L \cdot J) \times (L \cdot (J-2))$; and the dimension of either $\boldsymbol{C}$ or $\boldsymbol{D}$ was $M \times L \equiv (L \cdot J) \times L$.

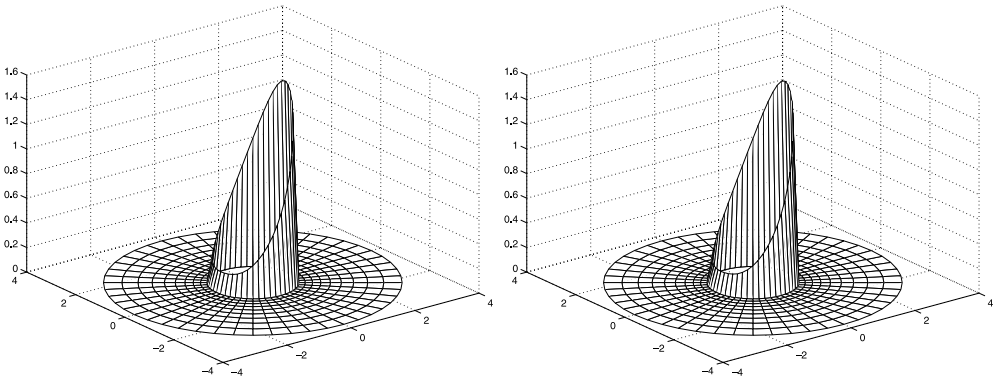
We have tried several numerical approaches for solving the corresponding minimization problems (4.11), starting with the algorithms available as a part of the standard optimization toolbox in MATLAB. However, our best numerical results were obtained with the software package SeDuMi by J. F. Sturm.[5] This is a numerical algorithm for optimization over cones [17]; it employs the ideas of interior point methods and the self-dual embedding technique of [25]; see also [13]. The algorithm allows for complex-valued entries, which is very important in our framework, and also for quasi-convex quadratic and positive semidefinite constraints. Of course, all the cases that we have been able to compute using SeDuMi (see above) can still be treated only as simple model examples on the scale of potential applications for noise control (see section 4.1). However, the optimal solutions that we have obtained all demonstrate a very coherent behavior that we discuss below. On Figures 4.1(a), 4.2(a), and 4.3(a) we plot magnitudes of the L_1 optimal solutions computed with SeDuMi [17]. Let us also note that SeDuMi is, in fact, a rather general procedure, and one may expect

[5] http://fewcal.kub.nl/sturm/software/sedumi.html



(a) L_1 optimal solution (b) Surface monopoles $g^{(\mathrm{h,surf})}_{\mathrm{monopole}}$

FIG. 4.1. *Control sources for* $L = 32$, $J = 7$, $k = 0.5$, *excitation* $\delta(\boldsymbol{x} - \boldsymbol{x}_1)$, $\boldsymbol{x}_1 = (5, 0)$.

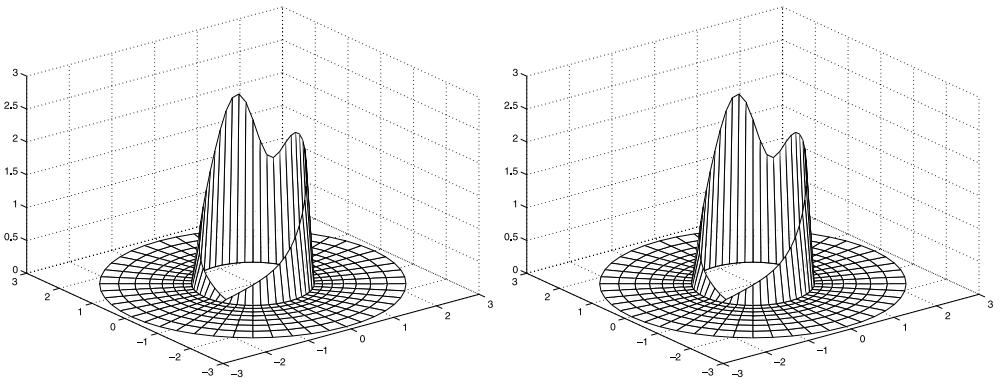


(a) L_1 optimal solution (b) Surface monopoles $g^{(\mathrm{h,surf})}_{\mathrm{monopole}}$

FIG. 4.2. *Control sources for* $L = 48$, $J = 11$, $k = 0.5$, *excitation* $\delta(\boldsymbol{x} - \boldsymbol{x}_1)$, $\boldsymbol{x}_1 = (5, 0)$.

better numerical performance from more focused algorithms, such as the one proposed by Andersen et al. in [1]. In the future, we may try the algorithm of [1] for solving the foregoing L_1 minimization problem.

Apparently, the most obvious observation that one can make by looking at Figures 4.1(a), 4.2(a), and 4.3(a) is that in all cases the optimal solution (i.e., the L_1 minimum) is concentrated on a single circumferential layer of grid nodes. This is the second to innermost circle of the grid $\mathbb{N}^-$ (see (4.14)), i.e., the grid line $j = 0$. It corresponds to the outer portion of the grid boundary γ^- (see (4.14)), and, in the continuous case, to the interface Γ itself, $\Gamma = \partial\Omega = \{(\rho, \theta) \,|\, \rho = r = 1\}$. In other words, the L_1-optimal solutions for control sources that we have computed can all be interpreted as layers of monopole sources on the perimeter of the protected region Ω. This clearly calls for comparing these optimal solutions with the densities of discrete



(a) L_1 optimal solution (b) Surface monopoles $g^{(\mathrm{h,surf})}_{\mathrm{monopole}}$

FIG. 4.3. *Control sources for* $L = 48$, $J = 9$, $k = 0.9$, *excitation* $\delta(\boldsymbol{x} - \boldsymbol{x}_1) + \delta(\boldsymbol{x} - \boldsymbol{x}_2)$, $\boldsymbol{x}_1 = (5, 0)$, $x_2 = (1, 2)$.

single-layer potentials introduced in section 3.4.

To conform the general definitions of section 3.4 to the specific geometric setup analyzed here, we need to construct on $\mathbb{N}^- \backslash \gamma^+$ the solution $w^{(h)}$ of the discrete exterior Dirichlet problem (3.12b) for the case when γ^- is the grid circle $j = 0$. (In this case, $\mathbb{M}^-_{\mathrm{int}}$ corresponds to $j > 0$.) Then, the operator $\boldsymbol{L}^{(h)}$ of (3.5) needs to be applied to the overall resulting function $w^{(h)}$, including its definition (3.12a) on the inner part γ^+ of the grid boundary: $w^{(h)}_n|_{n\in\gamma^+} = u^{(h)}_n|_{n\in\gamma^+}$. Since we know ahead of time that the resulting controls will differ from zero only on γ^- (which is an equivalent of $\mathbb{M}^- \backslash \mathbb{M}^-_{\mathrm{int}}$ in this case), we need not consider $w^{(h)}$ anywhere beyond $j = 1$. Consequently, for the purpose of constructing surface monopoles, we may simply set $J = 1$ and consider $w^{(h)}$ on three grid circles only: $j = -1$, $j = 0 \equiv J - 1$, and $j = 1 \equiv J$. In so doing, we obviously have to specify the ABCs (3.11) right on the interface; in other words, the input for the ABCs will be on γ^-, i.e., at $j = 0$, and the output will be on the outermost circle $j = J = 1$. Clearly, specifying the ABCs on γ^- allows us to reconstruct $w^{(h)}_{s,j}$ for $j = J$ directly by formula (3.11), i.e., without actually solving the aforementioned exterior Dirichlet problem. And once we know $w^{(h)}$ for $j = -1$, 0, and 1, we can easily obtain the discrete surface monopole controls on γ^-, i.e., for $j = 0$. As in section 3.4, we will denote these control sources $g^{(\mathrm{h,surf})}_{\mathrm{monopole}}$.

In Figures 4.1(b), 4.2(b), and 4.3(b), we plot magnitudes of the discrete surface controls $g^{(\mathrm{h,surf})}_{\mathrm{monopole}}$ for the exact same cases, for which we have explicitly computed the L_1 minimal solutions using SeDuMi. Visually comparing Figures 4.1(a), 4.2(a), and 4.3(a) with respective Figures 4.1(b), 4.2(b), and 4.3(b), we conclude that there is virtually no difference between them. In other words, the L_1-optimal solutions coincide with the surface monopole control sources $g^{(\mathrm{h,surf})}_{\mathrm{monopole}}$. To further corroborate this conclusion, we evaluate the L_1 norm of the difference on the grid $\mathbb{M}^-$ between each L_1-optimal solution and the corresponding surface monopole layer $g^{(\mathrm{h,surf})}_{\mathrm{monopole}}$, assuming that $g^{(\mathrm{h,surf})}_{\mathrm{monopole}} = 0$ everywhere except on γ^-. The results are presented in Table 4.1.

The data in Table 4.1, which take into account both magnitude and phase, do

TABLE 4.1
Comparison of the computed L_1-optimal solutions with surface monopoles.

Case	$\min_{w^{(h)}} \|g^{(h)}\|_{1,\mathbb{M}^-}$	$\|g^{(\mathrm{h,surf})}_{\mathrm{monopole}}\|_{1,\mathbb{M}^-}$	$\|g^{(h)}_{\mathrm{min}} - g^{(\mathrm{h,surf})}_{\mathrm{monopole}}\|_{1,\mathbb{M}^-}$	Relative diff.
Figure 4.1	0.5764	0.5761	0.0067	0.0117
Figure 4.2	0.5769	0.5761	0.0036	0.0063
Figure 4.3	0.9760	0.9750	0.0083	0.0085

corroborate that the respective solutions are close to one another. Moreover, by comparing the second and third rows in Table 4.1, we can apparently observe the phenomenon of grid convergence. Indeed, the case of Figure 4.2 is computed on a grid which is 1.5 times finer in each direction than the grid of Figure 4.1. For a second-order scheme, we can consequently expect a drop in the error by a factor of ~ 2.25, which we indeed see in Table 4.1 for both the absolute and relative difference between the L_1 minimum $g^{(h)}_{\mathrm{min}}$ and surface monopoles $g^{(\mathrm{h,surf})}_{\mathrm{monopole}}$. Even though the solutions that we are computing are obviously not smooth, and thus the grid convergence may be difficult to justify analytically, the foregoing experimental observation certainly makes our point about the coincidence of $g^{(h)}_{\mathrm{min}}$ and $g^{(\mathrm{h,surf})}_{\mathrm{monopole}}$ even more convincing.

Summarizing the foregoing numerical results, we can see that in all the cases analyzed the minimum for the L_1 norms of the control sources $g^{(h)}$ on $\mathbb{M}^-$ (see (4.5)) is actually given by the L_1 norm of $g^{(\mathrm{h,surf})}_{\mathrm{monopole}}$:

$$\min_{w^{(h)}} \|g^{(h)}\|_{1,\mathbb{M}^-} = \|g^{(\mathrm{h,surf})}_{\mathrm{monopole}}\|_{1,\mathbb{M}^-}.$$

The right-hand side of the previous equality can be recast into a more natural form by noticing that surface controls $g^{(\mathrm{h,surf})}_{\mathrm{monopole}}$ are defined only on γ^-. Then we can replace the L_1 norm over the two-dimensional grid domain $\mathbb{M}^-$ by the L_1 norm over the "one-dimensional" grid subset γ^-. This will bring about a factor of $\Delta\rho_1$ because obviously the cell areas V_m (see (4.5)) that correspond to nodes $j = 0$ are all equal and proportional to $\Delta\rho_1$. As such, we obtain

$$\min_{w^{(h)}} \|g^{(h)}\|_{1,\mathbb{M}^-} = \|g^{(\mathrm{h,surf})}_{\mathrm{monopole}}\|_{1,\gamma^-} \Delta\rho_1. \tag{4.15}$$

Equality (4.15) basically conjectures *global minimality of the surface monopole solution for active controls in the sense of* L_1.

As of yet, of course, we can only claim that equality (4.15) holds because it has been corroborated by a particular collection of numerical experiments that we have conducted. However, motivated by the consistency of our experimental observations that all suggest (4.15) (see Figures 4.1, 4.2, and 4.3 and Table 4.1), we have been able to prove a general result on the global L_1-optimality of the surface monopole solution for controls in both continuous and discrete formulation in the one-dimensional case. As has already been mentioned, we interpret the combination of the foregoing numerical results and the forthcoming analytic one-dimensional proof as an indication that surface monopoles may provide a universal global optimum in the sense of L_1. This, in particular, means that no numerical optimization will be needed for constructing the L_1-optimal control sources; they can simply be obtained by solving the boundary-value problem (3.12b) for the generating function $w^{(h)}$.

4.3. One-dimensional proof of global L_1-optimality. It will be convenient to consider simultaneously both the continuous and discrete formulations of the one-dimensional noise control problem. Let's denote the independent variable $x \in \mathbb{R}$ and

introduce the one-dimensional Helmholtz equation for the field variable $u = u(x)$ (cf. (1.1)):

$$\boldsymbol{L}u \equiv \frac{d^2u}{dx^2} + k^2u = f(x). \tag{4.16}$$

Then we introduce a uniform grid $x_n = n \cdot h$, $n = 0, \pm 1, \pm 2, \dots$, of variable x, and approximate (4.16) with the second-order central-difference scheme

$$\boldsymbol{L}^{(h)}u^{(h)} = \frac{u^{(h)}_{n+1} - 2u^{(h)}_n + u^{(h)}_{n-1}}{h^2} + k^2u^{(h)}_n = f^{(h)}_n. \tag{4.17}$$

Note that here we are using the same subscript "n" for both the discrete unknown function $u^{(h)}$ and the discrete right-hand side $f^{(h)}$, because for the particular scheme (4.17) they are defined on the same grid. We will still need to distinguish, however, between the grids $\mathbb{M}$ and $\mathbb{N}$ when constructing the necessary grid subsets.

Let us assume that our protected region Ω corresponds to $x < 0$ and accordingly, the complementary region Ω_1 corresponds to $x \geq 0$. Then, the continuous control sources will be given by (cf. formula (2.1))

$$g(x) = -\frac{d^2w}{dx^2} - k^2w\bigg|_{x\geq 0}, \tag{4.18}$$

where the auxiliary function $w(x)$, $x \geq 0$, is supposed to satisfy the interface conditions (cf. formula (2.2))

$$w(0) = u(0), \qquad \frac{dw}{dx}\bigg|_{x=0} = \frac{du}{dx}\bigg|_{x=0} \tag{4.19}$$

and the appropriate ABC, i.e., the radiation boundary condition, as $x \longrightarrow +\infty$. The quantities $u(0)$ and $\frac{du}{dx}\big|_{x=0}$ in (4.19) are assumed to be given.

Next, applying the definitions of section 3.1 to a particular stencil given by (4.17), we will have $\mathbb{M}^+ = \{m \,|\, m \equiv n = -1, -2, \dots\}$, $\mathbb{M}^- = \{m \,|\, m \equiv n = 0, 1, 2, \dots\}$, $\mathbb{N}^+ = \{n \,|\, n = 0, -1, -2, \dots\}$, $\mathbb{N}^- = \{n \,|\, n = -1, 0, 1, 2, \dots\}$, and $\gamma = \{n \,|\, n = -1, 0\}$. Accordingly, the discrete one-dimensional control sources will be given by (cf. formula (3.3))

$$g^{(h)}_n = -\frac{w^{(h)}_{n+1} - 2w^{(h)}_n + w^{(h)}_{n-1}}{h^2} - k^2w^{(h)}_n\bigg|_{n\geq 0}, \tag{4.20}$$

where the auxiliary grid function $w^{(h)}_n$ is supposed to satisfy the interface conditions on $\gamma = \{n \,|\, n = -1, 0\}$ (cf. formula (3.4))

$$w^{(h)}_{-1} = u^{(h)}_{-1}, \qquad w^{(h)}_0 = u^{(h)}_0, \tag{4.21}$$

and the appropriate ABC at infinity, or in other words, for large n's. Again, the quantities $u^{(h)}_{-1}$ and $u^{(h)}_0$ in (4.21) are considered as given.

To obtain the continuous ABC, we assume that the auxiliary function $w(x)$ satisfies the homogeneous version of (4.16): $\boldsymbol{L}w = 0$ for $x \geq X > 0$. This equation has two linearly independent solutions: e^{-ikx} is a right-traveling wave, and e^{ikx} is a left-traveling wave. In the one-dimensional framework, we obviously need to treat

the right-traveling wave as outgoing for the artificial outer boundary $x = X$. Therefore, employing the same mode selection principle as in section 2.2, we arrive at the following ABC (cf. formulae (2.13) and (2.17)):

$$\left.\frac{dw}{dx}\right|_{x=X} = -ikw(X), \tag{4.22}$$

which guarantees that only one of the two aforementioned linearly independent modes, namely e^{-ikx}, will remain in the composition of $w(x)$ for $x \geq X$. Note that boundary condition (4.22) can, in fact, be interpreted as the Sommerfeld radiation condition. In the one-dimensional case, it can be specified at a finite location, in contradistinction to the multidimensional case when these boundary conditions can only be specified at infinity; see formulae (1.2a), (1.2b). Altogether, the continuous auxiliary function $w = w(x)$ that defines the control sources $g(x)$ by formula (4.18) is specified on the interval $[0, X]$ and satisfies boundary conditions (4.19) and (4.22).

To obtain the discrete one-dimensional ABC, we will not approximate (4.22) with finite differences, as we did in section 3.3 for the polar case, when it was basically the only option. We will rather use a genuine finite-difference approach, which has shown efficient in many cases (see the review [19]), and which was studied in our recent paper [4] for a more complex formulation that involves a high-order approximation to the Helmholtz equation. Let's assume that the auxiliary grid function $w^{(h)}$ satisfies the homogeneous version of (4.17): $\boldsymbol{L}^{(h)} w^{(h)}\big|_n = 0$ for $n \geq N > 0$ (one may think that $X = (N-1) \cdot h$). This homogeneous finite-difference equation has two linearly independent solutions: q^n and q^{-n}, where q and q^{-1} are roots of the corresponding algebraic characteristic equation

$$q^2 - (2 - k^2h^2)q + 1 = 0. \tag{4.23}$$

These roots are given by the formulae:

$$q = 1 - \frac{1}{2}k^2h^2 - ikh\sqrt{1 - \frac{1}{4}k^2h^2}, \qquad q^{-1} = 1 - \frac{1}{2}k^2h^2 + ikh\sqrt{1 - \frac{1}{4}k^2h^2}. \tag{4.24}$$

It is easy to see from (4.24) that for small h the discrete wave q^n approximates the continuous right-traveling wave e^{-ikx}, and the discrete wave q^{-n} approximates the continuous left-traveling wave e^{ikx}. Therefore, the solution q^n shall be interpreted as a discrete outgoing wave, and q^{-n} shall be interpreted as a discrete incoming wave, for the external artificial boundary $n = N$. To guarantee the radiation of waves, we need to select q^n and prohibit q^{-n}, or in other words, require that $w_n^{(h)} = c \cdot q^n$ for $n \geq N$, where $c = \text{const}$. Accordingly, we arrive at the following discrete ABC:

$$w_N^{(h)} = q \cdot w_{N-1}^{(h)}, \tag{4.25}$$

which guarantees that only one of the two aforementioned linearly independent solutions, namely q^n, will remain in the composition of $w_n^{(h)}$ for $n \geq N$. Altogether, the auxiliary grid function $w^{(h)} = w_n^{(h)}$ that defines the control sources $g^{(h)}$ by formula (4.20) is specified on the grid subset $\{n \,|\, n = -1, 0, 1, \ldots, N\}$ and satisfies boundary conditions (4.21) and (4.25).

From now on, we will be considering only the situation with no interior sources. In other words, the only field present in the model *before control* will be the incoming field with respect to the protected region $\Omega = \{x \in \mathbb{R} \,|\, x < 0\}$. In the continuous case it can be expressed as $u(x) \equiv u^-(x) = Ae^{ikx}$, and in the discrete case as $u_n^{(h)} \equiv u_n^{(h)-} = Aq^{-n}$, where $A = \text{const}$. This restriction, in fact, presents no loss of generality. Indeed, if we had both components, $u(x) = u^-(x) + u^+(x) = Ae^{ikx} + Be^{-ikx}$, and had chosen $w(x)$ according to (4.19) and (4.22), then we could have replaced this

$w(x)$ by $\tilde{w}(x) = w(x) - Be^{-ikx} \equiv w(x) - u^+(x)$. The function $\tilde{w}(x)$ would satisfy the new interface conditions $\tilde{w}(0) = u^-(0)$, $\frac{d\tilde{w}}{dx}\big|_{x=0} = \frac{du^-}{dx}\big|_{x=0}$ instead of (4.19), and the same original ABC (4.22). Most important, the control sources generated by this new auxiliary function according to (4.18) will be the exact same control sources as those generated by $w(x)$: $\boldsymbol{L}w = \boldsymbol{L}[\tilde{w} + Be^{-ikx}] = \boldsymbol{L}\tilde{w}$, $x \geq 0$. Similarly in the discrete case, if we had $u_n^{(h)} = u_n^{(h)-} + u_n^{(h)+} = Aq^{-n} + Bq^n$, then the auxiliary functions $w_n^{(h)}$ and $\tilde{w}_n^{(h)} = w_n^{(h)} - Bq^n \equiv w_n^{(h)} - u_n^{(h)+}$ would generate the exact same discrete control sources according to (4.20). Of course, the foregoing argument is in complete agreement with the general discussion of section 2 on insensitivity of the control sources to the interior sound.

In the continuous one-dimensional case, the interface between the protected region $\Omega = \{x \in \mathbb{R} \,|\, x < 0\}$ and its complement $\Omega_1 = \{x \in \mathbb{R} \,|\, x \geq 0\}$ is obviously one point, $x = 0$. To construct the corresponding "surface" monopole controls, we consider a special form of the auxiliary function $w(x)$. Namely, if the original field to be controlled is the left-traveling wave that propagates into Ω, $u(x) = u^-(x) = Ae^{ikx}$, $x < 0$, then we take $w(x)$ in the form of the right-traveling wave: $w(x) = Ae^{-ikx}$, $x \geq 0$. Obviously, this function $w(x)$ solves the homogeneous equation on Ω_1, $\boldsymbol{L}w = 0$, $x \geq 0$, and satisfies the ABC (4.18). It also satisfies the Dirichlet boundary condition at the interface $x = 0$: $w(0) = u(0)$. Altogether, we see that $w(x)$ selected this way solves the one-dimensional counterpart of the exterior Dirichlet problem (2.23) that we used in section 2.3 to obtain the control sources in the form of surface monopoles. According to the analysis of section 2.3, surface monopoles are obtained by applying the operator $-\boldsymbol{L}$ to the function v of (2.21), which has discontinuous first derivative across the interface. In the specific one-dimensional case that we are studying here, this function is given by

$$v(x) = \begin{cases} Ae^{ikx} & \text{for } x < 0, \\ Ae^{-ikx} & \text{for } x \geq 0. \end{cases} \tag{4.26}$$

Applying the operator $-\boldsymbol{L}$ (see (4.16)) to the function $v(x)$ of (4.26) in the sense of distributions (see [24]), we obtain the following "surface" (in fact, point) monopole control source (cf. formula (2.30)):

$$g_{\text{monopole}}^{(\text{surf})} = 2Aik\delta(x). \tag{4.27}$$

To obtain the discrete "surface" monopoles in the one-dimensional case, we need to consider $u_n^{(h)} = Aq^{-n}$ for $n \leq 0$, and $w_n^{(h)} = Aq^n$ for $n \geq 0$; we also set $w_{-1}^{(h)} = u_{-1}^{(h)}$. The aforementioned $w_n^{(h)}$ solves the discrete homogeneous equation $\boldsymbol{L}^{(h)}w^{(h)} = 0$ for $n > 0$, satisfies the discrete ABC (4.25), and the interface conditions $w_\gamma^{(h)} = u_\gamma^{(h)}$, or equivalently (4.21). In other words, the selected $w^{(h)}$ solves the one-dimensional version of the exterior Dirichlet-type problem (3.12b), (3.12a) that we used in section 3.4 to derive surface monopole controls in the discrete framework. Applying the operator $-\boldsymbol{L}^{(h)}$ (see (4.17)) to the foregoing function $w^{(h)}$, we obtain the discrete surface control source

$$g_{\text{monopole}}^{(h,\,\text{surf})} = \begin{cases} -A\left(2\frac{q-1}{h^2} + k^2\right) & \text{for} \quad n = 0, \\ 0 & \text{for} \quad n > 0. \end{cases} \tag{4.28}$$

We are now prepared to formulate our central result on the global L_1-optimality of surface monopoles in the one-dimensional discrete framework. For any function $w^{(h)} = w_n^{(h)}$ that satisfies the ABC (4.25) and the interface conditions (4.21), where $u_n^{(h)} = Aq^{-n}$, $n \leq 0$, the L_1 norm of the corresponding control sources (4.20) will

always be greater than or equal to the magnitude of the surface monopole (4.28) times the grid size h: $\|g^{(h)}\|_1 \geq |g^{(h,\,\mathrm{surf})}_{\mathrm{monopole}}|h$. In the case $w^{(h)}_n = Aq^n$ the equality is achieved, and thus $\min_{w^{(h)}} \|g^{(h)}\|_1 = |g^{(h,\,\mathrm{surf})}_{\mathrm{monopole}}|h$. In other words, the following theorem holds.

THEOREM 4.1. *Let a complex-valued function $w^{(h)} = w^{(h)}_n$ be defined on the grid $n = -1, 0, 1, \ldots, N$, where $N > 0$ can be arbitrary. Let $w^{(h)}_0 = A$, where $A \in \mathbb{C}$ is a given constant, and $w^{(h)}_{-1} = qw^{(h)}_0$ and $w^{(h)}_N = qw^{(h)}_{N-1}$, where q is defined by formula (4.24). Then,*

$$\min_{w^{(h)}_n} \sum_{n=0}^{N-1} \left| \frac{w^{(h)}_{n+1} - 2w^{(h)}_n + w^{(h)}_{n-1}}{h^2} + k^2 w^{(h)}_n \right| = |A| \left| 2\frac{q-1}{h^2} + k^2 \right|. \tag{4.29}$$

Proof. Let us introduce new quantities p_0, $p_1, \ldots, p_{N-2}$ so that $w^{(h)}_1 = p_0 w^{(h)}_0$, $w^{(h)}_2 = p_1 w^{(h)}_1, \ldots, w^{(h)}_{N-1} = p_{N-2} w^{(h)}_{N-2}$. Then the sum on the left-hand side of (4.29) can be recast in the following form (taking into account that $w^{(h)}_{-1} = qw^{(h)}_0$ and $w^{(h)}_N = qw^{(h)}_{N-1}$):

$$\begin{aligned}
&\sum_{n=0}^{N-1} \left| \frac{w^{(h)}_{n+1} - 2w^{(h)}_n + w^{(h)}_{n-1}}{h^2} + k^2 w^{(h)}_n \right| \\
&\quad = |A| \left| \frac{p_0 + q - 2}{h^2} + k^2 \right| + |A| \left| \frac{p_0 p_1 - 2p_0 + 1}{h^2} + k^2 p_0 \right| \\
&\quad + |A||p_0| \left| \frac{p_1 p_2 - 2p_1 + 1}{h^2} + k^2 p_1 \right| + |A||p_0||p_1| \left| \frac{p_2 p_3 - 2p_2 + 1}{h^2} + k^2 p_2 \right| + \cdots \\
&\quad + |A||p_0||p_1| \cdots |p_{N-3}| \left| \frac{p_{N-2} q - 2p_{N-2} + 1}{h^2} + k^2 p_{N-2} \right|.
\end{aligned}$$

Next, we introduce new notations: $p_0 = q + z_0$, $p_1 = q + z_1, \ldots, p_{N-2} = q + z_{N-2}$, where q is defined by (4.24) and all the quantities are generally assumed to be complex. Using these new notations, we can rewrite the generic term on the right-hand side of the previous equality as follows:

$$\begin{aligned}
&|A||p_0||p_1| \cdots |p_n| \left| \frac{p_{n+1} p_{n+2} - 2p_{n+1} + 1}{h^2} + k^2 p_{n+1} \right| \\
&\quad = |A||q + z_0||q + z_1| \cdots |q + z_n| \left| \frac{(q + z_{n+1})(q + z_{n+2}) - 2(q + z_{n+1}) + 1}{h^2} \right. \\
&\qquad\qquad \left. + k^2 (q + z_{n+1}) \right| \\
&\quad = |A||q + z_0||q + z_1| \cdots |q \\
&\qquad + z_n| \Bigg| \underbrace{\frac{q^2 - 2q + 1}{h^2} + k^2 q}_{0} + \frac{z_{n+1} z_{n+2} + q(z_{n+1} + z_{n+2}) - 2z_{n+1}}{h^2} + k^2 z_{n+1} \Bigg| \\
&\quad = |A||q + z_0||q + z_1| \cdots |q + z_n| \left| \frac{z_{n+2}(q + z_{n+1})}{h^2} + z_{n+1} \left(\frac{q-2}{h^2} + k^2 \right) \right| \\
&\quad = |A||q + z_0||q + z_1| \cdots |q + z_n| \left| \frac{z_{n+2}(q + z_{n+1})}{h^2} + \frac{z_{n+1}}{h^2} \mu \right|.
\end{aligned}$$

In the last chain of equalities, expression $\frac{q^2-2q+1}{h^2} + k^2 q$ turns into zero by virtue of the characteristic equation (4.23), and $\mu = q - 2 + k^2h^2$. Using the definition of q from (4.24), we can obtain

$$|\mu|^2 = \left|-1 - \frac{1}{2}k^2h^2 - ikh\sqrt{1 - \frac{1}{4}k^2h^2} + k^2h^2\right| = \left|-1 + \frac{1}{2}k^2h^2 - ikh\sqrt{1 - \frac{1}{4}k^2h^2}\right|$$

$$= \left(-1 + \frac{1}{2}k^2h^2\right)^2 + k^2h^2\left(1 - \frac{1}{4}k^2h^2\right) = 1 - k^2h^2 + \frac{1}{4}k^4h^4 + k^2h^2 - \frac{1}{4}k^4h^4 = 1.$$

Finally, collecting all terms, we can now have

$$\sum_{n=0}^{N-1}\left|\frac{w_{n+1}^{(h)} - 2w_n^{(h)} + w_{n-1}^{(h)}}{h^2} + k^2 w_n^{(h)}\right|$$

$$= |A|\left|\frac{z_0}{h^2} + 2\frac{q-1}{h^2} + k^2\right| + |A|\left|\frac{z_1(q+z_0)}{h^2} + \frac{z_0}{h^2}\mu\right|$$

$$+ |A||q+z_0|\left|\frac{z_2(q+z_1)}{h^2} + \frac{z_1}{h^2}\mu\right| + |A||q+z_0||q+z_1|\left|\frac{z_3(q+z_2)}{h^2} + \frac{z_2}{h^2}\mu\right| + \cdots$$

$$+ |A||q+z_0||q+z_1|\ldots|q+z_{N-3}|\left|\frac{z_{N-2}}{h^2}\mu\right|$$

$$\geq |A|\left|2\frac{q-1}{h^2} + k^2\right| - |A|\left|\frac{z_0}{h^2}\right| + |A|\left|\frac{z_0}{h^2}\right||\mu| - |A||q+z_0|\frac{|z_1|}{h^2} + |A||q+z_0|\frac{|z_1|}{h^2}|\mu|$$

$$- |A||q+z_0||q+z_1|\frac{|z_2|}{h^2} + |A||q+z_0||q+z_1|\frac{|z_2|}{h^2}|\mu| - \cdots$$

$$+ |A||q+z_0||q+z_1|\ldots|q+z_{N-3}|\frac{|z_{N-2}|}{h^2}|\mu|$$

$$= |A|\left|2\frac{q-1}{h^2} + k^2\right|.$$

In other words, we have obtained the inequality

$$\sum_{n=0}^{N-1}\left|\frac{w_{n+1}^{(h)} - 2w_n^{(h)} + w_{n-1}^{(h)}}{h^2} + k^2 w_n^{(h)}\right| \geq |A|\left|2\frac{q-1}{h^2} + k^2\right|. \tag{4.30}$$

To establish the result of the theorem, i.e., formula (4.29), it remains to show only that there will be a particular $w^{(h)} = w_n^{(h)}$ for which inequality (4.30) transforms into the equality. Clearly, the equality in formula (4.30) is achieved for $w_n^{(h)} = Aq^n$, $n \geq 0$, because in this case $z_0 = z_1 = \cdots = z_{N-2} = 0$. This completes the proof. ∎

Let us also recall that if we multiply the sum on the left-hand side of either formula (4.29) or formula (4.30) by the grid size h, we obtain the discrete L_1 norm of the control sources $g^{(h)}$. Therefore, inequality (4.30) transforms into

$$\left\|g^{(h)}\right\|_1 \geq \left|g_{\text{monopole}}^{(h,\,\text{surf})}\right| h, \tag{4.31}$$

and consequently, we have, in effect, demonstrated the global L_1 minimality of the surface control sources:

$$\min_{w^{(h)}}\left\|g^{(h)}\right\|_1 = \left|g_{\text{monopole}}^{(h,\,\text{surf})}\right| h. \tag{4.32}$$

Equality (4.32) is a one-dimensional counterpart of (4.15), but unlike the experimentally established formula (4.15), equality (4.32) has been proven rigorously.

The foregoing proof of Theorem 4.1 also reveals the mechanism of discrepancy between the optimal control $g^{(h,\,\mathrm{surf})}_{\mathrm{monopole}}$ and all other suboptimal controls. Namely, every time the auxiliary function $w^{(h)}_n$ "departs" from the pure right-traveling wave Aq^n, which is equivalent to having $z_n \neq 0$, we may pick up additional value of $\|g^{(h)}\|_1$ in case the actual estimate based on the triangle inequality that we use,

$$|A||q+z_0||q+z_1|\cdots|q+z_n|\left|\frac{z_{n+2}(q+z_{n+1})}{h^2}+\frac{z_{n+1}}{h^2}\mu\right|$$

$$\geq |A||q+z_0||q+z_1|\cdots|q+z_n|\frac{|z_{n+1}|}{h^2}|\mu|$$

$$-|A||q+z_0||q+z_1|\cdots|q+z_n||q+z_{n+1}|\frac{|z_{n+2}|}{h^2},$$

happens to be "strictly greater" rather than "greater or equal" for this given term. It is also interesting to look into the role of the ABC (4.25). This boundary condition "swallows" the last term in the sum so that all the previous terms can cancel one another in pairs, and only the first term $|g^{(h,\,\mathrm{surf})}_{\mathrm{monopole}}|$ will remain.

Next, we will analyze the continuous case. Assume that $w(x)$ is a regular smooth function, which is defined on the interval $[0, X]$ and satisfies boundary conditions (4.19) and (4.22). Let us also assume that the grid function $w^{(h)} = w^{(h)}_n$, $n = 0, 1, \ldots, N-1$, is the trace of $w(x)$ on the aforementioned uniform grid with size h: $w^{(h)}_n = w(x_n) \equiv w(n \cdot h)$. Note that we need to require sufficient smoothness of $w(x)$ in order to guarantee the consistency of the finite-difference scheme: $\boldsymbol{L}^{(h)}w^{(h)} = \boldsymbol{L}w + O(h^2)$ (see (4.16) and (4.17)). Let us additionally define $w^{(h)}_{-1} = qw^{(h)}_0$ and $w^{(h)}_N = qw^{(h)}_{N-1}$, in accordance with the boundary conditions (4.21) and (4.25), respectively, like in the formulation of Theorem 4.1. Then for small h we can disregard the quadratic terms in the definition of q (see (4.24)) and have for the right endpoint

$$\frac{w^{(h)}_N - w^{(h)}_{N-1}}{h} = -ikw^{(h)}_{N-1},$$

which is obviously an approximation of the continuous ABC (4.22) with the accuracy $O(h)$. Similarly, for the left endpoint we obtain

$$\frac{w^{(h)}_0 - w^{(h)}_{-1}}{h} = ikw^{(h)}_0,$$

which is an $O(h)$ accurate approximation of the second boundary condition (4.19) under the assumption that the field to be controlled is $u(x) = u^-(x) = Ae^{ikx}$, $x < 0$, and consequently, $\frac{du}{dx}\big|_{x=0} = iku(0)$. Altogether, we have constructed a grid function $w^{(h)} = w^{(h)}_n$, $n = -1, 0, \ldots, N$, that satisfies the conditions of Theorem 4.1 and also approximates on the grid all the continuous requirements of the function $w = w(x)$.

Let us now again multiply both sides of inequality (4.30) by the positive quantity h and consider independently the limit on its right-hand side and the limit on its left-hand side as $h \longrightarrow +0$. First, we obtain

$$|A|\left|2\frac{q-1}{h^2}+k^2\right|h = |A|\left|-k^2-\frac{2ik}{h}\sqrt{1-\frac{1}{4}k^2h^2}+k^2\right|h \longrightarrow 2|A|k \quad \text{as} \quad h \longrightarrow +0.$$

Note that the limit is equal to the magnitude of the surface monopole in the continuous formulation (see (4.27)). Next, on the left-hand side we have

$$\sum_{n=0}^{N-1}\left|\frac{w_{n+1}^{(h)}-2w_n^{(h)}+w_{n-1}^{(h)}}{h^2}+k^2w_n^{(h)}\right|h=\sum_{n=0}^{N-1}\left|\frac{d^2w}{dx^2}(x_n)+k^2w(x_n)\right|h+O(h^2)$$

$$\longrightarrow\int_0^X\left|\frac{d^2w}{dx^2}(x)+k^2w(x)\right|dx\qquad\text{as}\quad h\longrightarrow+0.$$

Note that the limit is equal to the L_1 norm $\|g\|_1$ of the continuous control sources $g(x)$ defined by formula (4.18). As inequality (4.30) holds for any given value of h, we can claim that it will also hold in the limit $h\longrightarrow+0$. Therefore, we have arrived at the following result.

COROLLARY 4.2. *Let a complex-valued function $w=w(x)$ be defined on $[0,X]$. Let $w(0)=A$, where $A\in\mathbb{C}$ is a given constant, and $w'(0)=ikw(0)$ and $w'(X)=-ikw(X)$. Then,*

$$\int_0^X\left|\frac{d^2w}{dx^2}(x)+k^2w(x)\right|dx\geq 2|A|k. \tag{4.33}$$

It is easy to see that the requirements of $w(x)$ formulated in Corollary 4.2 are equivalent to the conditions that guarantee the appropriateness of $w(x)$ for constructing the control sources $g(x)$ using (4.18); see formulae (4.19) and (4.22). Therefore, the result of Corollary 4.2, i.e., inequality (4.33), can be recast as

$$\|g\|_1\geq 2|A|k. \tag{4.34}$$

On the right-hand side of inequality (4.34) we have the magnitude of the "surface" monopole $g_{\text{monopole}}^{(\text{surf})}$ defined by formula (4.27). Note that, unlike in the previously considered discrete case, when the minimal solution $g_{\text{monopole}}^{(h,\,\text{surf})}$ of (4.28) was an element of the same class of control sources $g^{(h)}$ defined by (4.20), here the minimum $g_{\text{monopole}}^{(\text{surf})}(x)$ defined by (4.27) belongs to a different class of functions, namely, singular (i.e., δ-type) distributions, as opposed to regular (i.e., $L_1^{(\text{loc})}$) distributions. In other words, $g_{\text{monopole}}^{(\text{surf})}(x)\notin L_1(\mathbb{R})$, and not even $L_1^{(\text{loc})}(\mathbb{R})$. As such, we cannot introduce the L_1 norm of $g_{\text{monopole}}^{(\text{surf})}(x)$. Therefore, inequality (4.34) formally has to stay the way it is. However, symbolically we can, of course, write

$$\int_{\mathbb{R}}\left|g_{\text{monopole}}^{(\text{surf})}(x)\right|dx=\int_{\mathbb{R}}|2Aik\delta(x)|\,dx=2|A|k, \tag{4.35}$$

which allows us to "informally" interpret inequality (4.34) as if $g_{\text{monopole}}^{(\text{surf})}(x)$ of (4.27) provided a lower bound in L_1 for all the control sources $g(x)$ defined by (4.18). Let us also note that "integration" with respect to x in (4.35), which "removes" the δ-function itself and leaves only its magnitude $2k|A|$, is a continuous analogue of multiplication by h on the right-hand side of formulae (4.31) or (4.32). Therefore, we conclude that the continuous inequality (4.34) is a direct counterpart of the discrete inequality (4.31).

Even though $g_{\text{monopole}}^{(\text{surf})}(x)\notin L_1^{(\text{loc})}(\mathbb{R})$, we will still show that there are regular control sources $g(x)\in L_1^{(\text{loc})}(\mathbb{R})$ that are arbitrarily close to $g_{\text{monopole}}^{(\text{surf})}(x)$ in the weak

sense. More precisely, we will construct a sequence of regular auxiliary functions $w_\epsilon(x)$ such that the corresponding $g_\epsilon(x)$ obtained according to (4.18) will converge to $g^{(\mathrm{surf})}_{\mathrm{monopole}}(x)$ of (4.27) in the sense of distributions. This will allow us to claim that although the minimal solution $g^{(\mathrm{surf})}_{\mathrm{monopole}}(x)$ is singular, it is, in fact, "on the borderline" of the class of regular solutions. In other words, it is a limiting point, in the sense of weak convergence, of the space of all $g(x)$ defined by formula (4.27). In addition, we will also show that the L_1 norms $\|g_\epsilon\|$ converge to the magnitude $2k|A|$ of the "surface" monopole $g^{(\mathrm{surf})}_{\mathrm{monopole}}(x)$ of (4.27). This will allow us to formulate in the continuous case the result similar to the minimality (4.32) but in the sense of "infimum" rather than "minimum."

Consider a regular function $w_\epsilon = w_\epsilon(x)$ that is defined on $[0, X]$ and satisfies boundary conditions (4.19) and (4.22) with $u(x) = Ae^{ikx}$ for $x < 0$. In addition, let us assume that not only for $x > X$, but also in between some (small) $\epsilon < X$ and X, the function $w_\epsilon(x)$ already coincides with a right-traveling wave: $w_\epsilon(x) = w_\epsilon(\epsilon)e^{-ik(x-\epsilon)}$, $\epsilon \le x \le X$. In other words, we require that $w_\epsilon(0) = A$, $w'_\epsilon(0) = ikA$, and $w'_\epsilon(\epsilon) = -ikw_\epsilon(\epsilon)$. For the purpose of obtaining the aforementioned convergent sequence, we will subsequently let $\epsilon \longrightarrow +0$. The control sources $g_\epsilon(x)$ are defined according to formula (4.18): $g_\epsilon(x) = -w''_\epsilon(x) - k^2 w_\epsilon(x)$, $0 \le x \le \epsilon$, and for $x > \epsilon$ we have $g_\epsilon(x) = 0$. If $\varphi = \varphi(x)$ is a test function on $\mathbb{R}$, i.e., a compactly supported infinitely smooth function (see [24]), then the corresponding functional, i.e., the distribution g_ϵ itself, can be represented as follows:

$$(g_\epsilon, \varphi) = \int_{\mathbb{R}} g_\epsilon(x)\varphi(x)dx = \int_0^\epsilon [-w''(x) - k^2 w(x)]\varphi(x)dx$$

$$= w(\epsilon)\varphi'(\epsilon) - w(0)\varphi(0) - [\varphi(\epsilon)w'(\epsilon) - \varphi(0)w'(0)] + \int_0^\epsilon [-w(x)\varphi''(x) + k^2 w(x)\varphi(x)]dx.$$

As $w(x) \in L_1^{(\mathrm{loc})}(\mathbb{R})$ and $\varphi(x)$ is a test function, the integral on the right-hand side of the previous equality vanishes as $\epsilon \longrightarrow +0$. Let us now additionally assume that the functions $w_\epsilon = w_\epsilon(x)$ are constructed so that $w_\epsilon(\epsilon) \longrightarrow w_\epsilon(0)$ when $\epsilon \longrightarrow +0$. In other words, we assume continuity at $x = 0$. Then, because of the boundary condition (4.22) at $x = \epsilon$, we have $w'_\epsilon(\epsilon) \longrightarrow -ikA$ as $\epsilon \longrightarrow +0$. Altogether, we obtain

$$(g_\epsilon, \varphi) \longrightarrow 2ikA\varphi(0) \quad \text{as} \quad \epsilon \longrightarrow +0. \tag{4.36}$$

The limit (4.36) implies that in the sense of distributions

$$g_\epsilon(x) \longrightarrow g^{(\mathrm{surf})}_{\mathrm{monopole}}(x) \equiv 2ikA\delta(x) \quad \text{as} \quad \epsilon \longrightarrow +0. \tag{4.37}$$

Let now specify a particular form of $w_\epsilon(x)$:

$$w_\epsilon(x) = \frac{-ikA}{\epsilon}x^2 + ikAx + A, \qquad x \in [0, \epsilon]. \tag{4.38}$$

For this function, we have $w_\epsilon(0) = w_\epsilon(\epsilon) = A$, $w'_\epsilon(0) = ikA$, and $w'_\epsilon(\epsilon) = -ikA$, and consequently, $w_\epsilon(x)$ of (4.38) meets all the previous conditions. For the L_1 norm of the corresponding control $g_\epsilon(x)$, we obtain

$$\|g_\epsilon\|_1 = \int_0^\epsilon |w''_\epsilon(x) + k^2 w_\epsilon(x)|dx = \int_0^\epsilon |A| \left[k^2 + \left(-\frac{k^3}{\epsilon}x^2 + k^3 x - \frac{2k}{\epsilon}\right)^2\right]^{1/2} dx.$$

As we always have $0 \le x \le \epsilon$, the dominant term in the last integral for small ϵ is $\frac{2k}{\epsilon}$, and therefore

$$\|g_\epsilon\|_1 \longrightarrow 2k|A| \quad \text{as} \quad \epsilon \longrightarrow +0. \tag{4.39}$$

Putting together the result of Corollary 4.2 with the limits (4.37) and (4.39), we arrive at the following result.

THEOREM 4.3. *Let a complex-valued function $w = w(x)$ be defined on $[0, X]$. Let $w(0) = A$, where $A \in \mathbb{C}$ is a given constant, and $w'(0) = ikw(0)$ and $w'(X) = -ikw(X)$. Then, in the class of regular control sources $g(x)$ defined by formula* (4.18) *for all such $w(x)$, one can identify a sequence $g_\epsilon(x)$ that would converge in the sense of distributions to the point monopole $g^{(\mathrm{surf})}_{\mathrm{monopole}}(x)$ defined by formula* (4.27):

$$g_\epsilon(x) \longrightarrow g^{(\mathrm{surf})}_{\mathrm{monopole}}(x) \equiv 2ikA\delta(x) \quad \textit{as} \quad \epsilon \longrightarrow +0.$$

Besides, the magnitude of the point monopole $g^{(\mathrm{surf})}_{\mathrm{monopole}}(x)$ of (4.27) *provides the greatest lower bound for L_1 norms of all the control sources $g(x)$ of* (4.18):

$$\inf_{w(x)} \|g\|_1 = 2k|A|. \tag{4.40}$$

Clearly, estimate (4.40) can be interpreted as global L_1 minimality of the "surface" control (4.27) among the continuous one-dimensional control sources (4.18). It is a continuous version of the previously established discrete result (4.32).

5. Discussion. For the problem of active control of sound, we have systematically described time-harmonic general solutions for the volume and surface control sources in the continuous and discrete formulation of the problem. These control sources guarantee the identical cancellation of unwanted noise on a predetermined region of interest. We have also proposed a criterion for optimization of the resulting control sources. This criterion chooses the overall absolute acoustic source strength as the cost function for minimization, and as such admits a clear physical interpretation. Mathematically, it translates into minimization of complex-valued functions in the sense of L_1, which is a very challenging problem from the standpoint of numerical implementation. We have still managed, though, to compute several two-dimensional numerical solutions using the algorithm SeDuMi. All these solutions demonstrate a coherent behavior—the minimum is achieved on the surface of the protected region. In other words, the minimum is delivered by the appropriate surface monopole controls. Therefore, the numerical evidence that we have received indicates that surface monopoles may provide a global L_1 minimum for the control sources in the general setting. We have been able to rigorously prove this result in the one-dimensional case for both continuous and discrete formulation of the problem. Even though we have not yet been able to prove a similar result for a general multi-dimensional framework, we still believe that it is true, because a combination of the two-dimensional numerical evidence and a one-dimensional accurate proof cannot, in our opinion, be a mere coincidence. Therefore, we put forward the minimization result in the form of a conjecture. Let us recall that according to (2.30) the surface monopole controls are given by

$$g^{(\mathrm{surf})}_{\mathrm{monopole}}(\boldsymbol{x}) = -\left[\frac{\partial w}{\partial \boldsymbol{n}} - \frac{\partial u}{\partial \boldsymbol{n}}\right]_\Gamma \delta(\Gamma) = -\left[\frac{\partial \tilde{w}}{\partial \boldsymbol{n}} - \frac{\partial u^-}{\partial \boldsymbol{n}}\right]_\Gamma \delta(\Gamma) \equiv \nu(\boldsymbol{x})|_{\boldsymbol{x}\in\Gamma} \cdot \delta(\Gamma), \tag{5.1}$$

where $\tilde{w}(x) = w(\boldsymbol{x}) - u^+(\boldsymbol{x})$, as before, and $w(\boldsymbol{x})$ is a solution to the exterior Dirichlet problem (2.23). Then, we can formulate the following.

CONJECTURE 5.1. *Let a complex-valued function $w = w(\boldsymbol{x})$ be defined on $\Omega_1 = \mathbb{R}^n \backslash \Omega$, and let it be sufficiently smooth so that the operator $\boldsymbol{L}$ of* (1.1) *can be applied to $w(\boldsymbol{x})$ on its entire domain in the classical sense, and the result $\boldsymbol{L}w$ can be locally absolutely integrable. Let, in addition, $w(\boldsymbol{x})$ satisfy the interface conditions* (2.2)*, where $u = u(\boldsymbol{x})$ is a given field to be controlled, and the appropriate Sommerfeld radiation boundary conditions at infinity,* (1.2a) *or* (1.2b)*. Then the greatest lower bound for the L_1 norms of all the control sources $g(\boldsymbol{x})$ obtained with such auxiliary functions $w(\boldsymbol{x})$ using formula* (2.1) *is given by the L_1 norm on Γ of the magnitude of surface monopoles* (5.1)*:*

$$\inf_{w(\boldsymbol{x})} \int_{\Omega_1} |g(\boldsymbol{x})| d\boldsymbol{x} = \int_{\Gamma} |\nu(\boldsymbol{x})| ds. \tag{5.2}$$

Alternatively, we can rewrite (5.2) as

$$\inf_{w(\boldsymbol{x})} \|g(\boldsymbol{x})\|_{1,\Omega_1} = \|\nu\|_{1,\Gamma}. \tag{5.3}$$

Equality (5.3) is a multidimensional generalization of (4.40). Let us also notice that equality (4.15), which was obtained on the basis of experimental observations in two space dimensions, can be considered a discrete two-dimensional prototype of (5.3).

In the formulation of Conjecture 5.1, we did not include the results on the convergence of a sequence of volumetric controls to the surface layer $\nu\delta(\Gamma)$ (see (5.1)) and on the convergence of the corresponding L_1 norms, as we did in Theorem 4.3 for the one-dimensional case. We believe, though, that these results can be easily formulated and justified in the multidimensional framework using an approach similar to the one that we have used in the one-dimensional case. The key missing part, however, that does not yet allow us to transform Conjecture 5.1 into a theorem, is proving that surface monopoles provide *a lower bound* for the volumetric controls in the sense of L_1, whereas showing that this is the greatest lower bound is more straightforward. The analysis of this problem will be a subject of our future research. In this connection we can mention only that, at least in the two-dimensional case, the geometry of the protected region Ω should not be a limitation when constructing a general proof. If one can prove the result for a constant-width linear strip with periodic boundary conditions on its sides, and with the interface Γ being a segment of the straight line normal to the sides of the strip, then for any other shape the same result can likely be obtained with the help of a conformal mapping.

Acknowledgments. It is our pleasure to acknowledge most useful discussions with Jan Hesthaven, Wu Li, Michael Overton, and Victor Ryaben'kii. We would also like to thank the reviewer of the paper, who drew our attention to several important points and thereby helped make the revised manuscript a substantial improvement over its original version.

REFERENCES

[1] K. D. ANDERSEN, E. CHRISTIANSEN, A. R. CONN, AND M. L. OVERTON, *An efficient primal-dual interior-point method for minimizing a sum of Euclidean norms*, SIAM J. Sci. Comput., 22 (2000), pp. 243–262.

[2] A. P. CALDERON, *Boundary-value problems for elliptic equations*, in Proceedings of the Soviet-American Conference on Partial Differential Equations, Novosibirsk, Moscow, 1963, Fizmatgiz, Moscow, pp. 303–304.

[3] S. J. Elliot, *Signal Processing for Active Control*, Academic Press, San Diego, 2001.
[4] G. Fibich and S. V. Tsynkov, *High-order two-way artificial boundary conditions for nonlinear wave propagation with backscattering*, J. Comput. Phys., 171 (2001), pp. 632–677.
[5] C. R. Fuller, S. J. Elliot, and P. A. Nelson, *Active Control of Vibration*, Academic Press, London, 1996.
[6] J. Lončarić, *Sensor/actuator placement via optimal distributed control of exterior Stokes flow*, in Computational Methods in Optimal Design and Control: Proceedings of the AFOSR Workshop on Optimal Design and Control, J. T. Borggaard, J. Burns, E. Cliff, and S. Schreck, eds., Progr. Systems Control Theory 24, Birkhäuser Boston, Cambridge, MA, 1998, pp. 303–322.
[7] J. Lončarić, V. S. Ryaben'kii, and S. V. Tsynkov, *Active shielding and control of noise*, SIAM J. Appl. Math., 62 (2001), pp. 563–596.
[8] J. Lončarić and S. V. Tsynkov, *Optimization of Power in the Problems of Active Control of Sound*, manuscript, 2003.
[9] J. Lončarić and S. V. Tsynkov, *Quadratic Optimization in the Problems of Active Control of Sound*, Technical report 2002-35, NASA/CR-2002-211939, ICASE, Hampton, VA, 2002 (also SIAM J. Appl. Math., submitted).
[10] C. L. Morfey, *Dictionary of Acoustics*, Academic Press, San Diego, 2001.
[11] P. A. Nelson and S. J. Elliot, *Active Control of Sound*, Academic Press, San Diego, 1999.
[12] Y. Nesterov and A. Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming*, SIAM Stud. Appl. Math. 13, SIAM, Philadelphia, 1994.
[13] Y. Nesterov and M. J. Todd, *Self-scaled barriers and interior-point methods for convex programming*, Math. Oper. Res., 22 (1997), pp. 1–42.
[14] T. W. Roberts, D. Sidilkover, and S. V. Tsynkov, *On the combined performance of non-local artificial boundary conditions with the new generation of advanced multigrid flow solvers*, Computers and Fluids, 31 (2001), pp. 269–308.
[15] V. S. Ryaben'kii, *Method of Difference Potentials and Its Applications*, Springer-Verlag, Berlin, 2002.
[16] R. T. Seeley, *Singular integrals and boundary value problems*, Amer. J. Math., 88 (1966), pp. 781–809.
[17] J. F. Sturm, *Using SeDuMi* 1.02, *A MATLAB toolbox for optimization over symmetric cones*, Optim. Methods Softw., 11–12 (1999), pp. 625–653; special issue on interior point methods (CD supplement with software).
[18] A. N. Tikhonov and A. A. Samarskii, *Equations of Mathematical Physics*, Pergamon Press, Oxford, 1963.
[19] S. V. Tsynkov, *Numerical solution of problems on unbounded domains. A review*, Appl. Numer. Math., 27 (1998), pp. 465–532.
[20] S. V. Tsynkov, *On the definition of surface potentials for finite-difference operators*, J. Sci. Comput., 18 (2003), pp. 155–189.
[21] R. J. Vanderbei, *Linear Programming: Foundations and Extensions*, Kluwer Academic Publishers, Boston, 2001.
[22] R. I. Veizman and V. S. Ryaben'kii, *Difference problems of screening and simulation*, Dokl. Akad. Nauk, 354 (1997), pp. 151–154.
[23] R. I. Veizman and V. S. Ryaben'kii, *Difference simulation problems*, Trans. Moscow Math. Soc., 58 (1997), pp. 239–248.
[24] V. S. Vladimirov, *Equations of Mathematical Physics*, Dekker, New York, 1971.
[25] Y. Ye, M. J. Todd, and S. Mizuno, *An $O(\sqrt{n}L)$-iteration homogeneous and self-dual linear programming algorithm*, Math. Oper. Res., 19 (1994), pp. 53–67.

SIAM J. APPL. MATH.
Vol. 63, No. 4, pp. 1184–1207

INTERMITTENCY IN THE TRANSITION TO TURBULENCE*

A. C. FOWLER† AND P. D. HOWELL†

Abstract. It is commonly known that the intermittent transition from laminar to turbulent flow in pipes occurs because, at intermediate values of a prescribed pressure drop, a purely laminar flow offers too little resistance, but a fully turbulent one offers too much. We propose a phenomenological model of the flow, which is able to explain this in a quantitative way through a hysteretic transition between laminar and turbulent "states," characterized by a disturbance amplitude variable that satisfies a natural type of evolution equation. The form of this equation is motivated by physical observations and derived by an averaging procedure, and we show that it naturally predicts disturbances having the characteristics of slugs and puffs. The model predicts oscillations similar to those which occur in intermittency in pipe flow, but it also predicts that stationary "biphasic" states can occur in sufficiently short pipes.

Key words. intermittency, transition, turbulence, slugs, puffs

AMS subject classifications. 76E30, 76F10

PII. S0036139900368893

1. Introduction.

Ever since Reynolds's (1883) seminal paper on the transition to turbulence in pipe flow, it has been known that the transition occurs in an intermittent fashion. As the Reynolds number increases beyond a value of around 2000 (although the precise value depends on the pipe used and on the experimental conditions at the inlet), intermittent flashes of turbulence can be seen in the pipe. Furthermore, the reason for this intermittency is well known, at least in a crude way (Prandtl and Tietjens (1934, pp. 36f.)). Turbulent flow at a given flow rate has a higher drag than laminar flow, and so, as the pressure drop driving the flow is increased, there arises a critical interval of flow rate within which laminar flow offers too low a resistance to the pressure drop but turbulent flow provides too high a resistance. In this intermediate case, the flow cycles between the two types of flow, and this is manifested in the pipe through the regular occurrence of turbulent "flashes"; this is Reynolds's term, but it has now become more customary to call the flashes "slugs" or "puffs" (Wygnanski and Champagne (1973)), depending on their provenance. The resultant flow then oscillates, producing an oscillatory (and indeed, periodic) outlet flow (Prandtl and Tietjens (1934, p. 37)).

It is perhaps unsurprising that there have been few attempts to recover these observations theoretically. Of necessity, any putative model must be semiempirical, and those that have been put forward (Bohr and Rand (1991), Deissler (1987a,b), Sakaguchi and Brand (1996)) serve as qualitative analogues rather than quantitative ones, and their aim has been to explain qualitatively the existence of turbulent slugs, rather than to draw a quantitative comparison; in addition, the resulting periodic solutions have not been found, although Deissler (1987b) hints at a mechanism similar to that suggested here.

Our aim in this paper is to provide a simple model which avoids the detailed complexities of three-dimensional turbulent flow, but which is nevertheless built solidly on

*Received by the editors March 10, 2000; accepted for publication (in revised form) August 9, 2002; published electronically March 26, 2003.
http://www.siam.org/journals/siap/63-4/36889.html
†Mathematical Institute, Oxford University, 24-29 St Giles', Oxford OX1 3LB, England (fowler@maths.ox.ac.uk, howell@maths.ox.ac.uk).

observed features of turbulence, and in particular, on the experimentally determined drag law. Although our motivating aim is to provide a predictive mechanism for intermittency and flow oscillations, we also find that we can make detailed comparisons with slug and puff dynamics. Wygnanski and Champagne (1973) distinguished these on the basis that slugs occurred at higher Reynolds number (above about 2700), were caused by low amplitude disturbances, and spread longitudinally as they propagated, the front and rear travelling at speeds respectively greater and less than the mean flow speed u, at least for large enough Reynolds number. In a very careful study, Lindgren (1957) showed that the front and rear wave speeds of slugs appeared to approach a value of about $0.9u$ as the Reynolds number decreased towards a value R_k (equal to about 2400 in his experiments), although in fact distinct slug measurements could be made only down to Reynolds number 2700.

At lower Reynolds number (below 2700 in Wygnanski and Champagne's experiments, presumably below 2400 in Lindgren's), puffs are seen. In contrast to slugs, puffs are generated by large amplitude disturbances at the inlet, and unlike slugs, which have relatively sharp leading and trailing edges, puffs have only a sharp trailing edge and a diffuse front. The trailing edge migrates backwards relative to the mean flow, with the difference between their two speeds tending to zero as the Reynolds number decreases towards a value $\underline{R}$ (about 2050 in Lindgren's experiments). Lindgren identifies a further Reynolds number $\overline{R}_k$, above which slugs grow as they propagate; presumably this is the Reynolds number at which stable slugs become viable. A final critical value is $\overline{R}$, above which fully developed turbulence can be maintained throughout the pipe.

2. A model for intermittency in turbulent flow.

2.1. Behavior of the wall friction. The Reynolds number for flow in a pipe of diameter d is

$$Re = \frac{\rho u d}{\mu}, \tag{2.1}$$

where u is the mean flow velocity, ρ is the density, and μ is the viscosity. In a pipe of length l, the pressure drop along its length in conditions of steady flow is given by

$$F = \frac{\Delta p}{l} = \frac{\lambda \rho u^2}{2d}, \tag{2.2}$$

where the drag coefficient λ is a function of the Reynolds number (Schlichting (1979)). In conditions of fully developed laminar flow, which pertain for $Re \lesssim 2300$,

$$\lambda_L = \frac{64}{Re}, \tag{2.3}$$

whereas Blasius's (1913) empirical relation for fully developed turbulent flow is

$$\lambda_T = \frac{0.3164}{Re^{1/4}} \tag{2.4}$$

and is approximately valid for $3300 < Re < 10^5$. At higher Reynolds number, a more sophisticated result based on Prandtl's mixing length theory can be used to define λ implicitly, but (2.4) will suffice for the present purpose, where such high Reynolds numbers are not relevant.

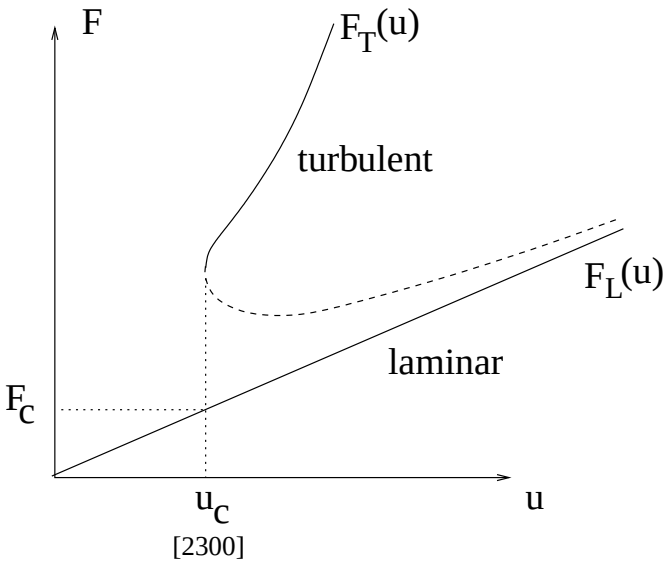


FIG. 1. *Schematic drag law for laminar and turbulent flow. The experimentally determined laminar and turbulent values are the solid curves, and we hypothesize that an unstable (dashed) branch separates the two, as shown.*

Between the onset of transition at $Re = 2300$ and the attainment of fully developed turbulence (throughout the pipe) at $Re \approx 3300$,[1] there is a region in which λ *increases* with Re. This odd behavior is associated with the phenomenon of *intermittency*. At Reynolds numbers above 2300, fluid can exist locally in a turbulent state, but for $Re < 3300$, turbulent slugs of fluid are interspersed with laminar plugs. The intermittency factor γ (the fraction of time at a fixed location for which the flow is turbulent) grows with distance downstream, and also with Reynolds number (Rotta (1956), Wygnanski and Champagne (1973)).

We wish to place a specific interpretation on this observed behavior. The local cross-sectionally averaged wall friction in a tube of diameter d is given from (2.2) by

$$F_L = \frac{32\mu u}{d^2} \tag{2.5}$$

for laminar flow (using (2.3)) and

$$F_T \approx \frac{0.16\rho^{3/4}\mu^{1/4}u^{7/4}}{d^{5/4}} \tag{2.6}$$

for turbulent flow.

As shown in Figure 1, the turbulent friction $F_T(u)$ exists as a local description down to Reynolds numbers of 2300, while the laminar expression $F_L(u)$ exists as a solution for all values of u. In particular, at values of Re above 2300, both behaviors are possible as locally stable solutions of the Navier–Stokes equation.

[1] This numerical value and that of the "onset" at 2300 depend on the level of the inlet disturbance, as well as the particular experimental set-up; these values are adopted from inspection of experimental drag measurements—see Schlichting (1979, Figure 20.1)—and will be used as typical values. In due course, we will relate them to the critical values discussed by Lindgren (1957).

Although it is not in fact essential to our argument, the existence of a jump in F between the two accessible solution branches suggests strongly that there is a third intermediate branch which joins the laminar and turbulent branches, as shown in Figure 1, and this is supported by other experimental work also (Huang and Huang (1989), Wygnanski and Champagne (1973), Darbyshire and Mullin (1995)). Moreover, the existence of an unstable intermediate branch is analogous to the existence of unstable equilibria in the transition to turbulence of plane and pipe Poiseuille flow (Orszag and Patera (1980), (1983)), and this lends support to the concept embodied in Figure 1.

We take the form of the dashed part of the $F(u)$ curve in Figure 1 as a hypothesis. It is then convenient to think of the local friction F as a state variable (rather like enthalpy), and to suppose that it is a measure of the laminar or turbulent "phase" of the fluid. And, rather like a phase change, the intermediate state is not accessible. In a fluid, boiling leads to dispersed phases at intermediate enthalpies, and transition to turbulence leads to intermittency at intermediate flow rates.

From the point of view of dynamical systems theory, the existence of a state variable F demarcating laminar and turbulent phases with a multiple valued equilibrium suggests that the simplest model beyond the mixing length theory which can describe transitions between laminar and turbulent states is one which embodies an evolution equation for F. In order to see how such an equation might be proposed, we need to study the way in which the basic mixing length theory produces the equilibrium structure of Figure 1.

2.2. Averaging. We start with the Navier–Stokes equations

$$\frac{\partial u_i}{\partial x_i} = 0,$$

$$\rho\left[\frac{\partial u_i}{\partial t} + \frac{\partial}{\partial x_j}(u_i u_j)\right] = -\frac{\partial p}{\partial x_i} + \mu\nabla^2 u_i. \tag{2.7}$$

Following common procedure (Mathieu and Scott (2000)), we define

$$u_i = \bar{u}_i + u_i', \tag{2.8}$$

where $\bar{u}_i$ is a local time average of u_i, and u_i' represents the fluctuating part. More specifically,

$$\bar{u}_i = \frac{1}{2T}\int_{t-T}^{t+T} u_i\, dt, \tag{2.9}$$

and we suppose formally that u_i' varies on a time scale $\ll T$, while we allow $\bar{u}_i$ to vary on times $\gg T$. (The necessity for this assumption can be removed by taking ensemble averages instead.)

Averaging of (2.7) thus leads to

$$\frac{\partial \bar{u}_i}{\partial x_i} = 0,$$

$$\rho\left[\frac{\partial \bar{u}_i}{\partial t} + \frac{\partial}{\partial x_j}(\bar{u}_i \bar{u}_j)\right] = -\frac{\partial \bar{p}}{\partial x_i} + \mu\nabla^2 \bar{u}_i + \frac{\partial}{\partial x_j}(-\rho\overline{u_i' u_j'}). \tag{2.10}$$

Next we define the cross-sectional average over the pipe as

$$\hat{h} = \frac{1}{S}\int_S h\, dS, \tag{2.11}$$

where S denotes the cross-sectional area. If we let $x\ (= x_1)$ denote distance down the pipe axis, then also

$$\widehat{\frac{\partial f_j}{\partial x_j}} = \frac{\partial \hat{f}_1}{\partial x} - \frac{1}{R}\langle f_n \rangle_w, \tag{2.12}$$

where $R = S/P$ is the hydraulic radius, P is the pipe perimeter, $f_n = \mathbf{f}.\mathbf{n}$ denotes the *inwards* normal component of $\mathbf{f}$ ($\mathbf{n}$ is the unit inward normal) and $\langle\ \rangle_w$ denotes the circumferential average,

$$\langle g \rangle_w = \frac{1}{P}\int_{\partial S} g\, ds. \tag{2.13}$$

We take a cross-sectional average of (2.10). We must have $\langle \bar{u}_n \rangle_w = 0$, thence $\partial \hat{\bar{u}}_1/\partial x = 0$; i.e., the mean flow is

$$\hat{\bar{u}}_1 = u(t), \tag{2.14}$$

say, and thus, considering the x component only,

$$\rho \dot{u} = -\frac{\partial p}{\partial x} - F + \frac{\partial}{\partial x}(-\rho \widehat{\overline{u_1'^2}}), \tag{2.15}$$

where we write $\hat{\bar{p}} = p$, and

$$F = \frac{\mu}{R}\left\langle \frac{\partial \bar{u}_1}{\partial n} \right\rangle_w + \frac{1}{R}\langle -\rho \overline{u_1' u_n'} \rangle_w. \tag{2.16}$$

Strictly, the second term on the right-hand side of (2.16) can be neglected, since $u_i' = 0$ at the wall. However, it is more common to define the "wall" in (2.13) to lie just outside the laminar sublayer, so that although $\bar{\mathbf{u}} \approx 0$ there, we allow the Reynolds stresses $(-\rho \overline{u_1' u_j'})$ to be nonzero at the wall. (In particular, this allows us to deal with rough walls.) We follow this practice here.

In conditions of steady uniform flow, (2.16) defines the wall drag, and Figure 1 represents the observed variation of F with u. In laminar flow,

$$\left\langle \frac{\partial \bar{u}_1}{\partial n} \right\rangle_w = \frac{8u}{d}, \tag{2.17}$$

while Prandtl's mixing length theory for turbulent flow also leads to $\langle \partial \bar{u}_1/\partial n \rangle_w \propto u/d$, although with a different coefficient. We will suppose that (2.17) applies in both cases, partly for simplicity, and partly because the laminar contribution is small in turbulent flow, so that the inaccuracy of (2.17) in that case is inconsequential. (A more realistic prescription for $\langle \partial \bar{u}_1/\partial n \rangle_w$ would be $L(A)u/d$, but we will persevere with (2.17).)

Given our constitutive prescription (2.17) for $\langle \partial \bar{u}_1/\partial n \rangle_w$, we now define a fluctuation velocity

$$A = \frac{d}{8\mu}\langle -\rho \overline{u_1' u_n'} \rangle_w. \tag{2.18}$$

It follows from (2.16) and (2.17) that (since the hydraulic radius of a pipe is $d/4$)

$$F = \frac{8\mu}{dR}[u + A]. \tag{2.19}$$

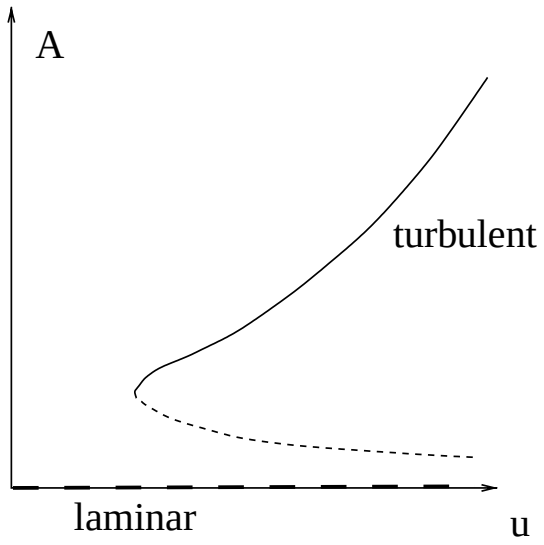


FIG. 2. *Diagram equivalent to Figure* 1 *for the postulated equilibrium, showing "amplitude" A versus flow rate u (or equivalently, Reynolds number).*

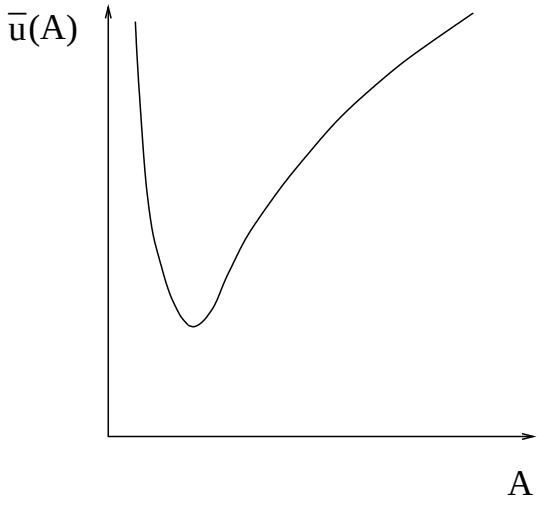


FIG. 3. *The function* $\bar{u}(A)$ *defined by reversing the axes in Figure* 2.

In terms of A, the equilibrium friction diagram Figure 1 now takes the form shown in Figure 2, and thus defines a single valued function $\bar{u}(A)$, such that in Figure 2, $u = \bar{u}(A)$ for $A \neq 0$, as shown in Figure 3.

The closure problem of turbulence is that of determining the Reynolds stresses $(-\rho\overline{u_i' u_j'})$. The idea of an eddy viscosity μ_T involves the closure assumption

$$-\rho\overline{u_i' u_j'} = \mu_T \left(\frac{\partial \bar{u}_i}{\partial x_j} + \frac{\partial \bar{u}_j}{\partial x_i} \right) - \frac{2}{3}\rho k \delta_{ij}, \tag{2.20}$$

where

$$k = \frac{1}{2}\overline{u_i' u_i'} \tag{2.21}$$

is the turbulent kinetic energy per unit mass. In unidirectional flows, Prandtl's mixing length hypothesis takes the form

$$\mu_T \propto \rho n^2 \left| \frac{\partial \bar{u}_1}{\partial n} \right|, \tag{2.22}$$

n representing distance from the wall. In a circular pipe, Schlichting (1979) shows that this empirical assumption actually leads to virtually perfect agreement with experimental measurements of wall drag. That is to say, in a steady uniform flow, F given by (2.16) together with an empirical assumption of eddy viscosity type for turbulent flow (and the equivalent exact molecular viscosity rule for laminar flow) is able to match the upper and lower branches in Figure 1 very well.

Prescription of (2.20) with (2.22) is an example of a zero-dimensional model closure. Such a closure is not good for three-dimensional flows, and this led to the development of more complicated "one-equation" or "two-equation" models. A typical example is the famous "k-ε" model (Mathieu and Scott (2000)), in which (2.20) is still used; one assumes (for example) that

$$\mu_T = \frac{Ck^2}{\varepsilon_D}, \tag{2.23}$$

where

$$\varepsilon_D = \frac{\mu}{2\rho} \overline{\left(\frac{\partial u_i'}{\partial x_j} + \frac{\partial u_j'}{\partial x_i} \right) \left(\frac{\partial u_i'}{\partial x_j} + \frac{\partial u_j'}{\partial x_i} \right)} \tag{2.24}$$

is the rate of dissipation of turbulent kinetic energy, and the model is closed by posing two evolution equations for k and ε_D (hence the term, two-equation model). In view of its similarity to the model we propose below, we give an example of a typical closure for k:

$$\frac{\partial k}{\partial t} + \bar{u}_k \frac{\partial k}{\partial x_k} = \Pi - \varepsilon_D + \frac{\partial}{\partial x_k} \left(\frac{\mu_T}{\sigma \rho} \frac{\partial k}{\partial x_k} \right); \tag{2.25}$$

Π is prescribed in terms of $\bar{u}$, σ is a constant. Note the basic type of advection-diffusion equation with source and sink terms.

In an analogous way, our purpose now is to consider how one might model slow time and space evolution of the fluctuation amplitude A. Derivation of an equation for A is not easy, nor is it our main purpose, and we will confine ourselves to providing a motivation for the form such an equation might take. We begin with the equations for the fluctuations u_i',

$$\frac{\partial u_i'}{\partial x_i} = 0,$$

$$\rho \left[\frac{\partial u_i'}{\partial t} + \frac{\partial}{\partial x_k} \{ \bar{u}_i u_k' + u_i' \bar{u}_k + u_i' u_k' - \overline{u_i' u_k'} \} \right] = -\frac{\partial p'}{\partial x_i} + \mu \nabla^2 u_i'. \tag{2.26}$$

Multiplying the second equation by u_j' and its u_j' equivalent by u_i', adding the two, and time averaging leads to the evolution equation for the Reynolds stress tensor

$$R_{ij} = -\rho\overline{u_i'u_j'} \tag{2.27}$$

in the form (see Launder, Reece, and Rodi (1975))

$$\begin{aligned}\frac{\partial R_{ij}}{\partial t} + \frac{\partial}{\partial x_k}[\bar{u}_k R_{ij}] = &-\left[R_{jk}\frac{\partial \bar{u}_i}{\partial x_k} + R_{ik}\frac{\partial \bar{u}_j}{\partial x_k}\right] + 2\mu\overline{\boldsymbol{\nabla} u_i' . \boldsymbol{\nabla} u_j'} - \overline{p'\left(\frac{\partial u_i'}{\partial x_j} + \frac{\partial u_j'}{\partial x_i}\right)} \\ &+ \frac{\partial}{\partial x_k}(\rho\overline{u_i'u_j'u_k'}) + \frac{\partial}{\partial x_k}[\overline{p'(u_i'\delta_{jk} + u_j'\delta_{ik})}] + \frac{\mu}{\rho}\nabla^2 R_{ij}.\end{aligned} \tag{2.28}$$

Launder, Reece, and Rodi (1975) characterize the terms in this equation in the following way. Of the six terms on the right-hand side of (2.28), the first represents generation and the second, dissipation; the last three represent transport, while the fourth (pressure strain) term is characterized by a spatial integral of the fluctuating velocity field. To see this, we take the divergence of the Navier–Stokes equation to find

$$\nabla^2 p = -\rho\frac{\partial^2(u_i u_j)}{\partial x_i \partial x_j}, \tag{2.29}$$

whence the fluctuating pressure field p' satisfies

$$\nabla^2 p' = -\rho\frac{\partial^2}{\partial x_i \partial x_j}[\bar{u}_i u_j' + u_i'\bar{u}_j + u_i'u_j' - \overline{u_i'u_j'}], \tag{2.30}$$

and p' can be written as a spatial integral convolving a suitable Green's function with the right-hand side of (2.30).

From (2.18), $A = (d/8\mu)\langle R_{1n}\rangle_w$, and in the absence of swirling motion, $\langle g\rangle_w = g$ for any of the terms g in (2.28). Also, on the wall $\bar{u}_i = 0$, and $\partial\bar{u}_1/\partial n$ is the only nonvanishing velocity derivative term ($\partial\bar{u}_n/\partial n = -\partial\bar{u}_1/\partial x = 0$). Setting $i = 1$, $j = n$, a wall average leads to

$$\begin{aligned}&\frac{\partial}{\partial t}\langle -\rho\overline{u_1'u_n'}\rangle_w + \frac{\partial}{\partial x}\langle -\rho\overline{u_1'^2u_n'}\rangle_w \\ &\qquad = \langle\rho\overline{u_n'^2}\rangle_w\left\langle\frac{\partial\bar{u}_1}{\partial n}\right\rangle_w + 2\mu\langle\overline{\boldsymbol{\nabla} u_1' . \boldsymbol{\nabla} u_n'}\rangle_w \\ &\qquad\quad - \left\langle\frac{\partial}{\partial n}(-\rho\overline{u_1'u_n'^2}) - \overline{u_1'\frac{\partial p'}{\partial n}} + \overline{p'\frac{\partial u_1'}{\partial n}} + \frac{\mu}{\rho}\frac{\partial^2}{\partial n^2}(-\rho\overline{u_1'u_n'})\right\rangle_w \\ &\qquad\quad + \frac{\partial}{\partial x}\left\langle\overline{p'u_n'}\right\rangle_w + \frac{\mu}{\rho}\frac{\partial^2}{\partial x^2}\langle -\rho\overline{u_1'u_n'}\rangle_w.\end{aligned} \tag{2.31}$$

It is not so easy to characterize the nature of the terms in (2.31) as we did for (2.28), because some of the transport terms in (2.28) involving normal derivatives can no longer be so categorized. Such terms, together with the pressure strain terms, form the third term on the right-hand side of (2.31). The other terms have direct analogy with those in (2.28). The second term on the left-hand side of (2.31) is a transport term and arises from the fourth term on the right-hand side of (2.28). The first and second terms on the right-hand side of (2.31) are source and sink terms, arising from the corresponding generation and dissipation terms in (2.28). The final two terms in (2.31) are transport terms.

2.3. Closure assumptions. We now face the daunting task of choosing constitutive laws for the various terms in (2.31). In choosing an eddy viscosity model, we specifically supposed (cf. (2.20))

$$\langle -\rho\overline{u_1' u_n'}\rangle_w = \mu_T \left(\frac{\partial \bar{u}_1}{\partial n} + \frac{\partial \bar{u}_n}{\partial x} \right)\Bigg|_w , \tag{2.32}$$

where μ_T itself depends on $\bar{u}_1$. In uniform flow ($\partial/\partial t = \partial/\partial x = 0$), this subsequently leads via Prandtl's mixing length theory to a functional relation between u and A. Given that we suppose u and A to be defined via (2.17) and (2.18), that is,

$$u = \frac{d}{8}\left\langle \frac{\partial \bar{u}_1}{\partial n} \right\rangle_w , \qquad A = \frac{d}{8\mu}\langle -\rho\overline{u_1' u_n'}\rangle_w , \tag{2.33}$$

it is natural to relate fluctuating terms in (2.31) to A, and terms in $\bar{u}_1$ to u. Furthermore, if we neglect $\partial/\partial t$ and $\partial/\partial x$, we should regain the equilibrium curve (with $A > 0$) of Figure 2, which we will suppose is written in the form $u = \bar{u}(A)$. At this point, we specifically ignore the problem of constituting the pressure fluctuation terms. As discussed above, we expect such terms to give rise to convolution integrals, but their omission will clarify the subsequent discussion without compromising the results. The above comments suggest (given (2.33)) that we choose

$$\begin{aligned} \langle \rho\overline{u_n'^2}\rangle_w \left\langle \frac{\partial \bar{u}_1}{\partial n} \right\rangle_w - \left\langle \frac{\partial}{\partial n}(-\rho\overline{u_1' u_n'^2}) - 2\mu\overline{\boldsymbol{\nabla} u_1' . \boldsymbol{\nabla} u_n'} + \frac{\mu}{\rho}\frac{\partial^2}{\partial n^2}(-\rho\overline{u_1' u_n'}) \right\rangle_w \\ = \frac{8\mu}{d} r(A)[u - \bar{u}(A)], \end{aligned} \tag{2.34}$$

where

$$r(A) = \frac{1}{\mu}\langle \rho\overline{u_n'^2}\rangle_w \tag{2.35}$$

is a positive function, with $dr/dA > 0$ and $r = 0$ when $A = 0$. The last term in (2.31) is a laminar diffusion term,

$$\frac{\mu}{\rho}\frac{\partial^2}{\partial x^2}\langle -\rho\overline{u_1' u_n'}\rangle_w = \frac{\mu}{\rho}.\frac{8\mu}{d}\frac{\partial^2 A}{\partial x^2}, \tag{2.36}$$

and it is plausible to expect an equivalent turbulent diffusive term to exist also.

There are two ingredients that we might expect to find in $\langle -\rho\overline{u_1'^2 u_n'}\rangle_w$. If we follow the (apparently arbitrary) recipe (2.32), which relates $\langle\overline{u_1' u_n'}\rangle_w$ to $\partial\bar{u}_1/\partial n$ and $\partial\bar{u}_n/\partial x$, we would equivalently write

$$\langle\overline{(-\rho u_1' u_n')u_1'}\rangle_w = -\frac{\mu_T}{\rho}\frac{\partial}{\partial x}\langle -\rho\overline{u_1' u_n'}\rangle_w + \cdots , \tag{2.37}$$

but also

$$\langle -\rho\overline{u_1'^2 u_n'}\rangle_w = \mu_T \frac{\partial}{\partial n}\langle\overline{u_1'^2}\rangle_w + \cdots . \tag{2.38}$$

The simplest choice is then

$$\langle -\rho\overline{u_1'^2 u_n'}\rangle_w = \frac{\mu_T}{\rho}\frac{\partial}{\partial x}\langle -\rho\overline{u_1' u_n'}\rangle_w + \mu_T\frac{\partial}{\partial n}\langle\overline{u_1'^2}\rangle_w . \tag{2.39}$$

Finally, it seems reasonable to propose

$$\mu_T \frac{\partial}{\partial n}\langle\overline{u_1'^2}\rangle_w = \frac{8\mu}{d}W(A,u), \tag{2.40}$$

with $W \geq 0$ as we expect $\partial\overline{u'^2}/\partial n$ to be positive at the wall: as for $r(A)$, we expect $W = 0$ when $A = 0$ and $\partial W/\partial A > 0$.

Multiplying (2.31) by $d/8\mu$, we have (ignoring the pressure fluctuation terms)

$$\begin{aligned}\frac{\partial A}{\partial t} + \frac{\partial}{\partial x}\left[-\frac{\mu_T}{\rho}\frac{\partial A}{\partial x} + W(A,u)\right]\\ = r(A)[u - \bar{u}(A)] + \frac{\mu}{\rho}\frac{\partial^2 A}{\partial x^2}.\end{aligned} \tag{2.41}$$

The functions $W(A,u)$ and $r(A)$ in (2.41) are matters of conjecture. To be specific, we now define

$$U = \frac{\partial W}{\partial A} = \frac{\partial\left[\mu_T \frac{\partial}{\partial n}\langle\overline{u_1'^2}\rangle_w\right]}{\partial\langle-\rho\overline{u_1' u_n'}\rangle_w}, \tag{2.42}$$

so that $U > 0$ is an advection velocity. If we define

$$D = \frac{(\mu + \mu_T)}{\rho} \tag{2.43}$$

in (2.41), then our simple model for A is

$$A_t + UA_x = r[u - \bar{u}(A)] + (DA_x)_x, \tag{2.44}$$

where $\bar{u}(A)$ indicates the single valued equilibrium curve with $A \neq 0$ in Figure 2. This equation represents the idea that A evolves with time while the fluid is advected, seeking the equilibria in Figure 2. The principal difference that a more detailed prescription might make would be the inclusion of integral terms in (2.44) corresponding to the pressure fluctuation terms which we have ignored.

We now discuss suitable choices for U, r, and D. It is generally thought that pipe Poiseuille flow is linearly stable at all Reynolds numbers (Drazin and Reid (1981)), although the smallest disturbance decay rates tend to zero as $Re \to \infty$. This will be the case here if $r\bar{u} \propto A$ and $r = o(A)$ as $A \to 0$. Experimental measurements (Wygnanski and Champagne (1973), Darbyshire and Mullin (1995)) are suggestive of an asymptotic relationship $\bar{u} \sim A^{-2}$ as $\bar{u} \to \infty$ (see Figure 4), and we will make this assumption here. It then follows that an appropriate choice for r is

$$r = kA^3. \tag{2.45}$$

Then (2.44) allows $A = 0$ to be linearly stable. An issue here concerns the measurement of experimental disturbance amplitudes. Darbyshire and Mullin (1995) indicate absolute amplitudes, whereas Wygnanski and Champagne's (1973) data (in Figure 4) apparently measures amplitude ratios. There are two points of observation to make. The first is that the data points at higher Re in Figure 4 will place the entire pipe flow in the inlet region, which jeopardizes the interpretation of the amplitude as a perturbation to fully developed flow. The other is that if the amplitude ratio did indeed decrease as weakly as suggested by Figure 4, then the absolute amplitude would

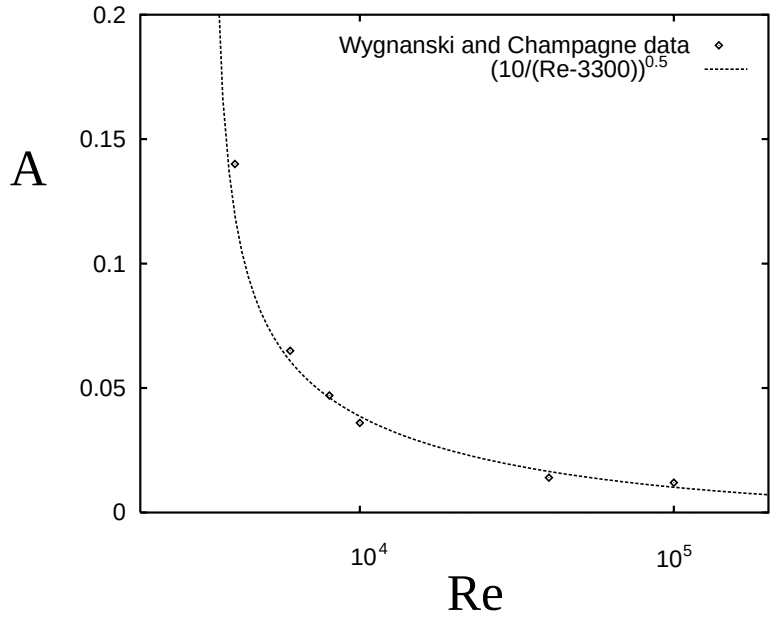


FIG. 4. *Data taken by hand from the lower threshold of the slug transition curve of Wygnanski and Champagne* (1973) *indicating threshold values of A (percent), together with the fitted curve* $A = \{10/(Re - 3300)\}^{1/2}$.

eventually increase with Re. This does not seem to be suggested by experiment, and is inconsistent with linear theory. We therefore hold to our interpretation of A in (2.44) as an absolute amplitude.

The constant k should represent the idea that the laminar-turbulent transition is associated with a rapid inviscid three-dimensional instability of slowly decaying two-dimensional disturbances (Orszag and Patera (1980)). This suggests that $t \sim d/u \sim 1/kA^2u$, and it is thus suitable to choose (for example)

$$k = \frac{1}{[A]^2 d}\,, \tag{2.46}$$

where $[A]$ is a representative value of A near transition.

The amplitude equation (2.44) is then in fact, when $A \ll u$, exactly that which can be derived using multiple scale methods (Davey and Nguyen (1971)), although the correctness of such equations is in doubt (Itoh (1977), Davey (1978)). However, it does seem likely that some such unstable solution branch bifurcates from infinity, as is suggested by the work of Rosenblat and Davis (1979) and Smith and Bodonyi (1982). We do not claim that (2.44) represents the last word in modelling slowly varying fluctuations, but its form is consistent with what is known about pipe flow. As a model, it serves the same purpose as Burgers's (1948) model, and indeed it bears some resemblance, although Burgers's model was more concerned with cross-stream transport of energy. The model (2.44) is also analogous (and serves a similar purpose) to the Swift–Hohenberg-type amplitude model, which has been used by some authors with a purpose similar to that of the present paper (Sakaguchi and Brand (1996), Deissler (1987a)).

The diffusion coefficient D in (2.43) has two parts representative of a laminar

and a (larger) turbulent diffusivity, and so we take $D \sim ud$ to represent the latter quantity. However, we should also allow D to depend on A, since when A is small, D is represented by molecular viscosity. We will find that the diffusive term allows for the existence of convectively growing slugs (Bohr and Rand (1991), Deissler (1987b)), with front and rear ends which travel respectively faster and slower than the mean flow, as is observed in practice.

The choice of a suitable advection velocity U is less clear. Most simply, we would take $U = u$, the mean flow. However, for marginally stable modes at high Re, the relevant linear stability advection velocity (Drazin and Reid (1981)) is less than u, and we therefore allow U to be less than u and to depend on A.

2.4. A model for intermittency. We can then provide a putative model for intermittent turbulent flow as follows. For an incompressible (thus $u = u(t)$) fluid flow subject to a mean pressure gradient F_m in a pipe $0 < x < l$, we solve

$$A_t + UA_x = kA^3[u - \bar{u}(A)] + (DA_x)_x. \tag{2.47}$$

In (2.15) we neglect the term $\frac{\partial}{\partial x}(-\rho\widehat{\overline{u_1'^2}})$, partly because we expect it to be small and partly because the natural constitution of $\overline{u_1'^2} \propto u^2$ leads to its absence, since $u = u(t)$. Then u and p satisfy

$$\begin{aligned} \rho\dot{u} &= -p_x - F, \\ \int_0^l -p_x\,dx &= F_m l, \end{aligned} \tag{2.48}$$

which together imply (using the definition of F in (2.19))

$$\rho\dot{u} = F_m - \frac{32\mu}{d^2}\left[u + \frac{1}{l}\int_0^l A\,dx\right]. \tag{2.49}$$

Deissler (1987b) has suggested that if F_m is prescribed, then the pressure drop feedback (i.e., (2.49)) may cause intermittency; we will seek to establish this suggestion here.

The pair of equations (2.47) and (2.49) require two boundary conditions for A (as well as initial conditions for A and u). A primary observation of transition in pipe flow ever since the experiments of Reynolds (1883) is that the level of inlet disturbance is instrumental in determining the nature of the flow. Therefore we prescribe

$$A = A_0 \quad \text{at } x = 0, \tag{2.50}$$

and we expect that the value of A_0 will be important in determining the dynamics of the flow.

The outlet condition is required only if the diffusion term is included, and, in order to exclude physically inappropriate boundary layers at the channel outlet, we prescribe the passive condition

$$A_x = 0 \quad \text{at } x = l \tag{2.51}$$

there.

2.5. Nondimensionalization. We choose scales $[F]$, $[A]$, $[u]$, $[t]$, $[x]$ as follows: $[F]$ is chosen as the critical value $F_c = 32\mu u_c/d^2$ in Figure 1, where the laminar flow can become unstable to turbulent bursts, and $[u] = u_c$ is the corresponding mean flow. Then we choose $[A] = [u]$, $[t] = \rho[u]/[F]$, and $[x] = [u][t]$; the advection velocity U is scaled with $[u]$. We then find that

$$[x] = \frac{d}{\varepsilon}, \tag{2.52}$$

where

$$\varepsilon = \frac{32}{[Re]}, \tag{2.53}$$

and the Reynolds number scale is

$$[Re] = \frac{\rho[u]d}{\mu}; \tag{2.54}$$

since $\varepsilon \ll 1$ ($\varepsilon \approx 0.015$ for $[Re] = 2300$), we see that $[x] \gg d$. The scale $[x]$, in fact, describes the inlet region of the pipe (Goldstein (1938, pp. 299f.)). The dimensionless model becomes (using the same notation for the dimensionless variables)

$$\begin{aligned} \varepsilon[A_t + UA_x] &= A^3[u - \bar{u}(A)] + \varepsilon^2(\kappa A_x)_x, \\ \dot{u} &= F^* - u - \frac{1}{L}\int_0^L A\,dx, \end{aligned} \tag{2.55}$$

where the dimensionless parameters are

$$L = \frac{\varepsilon l}{d}, \quad \kappa = \frac{D}{d[u]}, \quad F^* = \frac{F_m}{F_c}, \tag{2.56}$$

and, in keeping with (2.46), we have defined

$$k = \frac{1}{u_c^2 d}. \tag{2.57}$$

The dimensionless drag F is given for laminar flow by $F = u$, and Blasius's law for turbulent flow (2.6) is

$$F \approx \frac{[Re]^{3/4}}{200}u^{7/4} \approx 1.66\,u^{7/4} \tag{2.58}$$

for $u \gtrsim 1$, where the Reynolds number scale is taken to be 2300, by choice of $[u]$. This yields

$$A \approx 1.66\,u^{7/4} - u \quad \text{for} \quad u \gtrsim 1 \tag{2.59}$$

for the upper part of the curve in Figure 2. Over the range $1 \lesssim u \lesssim 4$, corresponding to Reynolds numbers up to 10^4, (2.59) is well approximated by $A \approx u^2 - \frac{1}{4}u$, i.e., $u \approx \frac{1}{8} + (A + \frac{1}{64})^{1/2}$ for $0.75 \lesssim A \lesssim 15$, and this in turn is well approximated by $u \approx 0.7 + 0.6\,A^{0.64}$. Our choice for $\bar{u}(A)$ is thus motivated by this expression, with an extra term $\propto 1/A^2$ added (in order to provide the unstable branch in Figure 2, and to

ensure linear decay at small A), with a coefficient δ chosen so that $\bar{u}$ has a minimum at $\bar{u} = 1$. Thus, we choose

$$\bar{u}(A) = a + bA^s + \frac{\delta}{A^2}, \tag{2.60}$$

and the values of the parameters motivated by the above discussion are

$$a = 0.7, \quad b = 0.6, \quad s = 0.64, \quad \delta = 0.0035. \tag{2.61}$$

It is a happy fact that the resulting decay rate of small disturbances $A \ll 1$ is $\dot{A}/A \approx -\delta$, which corresponds to the slow viscous decay rate ($\delta \sim \varepsilon \approx 0.014$), as is appropriate. Figure 16.3 of Schlichting (1979) indicates that a natural pipe length scale over which the intermittency factor γ grows is $l/d \sim 300$. For such pipe lengths (in a 1 cm diameter pipe this is $l = 3$ m), $L \approx 4.5$, so that values of $L \gtrsim 1$ are appropriate.

The choice of the advection velocity U and the diffusivity κ are as follows. For turbulent flow, we expect $\mu_T \sim \rho[u]d$; hence (2.43) implies $\kappa \sim 1$, but for small A it should be $O(\varepsilon)$, corresponding to the laminar viscosity term. The simplest choice compatible with these criteria is

$$\kappa = A + \varepsilon, \tag{2.62}$$

although in our numerical illustrations we will be content with the choice $\kappa = 1$. For the advection velocity, the choice $U = u$ corresponds to the mean flow. We define

$$U = u - V(A) \tag{2.63}$$

and will find that the corrective term to the mean flow allows us to include a realistic description of puffs within this simple theory, providing $V > 0$, as we expect.

3. Analysis.

3.1. Constant velocity.

We begin our analysis of (2.55) by supposing that the inlet velocity u is constant. We write the amplitude equation in the form

$$\varepsilon(A_t + UA_x) = f(A; u) + \varepsilon^2(\kappa A_x)_x, \tag{3.1}$$

where $U = u - V(A)$, $\kappa = A + \varepsilon$, and

$$f(A; u) = (u - a)A^3 - bA^{3+s} - \delta A. \tag{3.2}$$

The form of slugs. Clearly, A evolves locally over a rapid time scale of $O(\varepsilon)$ to a steady state of (3.1), i.e., to $A = 0$ (laminar) or to the stable (turbulent) branch of $u = \bar{u}(A)$ (see Figure 2). As shown in Figure 5, f is pseudocubic, and for $u > 1$ (the minimum value of $\bar{u}(A)$, corresponding here to $Re = 2300$) there are three roots of f: $A = 0$ and the stable and unstable branches, which are defined to be $A = A_m(u)$ and $A_M(u)$, respectively.

We change coordinates to the local moving frame coordinates (X, T) given by

$$x = ut + \varepsilon X, \quad t = \varepsilon T, \tag{3.3}$$

whence (3.1) becomes

$$A_T - V(A)A_X = f(A) + \{\kappa(A)A_X\}_X. \tag{3.4}$$

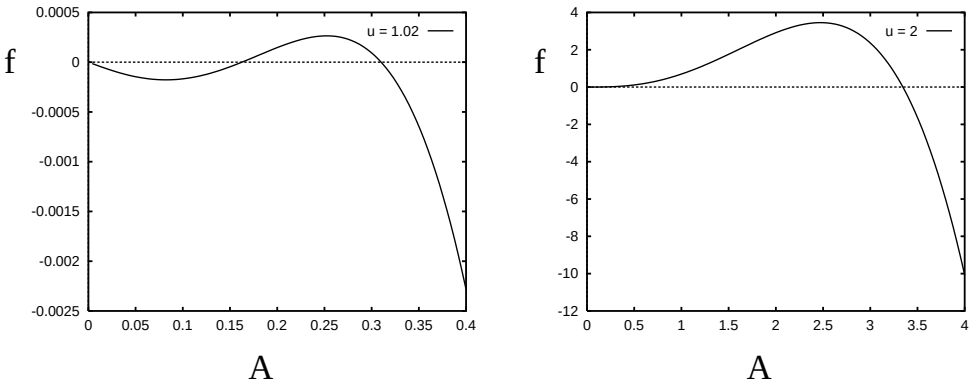


FIG. 5. *The function $f(A,u)$ for $u = 1.02$ and $u = 2$. Note the different vertical and horizontal scales.*

It is well known from reaction-diffusion theory (Murray 1993) that a local perturbation $A > A_m$ will evolve to a slug within which $A = A_M$, and the front and rear move *outward* at speeds v_+ and v_-, respectively, given by

$$v_\pm = \frac{\int_0^{A_M} \kappa(A) f(A)\,dA \mp \int_{-\infty}^{\infty} \kappa(A) V(A) A'^2\,dz}{\int_{-\infty}^{\infty} \kappa(A) A'^2\,dz}, \tag{3.5}$$

where $A(z)$ is the wave front solution of

$$-v_\pm A' \mp V(A)A' = f(A) + \{\kappa(A)A'\}', \tag{3.6}$$

together with

$$\begin{aligned} A &\to A_M \quad \text{as} \quad z \to -\infty, \\ A &\to 0 \qquad \text{as} \quad z \to +\infty. \end{aligned} \tag{3.7}$$

Note that the two wave front profiles are different if $V \neq 0$ (and then $v_+ \neq v_-$). The coordinate $z = \pm X - v_\pm T$ is the wave front variable. From (2.30), we define the front and rear slug boundary speeds u_f and u_r as

$$\begin{aligned} u_f &= u + v_+, \\ u_r &= u - v_-. \end{aligned} \tag{3.8}$$

If $V = 0$, so that the wave profiles are symmetric, then $v_+ = v_-$ and the mean slug speed is that of the mean flow; however, this is not observed.

The data in Figure 7 of Wygnanski and Champagne (1973), and in Figure 4.15 of Lindgren (1957), can be used to constrain, to some extent, our choices for $\kappa(A)$ and, particularly, $V(A)$. First, note that if $v_+ + v_- < 0$, then slugs contract and are not viable. If V is small, then this occurs if $\int_0^{A_M} \kappa(A) f(A)\,dA < 0$. We can see from the form of Figure 5 that there will be a critical value of $u = u^* > 1$ where $\int_0^{A_M(u^*)} \kappa(A) f(A;u^*)\,dA = 0$, so that slugs are only viable if $u > u^*$. For $V \neq 0$, we can expect u^* to depend on V. In Wygnanski and Champagne's data, the corresponding Reynolds number is 2700, whence $u^* \approx 1.17$ (based on a Reynolds number scale of 2300). This is also the minimum value for which Lindgren provides identifiable slug data. For $u > u^*$, the data indicate that u_f/u and u_r/u are functions

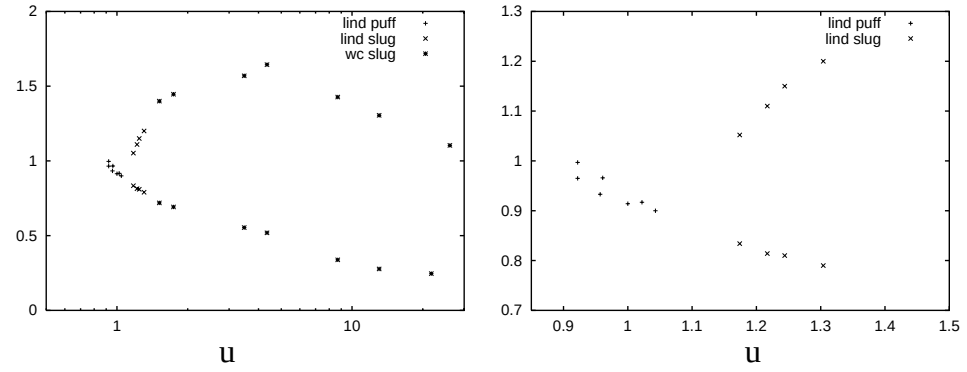


FIG. 6. *Estimates for the relative front and rear speeds u_f/u and u_r/u, redrawn from the data of Wygnanski and Champagne* (1973) *and Lindgren* (1957). *On the right is a close-up of (some of) Lindgren's data.*

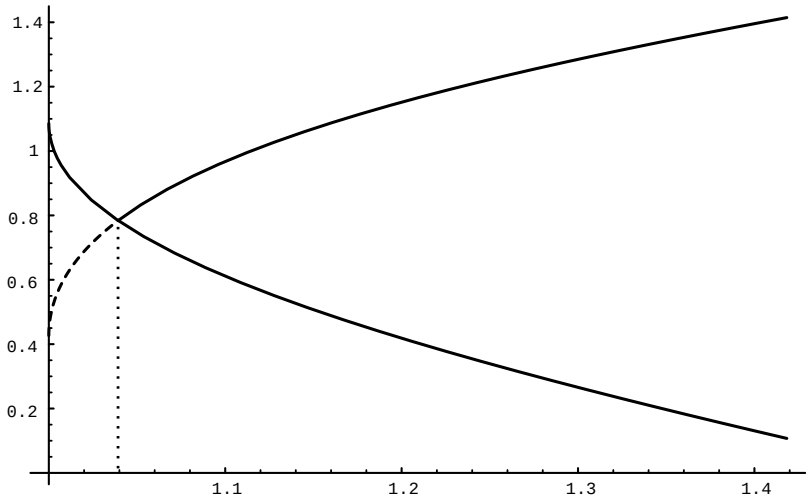


FIG. 7. *Front and rear relative wave speeds u_f/u and u_r/u, as a function of u. The parameter values used are $b = 0.6$, $s = 0.64$, and $\delta = 0.1$, with the drift velocity $V(A)$ being taken as $V = 0.4A$ and the diffusivity $\kappa = 1$. The critical value at which the two speeds coalesce is approximately* 1.04 *in this case, and the critical relative speed is about* 0.8. *With* $[Re] = 2300$, *this corresponds to a value of R_k (in Lindgren's* (1957) *notation) of* 2392.

of u, as shown in Figure 6; they appear to coalesce when $Re = R_k \approx 2400$, at a value of about 0.9. We have solved the travelling wave equation numerically, to reproduce data equivalent to that shown in Figure 6. Figure 7 shows the results of these calculations; they are in good qualitative agreement with Lindgren's (1957) results.

A feature of a slug is its sharp interface. In the travelling wave solutions of (3.6), this sharpness is represented both by the short wavelength ε (corresponding to a tube diameter) and also by the fact that (approximately) $\kappa \to 0$ as $A \to 0$; the diffusivity is degenerate and this causes A to reach (approximately) zero in a finite distance.

It is straightforward to compute the value of u^* (if $V = 0$) explicitly from the

expression (3.2) for f. We suppose $\kappa = A$, so that A_M and u^* are determined from

$$\begin{aligned} &(u-a)A^3 - bA^{3+s} - \delta A = 0, \\ &\frac{(u-a)}{5}A^5 - \frac{b}{5+s}A^{5+s} - \frac{\delta A^3}{3} = 0; \end{aligned} \tag{3.9}$$

from these we find

$$\begin{aligned} A_M &= \left[\frac{2\delta(5+s)}{3sb}\right]^{\frac{1}{2+s}}, \\ u^* &= a + 5(2+s)\left(\frac{\delta}{3s}\right)^{\frac{s}{2+s}}\left(\frac{b}{2(5+s)}\right)^{\frac{2}{2+s}}, \end{aligned} \tag{3.10}$$

and for the numerical values given in (2.61), we find $u^* \approx 1.01$, $A_M \approx 0.28$.

In this theory, u^* is the critical velocity required for slug propagation. Because $\delta \ll 1$, we see that, for $u > u^*$, the unstable zero A_m of f is approximately given by

$$A_m \approx \left(\frac{\delta}{u-a}\right)^{1/2}. \tag{3.11}$$

Note the similarity to the fitted curve in Figure 4. Thus small amplitude disturbances of $O(\delta^{1/2})$ spontaneously generate slugs, which grow as they propagate.

The form of puffs. When $u < u^*$, slugs cannot be maintained. In this case, large disturbances will cause puffs to propagate. These migrate slowly backwards relative to the mean flow, with a sharp trailing edge and a diffuse advancing boundary. In order to explain their characteristics, we observe first that when $u < u^*$, then f is small and A_M is as well (see Figure 5). In fact, (3.10) suggests that we write

$$A = \delta^{\frac{1}{2+s}}\alpha, \quad u - a = \delta^{\frac{s}{2+s}}w \tag{3.12}$$

in this case, and then

$$\begin{aligned} f(A,u) &= \delta^{\frac{3+s}{2+s}}F(\alpha,w), \\ F &= w\alpha^3 - b\alpha^{3+s} - \alpha, \end{aligned} \tag{3.13}$$

and (3.4) is

$$\alpha_T - V[\delta^{\frac{1}{2+s}}\alpha]\alpha_X = \delta F(\alpha,w) + \{\kappa[\delta^{\frac{1}{2+s}}\alpha]\alpha_X\}_X. \tag{3.14}$$

Now let us suppose, for example, that $V(A) = cA$ (as in Figure 7), $\kappa(A) \approx A$. If we define the slow time scale τ by

$$T = \frac{\tau}{\delta^{\frac{1}{2+s}}}, \tag{3.15}$$

then α satisfies the approximate equation

$$\alpha_\tau - c\alpha\alpha_X = (\alpha\alpha_X)_X + \delta^{\frac{1+s}{2+s}}F(\alpha;w), \tag{3.16}$$

which is, to leading order, a degenerate nonlinear diffusion equation of Burgers type.

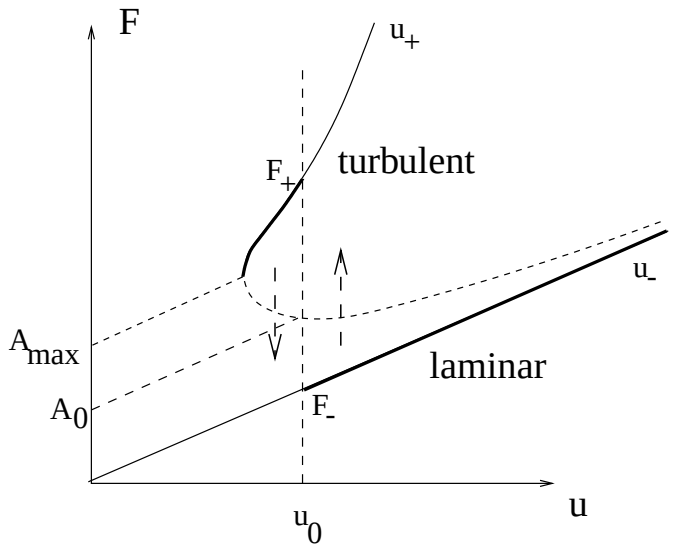


FIG. 8. *Diagrammatic representation of intermittency. The inlet disturbance amplitude A_0 determines a threshold velocity u_0 on the unstable branch, such that the turbulent branch is stable for $u > u_0$ and the laminar branch is stable for $u < u_0$, as indicated by the arrows. If the prescribed pressure drop F^* lies between the corresponding intersection values $F_\pm$ of $u = u_0$ with the turbulent and laminar branches, then an intermittent flow will ensue.*

Solutions behave as follows. Let $\alpha_m = O(1)$ denote the lower positive zero of F in (3.13). For prolonged perturbations to A in excess of $\delta^{\frac{1}{2+s}}\alpha_m$, a puff will form. The turbulent flow within the puff will be locally stable, but will eventually disappear by wastage of the profile. However, before this happens, the disturbance amplitude α will evolve over a time scale $\tau \sim 1$, i.e., $T \sim 1/\delta^{\frac{1}{2+s}}$, according to

$$\alpha_\tau - c\alpha\alpha_X = (\alpha\alpha_X)_X, \tag{3.17}$$

and thus into a profile with a sharp upstream (shock) profile and a diffuse downstream profile. The speed of the upstream front, relative to the mean velocity, is negative and of order $dX/dT \sim \delta^{\frac{1}{2+s}}$. All of these features are consistent with puffs. In particular, note that the necessary disturbance amplitude is $O(\delta^{\frac{1}{2+s}})$, as opposed to the smaller threshold $O(\delta^{1/2})$ for slugs. The formation of a backward propagating puff relative to the mean flow relies on the sign of V being positive. If this is the case, then the simple theory based on (3.4) is sufficient to explain many of the pertinent facts concerning slugs and puffs.

3.2. A mechanism for intermittency. The above discussion indicates that the amplitude evolution equation (3.1) can explain both artificially generated slugs and puffs if the flow rate is constant. In order to explain how intermittency can arise spontaneously, we must return to the pressure driven flow model given by (2.55).

Suppose that the inlet disturbance level is $A = A_0$. Since f is an increasing function of u, there is a unique value $u = u_0$ for which $A_m(u_0) = A_0$. Suppose that $u < u_0$. Then $A \to 0$ in $x > 0$ (see Figure 8), and (eventually) $u \to F^*$ from (2.55). However, if $F^* > u_0$, then at some point u reaches u_0, and for $u > u_0$, $A \to A_M(u)$ near the inlet: a slug is generated. Again (eventually) u tends to the positive root

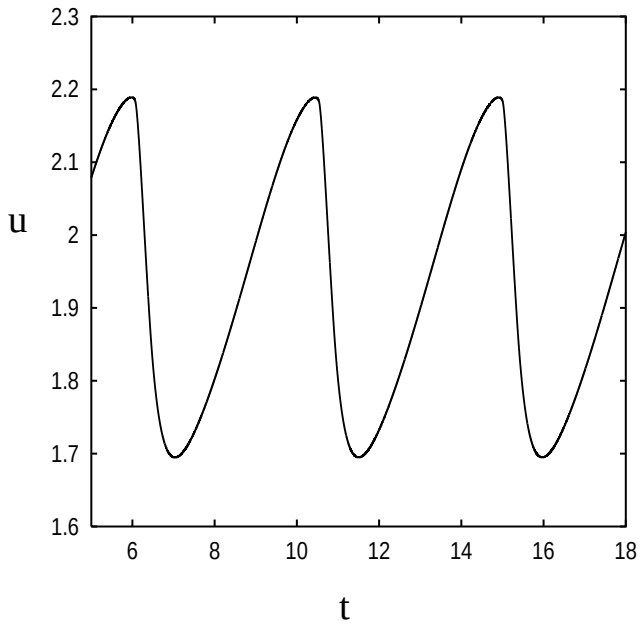


FIG. 9. *Mean velocity fluctuations in the solution of* (2.55). *The parameter values used are* $L = 4$, $A_0 = 0.3$, $\varepsilon = 0.05$, $\kappa(A) = 1$, $U(A,u) = u$, $a = 0.7$, $b = 0.6$, $s = 0.64$, $\delta = 0.1$, $F^* = 3$, $u(0) = 3.1$. *The space step was* 0.01, *and the time step was* 0.001.

of $F^* = u + A_M(u)$. (Note that this is simply the upper branch of the $F(u)$ curve in Figure 1.) But if the corresponding value of $u < u_0$, then again A decreases towards zero (a slug is terminated). It is fairly clear that this sequence will continue to oscillate, and also that the dependence of u on the spatial integral of A will cause a finite sequence of slugs to propagate downstream. The intermittency factor γ of Rotta (1956) will increase with x due to the spreading of the slugs.

Figure 8 illustrates this description graphically. Observe that the preceding paragraph indicates that $u \to u_-(F^*)$ if $u < u_0(A_0)$ and $u \to u_+(F^*)$ if $u > u_0(A_0)$, where $u_\pm(F^*)$ are the turbulent and laminar branches of Figure 8. Thus, intermittency should occur, given F^* and A_0, if

$$u_+(F^*) < u_0(A_0) < u_-(F^*) \tag{3.18}$$

in Figure 8. This depends on both the prescribed pressure drop F^* *and* the inlet disturbance amplitude A_0. The level curves of A $(= F - u)$ are simply lines parallel to the laminar branch u_-. Furthermore, $u_0(A_0)$ is simply the unstable branch of the $\bar{u}(A)$ curve. Thus, given A_0, we determine u_0 as the value of u at which the line $F = u + A_0$ intersects the unstable branch in Figure 8. So long as $A_0 < A_{\max}$ (the value of $F - u$ at the nose of the curve), the value u_0 defines two values F_+ and F_- on the stable turbulent and laminar branches, respectively. Intermittency then occurs if

$$F_- < F^* < F_+, \tag{3.19}$$

which is equivalent to (3.18).

Numerical results. We have solved the system (2.55) numerically, taking $U = u$ and $\kappa = 1$. Figure 9 shows the resulting periodic variations in the mean velocity. This

figure can be favorably compared with Figure 19 of Prandtl and Tietjens (1934). We have not had to tune the model, and so we believe this behavior to be robust. The choice of functions U and κ is immaterial to the phenomenon of intermittency; the advection U affects the specific form of puffs, while we avoid the degenerate $\kappa = A$ in order to avoid numerical awkwardness at slug boundaries. Similarly we choose a relatively high value of ε (0.05) (and hence also δ) in order to avoid the demands of excessively small space steps.

Figure 10 shows the space-time evolution of $A(x,t)$ for the same parameter values as used in Figure 9. After an initial transient, a periodic sequence of slugs is generated at some distance from the inlet and propagates downstream towards the outlet.

4. Discussion. Our primary purpose in this paper was to develop a simple model which could predict the intermittent transition to turbulence which is seen in pipe flow, and we have shown that this can be done using the observed fact that there is a sudden increase in friction at the onset of turbulent flow. We associate this with a hysteretic transition between laminar and turbulent states, characterized by a turbulent fluctuation amplitude A. We have then used the observed drag law to build the simplest evolution equation for A that is consistent with both the drag law and an inferred linear stability at small amplitudes. We have also shown that this form of amplitude equation is consistent with a one-equation closure of the time-averaged Navier–Stokes equations resembling those of R_{ij}-ε type (Launder, Reece, and Rodi (1975), Mathieu and Scott (2000)).

The inclusion of realistic diffusive and advective terms then allows us to describe, within the confines of this evolution equation model, phenomena which can be characterized as slugs or puffs, and we have shown that many of their peculiar features arise naturally from the simple ingredients of the model.

In this discussion, we wish to focus further on two particular experimental observations of the parameter ranges in which laminar, intermittent, or turbulent behavior occurs.

Lindgren (1957) identified four particular transition values of the Reynolds number, which he denoted as $\underline{R}$, R_k, $\overline{R}_k$, and $\overline{R}$. The value of $\underline{R}$ occurs when the first self-maintaining puffs are seen (i.e., with an identifiable tail velocity). In Lindgren's experiments, this value is about 2050. The value of R_k (about 2400) occurs when slugs are first seen (with identifiable fronts), and the value of $\overline{R}_k$ (about 2700) is where these become coherent, that is, they do not split as they propagate. Finally, $\overline{R}$ (about 3300?) denotes the onset of fully developed turbulence. The first three of these values correspond in Figure 6 to values $u \approx 0.9$, $u \approx 1.04$, and $u \approx 1.17$. In our model (cf. Figure 5) we can identify two critical values of u (and hence Re). Denoting the smaller and larger positive roots of (3.2) by $A_m(u)$ and $A_M(u)$, respectively, there is a critical value u_1 at which $A_m = A_M$, and a larger value u_2 for which the wave speeds v_+ and v_- given by (3.6) and (3.7) sum to zero. When V is small, $u_2 \approx u^*$, as discussed following (3.8), and explicit formulae for these critical values are

$$u_1 = a + (2+s)\left(\frac{\delta}{s}\right)^{\frac{s}{2+s}}\left(\frac{b}{2}\right)^{\frac{2}{2+s}},$$

$$u_2 \approx u^* = a + 5(2+s)\left(\frac{\delta}{3s}\right)^{\frac{s}{2+s}}\left(\frac{b}{2(5+s)}\right)^{\frac{2}{2+s}}. \tag{4.1}$$

We identify the values of u_1 with $\underline{R}$, and u_2 with R_k. There is no mechanism in our model for slug splitting, and we cannot produce any equivalent for $\overline{R}_k$. Finally, the onset of fully developed turbulence at $\overline{R}$ necessarily must depend on the inlet

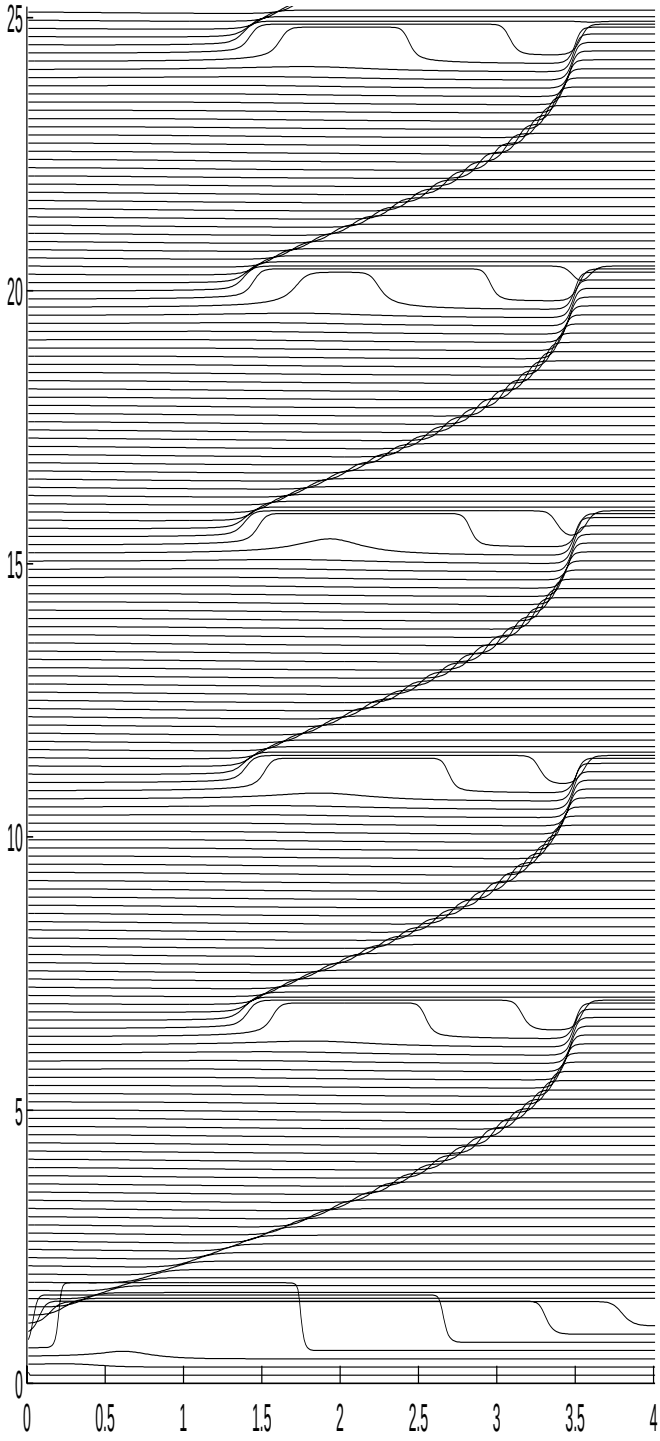


FIG. 10. *Space-time plot of $A(x,t)$, using the same parameters as in Figure 9. A sequence of slugs is generated midway down the pipe, each of which grows rapidly and then propagates down the pipe.*

disturbance amplitude level A_0 (since if this is zero, then the flow may be maintained as laminar indefinitely). In our model, this defines a third transition value u_4 (the missing u_3 would correspond to $\overline{R}_k$), such that $A_m(u_4) = A_0$, and thus

$$u_4 = a + \frac{\delta}{A_0^2} + bA_0^s. \tag{4.2}$$

This u_4 is a monotonically decreasing function of A_0 in the admissible range $0 < A < A_{\max}$, which is where $A_m = A_M$ and $u_4 = u_1$. Lindgren's apparent value of 3300 corresponds to $u_4 = 1.43$, and when A_0 is small, (3.11) applies, so that the inferred corresponding inlet disturbance amplitude is ≈ 0.02. It is also evident that, by adjusting the parameters b, a, s, and particularly δ, we could find reasonably accurate values for the predicted u_1 and u_2; such an exercise is rather alien to our present purpose, however.

The second comparison we want to make is to the amplitude-velocity transition curve in Figure 2 of Wygnanski and Champagne (1973). This figure plots inlet disturbance amplitude versus Reynolds number and delineates a laminar region from a fully turbulent region. The demarcation boundary essentially consists of two separate curves, between which the flow is intermittent. At high Re (and low A_0), slugs occur in the intermittent region, while at low Re (and high A_0), puffs occur. The slug and puff regions become "uncertain" around $Re = 2700$, corresponding in our model to the value $u = u_2$.

In our model, the critical inlet amplitude, which separates flows that can be fully laminar from those that can be fully turbulent, is $A_0 = A_m(u)$. However, whether fully developed turbulent or laminar flow can be obtained throughout the pipe depends on the applied pressure drop F^*, and it is appropriate to use this as the control variable. Denote the turbulent and laminar branches of the pressure drop curve as $F_T(u)$ and $F_L(u)$, these being given by (2.60) and $F_L = u$, respectively. Our predictions for the upper and lower transition curves in A_0-F^* space are then, for the upper curve,

$$\begin{aligned} F^* &= F_T(u), \\ A_0 &= A_m(u), \end{aligned} \tag{4.3}$$

and for the lower,

$$\begin{aligned} F^* &= F_L(u), \\ A_0 &= A_m(u). \end{aligned} \tag{4.4}$$

These curves are portrayed in Figure 11, and they show qualitative agreement with Wygnanski and Champagne's (1973) figure (bearing in mind that theirs is a log-linear plot). Indeed we have already seen (compare Figure 4 and (3.11)) that there is some quantitative resemblance also.

The fact that the very simple model used here compares so well with experimental observations is encouraging, but it must be pointed out that the guts of the process are entirely missing, that is to say, the generation of the turbulent chaotic eddies themselves. What we have shown is that some macroscopic features of the transition to turbulence in pipe flow can be understood more or less entirely through the well-founded postulate of a hysteretic transition between laminar and turbulent "states" of the fluid. Precisely what the turbulent state consists of is not addressed in this theory.

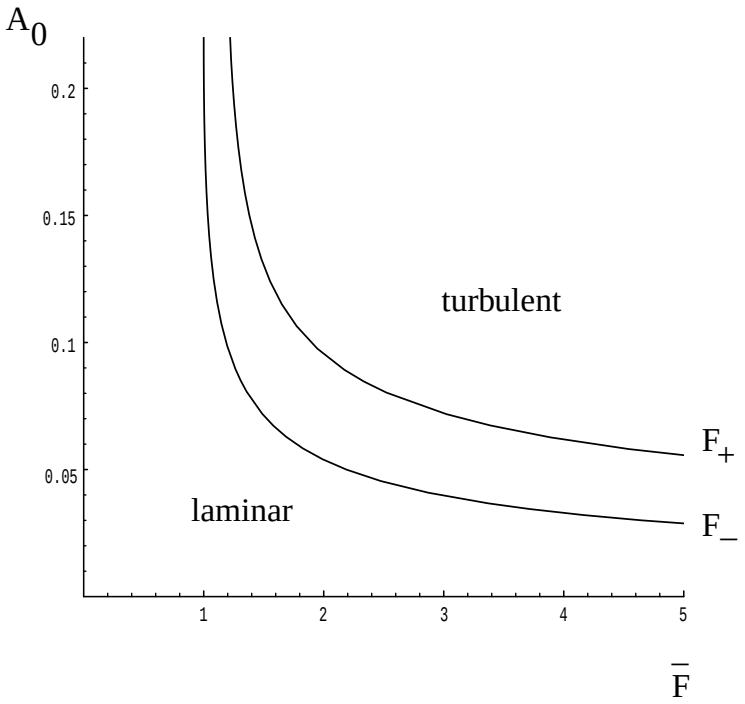


FIG. 11. *Parameter space of inlet disturbance A_0 versus applied pressure drop F^*, with $b = 0.6$, $s = 0.64$, $\delta = 0.0035$. Intermittency occurs for points between the two solid curves. Of these, the bottom is $F_- = a + bA_0^s + \delta/A_0^2$; the top is given by $F_+ = a + A_M + bA_0^s + \delta/A_0^2$, where A_M is the solution $> A_0$ of $bA_M^s + \delta/A_M^2 = bA_0^s + \delta/A_0^2$.*

REFERENCES

H. BLASIUS (1913), *Das Ähnlichkeitsgesetz bei Reibungsvorgängen in Flüssigkeiten*, Mitt. Forsch. Arb., 131, pp. 1–39.

T. BOHR AND D. A. RAND (1991), *A mechanism for localised turbulence*, Phys. D, 52, pp. 532–543.

J. M. BURGERS (1948), *A mathematical model illustrating the theory of turbulence*, Adv. Appl. Math., 1, pp. 171–199.

A. G. DARBYSHIRE AND T. MULLIN (1995), *Transition to turbulence in constant-mass-flux pipe flow*, J. Fluid Mech., 289, pp. 83–114.

A. DAVEY (1978), *On Itoh's finite amplitude stability for pipe flow*, J. Fluid Mech., 86, pp. 695–703.

A. DAVEY AND H. P. F. NGUYEN (1971), *Finite amplitude stability of pipe flow*, J. Fluid Mech., 45, pp. 701–720.

R. J. DEISSLER (1987a), *Turbulent bursts, spots and slugs in a generalised Ginzburg–Landau equation*, Phys. Lett. A, 120, pp. 334–340.

R. J. DEISSLER (1987b), *Spatially growing waves, intermittency, and convective chaos in an open-flow system*, Phys. D, 25, pp. 233–260.

P. G. DRAZIN AND W. H. REID (1981), *Hydrodynamic Stability*, Cambridge University Press, Cambridge, England.

S. GOLDSTEIN, ED. (1938), *Modern Developments in Fluid Dynamics*, Vol. 1, Clarendon Press, Oxford, England.

Y.-N. HUANG AND Y.-D. HUANG (1989), *On the transition to turbulence in pipe flow*, Phys. D, 37, pp. 153–159.

N. ITOH (1977), *Nonlinear stability of parallel flows with subcritical Reynolds numbers. Part 2. Stability of pipe Poiseuille flow to finite axisymmetric disturbance*, J. Fluid Mech., 82, pp. 469–479.

B. E. Launder, G. J. Reece, and W. Rodi (1975), *Progress in the development of a Reynolds-stress turbulence closure*, J. Fluid Mech., 68, pp. 537–566.

E. R. Lindgren (1957), *The transition process and other phenomena in viscous flow*, Ark. Fysik, Bd. 12, pp. 1–169.

J. Mathieu and J. Scott (2000), *An Introduction to Turbulent Flow*, Cambridge University Press, Cambridge, England.

J. D. Murray (1993), *Mathematical Biology*, 2nd ed., Springer-Verlag, New York.

S. A. Orszag and A. T. Patera (1980), *Subcritical transition to turbulence in plane channel flows*, Phys. Rev. Lett., 45, pp. 989–993.

S. A. Orszag and A. T. Patera (1983), *Secondary instability of wall-bounded shear flows*, J. Fluid Mech., 128, pp. 347–385.

L. Prandtl and O. G. Tietjens (1934), *Applied Hydro- and Aeromechanics*, Dover, New York.

O. Reynolds (1883), *An experimental investigation of the circumstances which determine whether the motion of water shall be direct or sinuous and of the law of resistance in parallel channels*, Proc. Roy. Soc. London A, 35, pp. 84–99.

S. Rosenblat and S. H. Davis (1979), *Bifurcation from infinity*, SIAM J. Appl. Math., 37, pp. 1–19.

J. Rotta (1956), *Experimenteller Beitrag zur Entstehung turbulenter Strömung im Rohr*, Ing.-Arch., 24, pp. 258–281.

H. Sakaguchi and H. R. Brand (1996), *Stable localized solutions of arbitrary length for the quintic Swift–Hohenberg equation*, Phys. D, 97, pp. 274–285.

H. Schlichting (1979), *Boundary-Layer Theory*, 7th ed., McGraw-Hill, New York.

F. T. Smith and R. J. Bodonyi (1982), *Amplitude dependent neutral modes in the Hagen–Poiseuille flow through a circular pipe*, Proc. Roy. Soc. London A, 384, pp. 463–489.

I. J. Wygnanski and F. H. Champagne (1973), *On transition in a pipe. Part* 1. *The origin of puffs and slugs and the flow in a turbulent slug*, J. Fluid Mech., 59, pp. 281–335.

SIAM J. APPL. MATH.
Vol. 63, No. 4, pp. 1208–1230

WHITE NOISE ANALYSIS OF COUPLED LINEAR-NONLINEAR SYSTEMS*

DUANE Q. NYKAMP†

Abstract. We present an asymptotic analysis of two coupled linear-nonlinear systems. Through measuring first and second input-output statistics of the systems in response to white noise input, one can completely characterize the systems and their coupling. The proposed model is similar to a widely used phenomenological model of neurons in response to sensory stimulation and may be used to help characterize neural circuitry in sensory brain regions.

Key words. neural networks, correlations, Weiner analysis, white noise

AMS subject classification. 92C20

PII. S0036139901397571

1. Introduction.

Most electrophysiology data from intact mammalian brains is recorded using an extracellular electrode which remains outside the neurons. When the electrode is positioned near a neuron, it can record the neuron's output events, called spikes, because spike magnitudes are sufficiently large. The internal state of a neuron, including small fluctuations in response to its inputs, cannot be measured.

When only output spikes are measurable, one cannot directly measure the effect of a connection from one neuron to another. If neuron 1 is connected to neuron 2, then an output spike of neuron 1 will perturb the internal state of neuron 2. If the internal state cannot be measured, this perturbation can be inferred only via its effect on the spike times of neuron 2. In general, the spike times of a neuron will be a function of many inputs coming from many other neurons. This complexity makes reliable inferences on the structure of neuronal circuits from spike time data a formidable challenge.

Explicit mathematical models may lead to tools that can address this challenge. Through model analysis, one may develop methods to infer aspects of network structure from spike times, subject to the validity of the underlying model. In this paper, we derive a method for reconstructing the connectivity between two isolated neurons based on a simple linear-nonlinear model (see below) of neural response to white noise. Although this model greatly simplifies the reality of the brain's neural networks, the results from this analysis can be used to analyze neurophysiology data, provided that they are interpreted within the limitations of the model [12].

Numerous researchers have used white noise analysis to describe the response of neurons to a stimulus. The most common use of white noise analysis has been to analyze the response properties of single neurons [11, 4, 5, 8, 9, 2, 3, 16, 18, 7, 6]. Recently, researchers have begun to apply the techniques of white noise analysis to simultaneous measurements of multiple neurons [15, 1, 19], although without explicitly modeling neural connectivity. In [12], we showed how, in white noise experiments, interpretation of spike time data is especially difficult because standard correlation

*Received by the editors November 6, 2001; accepted for publication (in revised form) September 19, 2002; published electronically April 9, 2003. This research was supported by an NSF Mathematical Sciences Postdoctoral Research Fellowship.

http://www.siam.org/journals/siap/63-4/39757.html

†Department of Mathematics, University of California, Los Angeles, CA 90095 (nykamp@math.ucla.edu).

measures confound stimulus and connectivity effects. We demonstrated correlation measures that remove the stimulus effects based on the linear-nonlinear model.

In this paper, we present the asymptotic analysis of the linear-nonlinear model that underlies the correlation measures of [12]. Subject to a first order approximation in the coupling magnitude, we derive a method to completely reconstruct the coupled system from first and second input-output statistics. As a consequence of this reconstruction, we obtain a correlation measure, which we call $\mathcal{W}$, that approximates the neuronal coupling.

Although the analysis below can be used for any pair of coupled linear-nonlinear systems, we refer to the systems as *neurons* since neuroscience was the motivation for this analysis and because this choice simplifies the description.

In section 2, we describe the linear-nonlinear model. In section 3, we derive expressions for the input-output statistics for the case when the neurons are uncoupled. We derive the corresponding expressions for the cases of unidirectional coupling in section 4 and generalize the results to mutual coupling in section 5. We demonstrate the method with simulations in section 6 and discuss the results in section 7.

2. The model. The standard model underlying most white noise analyses of neural function is the linear-nonlinear model of neural response to an input $\mathbf{X}$,

$$\Pr\big(R^i = 1 \big| \mathbf{X} = \mathbf{x}\big) = g(\mathbf{h}^i \cdot \mathbf{x}), \tag{2.1}$$

where the response R^i at discrete time point i is one if the neuron spiked, and zero otherwise. The neural response depends on the convolution of the kernel $\mathbf{h}$ with the stimulus. The stimulus $\mathbf{X}$ is a vector whose components represent the spatio-temporal sequence of stimulus values, such as the sequence of pixel values for each refresh of a computer monitor.

The neural response depends on the convolution of a stimulus with a kernel $\mathbf{h}$, normalized so that $|\mathbf{h}| = 1$. The kernel can be viewed as sliding along the stimulus with time; it represents the spatio-temporal stimulus features to which the neuron responds. We let $\mathbf{h}^i$ denote the kernel shifted for time point i and write the convolution of the kernel with the stimulus as the dot product $\mathbf{h}^i \cdot \mathbf{X}$ (implicitly viewing the temporal index of the stimulus as going backward in time). The function $g(\cdot)$ is the neuron's output nonlinearity (representing, for example, its spike generating mechanism). Although the linear-nonlinear system is only a phenomenological approximation of complex biology, it can be simply characterized by standard white noise analysis [13]. The ease of an explicit mathematical analysis is a prime motivation for choosing the linear-nonlinear model and white noise input.

We propose a model that augments the linear-nonlinear framework to include the effects of neural connections between two neurons. After neuron q spikes, the probability that neuron p spikes j time steps later is modified by the connectivity factor $\bar{W}^j_{qp}$. In a caricature of synaptic input acting at subthreshold levels (of the voltage, or internal state, of a neuron), the term $\bar{W}^j_{qp}$ is added underneath the nonlinearity so that

$$\Pr\big(R^i_p = 1 \big| \mathbf{X} = \mathbf{x}, \mathbf{R}_q = \mathbf{r}_q\big) = g_p\bigg(\mathbf{h}^i_p \cdot \mathbf{x} + \sum_{j \geq 0} \bar{W}^j_{qp} r^{i-j}_q\bigg), \tag{2.2}$$

where $p, q \in \{1, 2\}$ represent the index of the neurons, $q \neq p$, and $R^i_p \in \{0, 1\}$ is the response of neuron p at time i.[1]

[1] With the exceptions of W and T, we will use capital variables to denote random quantities. In addition, we will use subscripts to denote neuron index, and superscripts to denote temporal indices.

We assume that the output nonlinearity can be approximated as an error function

$$g_p(s) = \frac{\hat{r}_p}{2}\left[1 + \operatorname{erf}\left(\frac{s - \bar{T}_p}{\epsilon_p\sqrt{2}}\right)\right], \tag{2.3}$$

where $\hat{r}_p$ is the maximum firing rate, $\bar{T}_p$ is the threshold, ϵ_p defines the steepness of the nonlinearity, and $\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$. Note that $\lim_{x\to\infty} g_p(x) = \hat{r}_p$ and $\lim_{x\to-\infty} g_p(x) = 0$. The error function nonlinearity is assumed so that we can derive analytic results. As demonstrated in section 6, the results apply to more general nonlinearities.

So that the input $\mathbf{X}$ is a discrete approximation of temporal or spatio-temporal white noise, we let each of its n components be standard normal random variables. We do not explicitly distinguish spatial versus temporal components of the input in our notation because they are treated identically in the analysis. To keep the notation simple, time is represented only by the temporal index of the kernels $\mathbf{h}_p^i$ and the spikes R_p^i. With this convention, the probability density function of $\mathbf{X}$ is simply

$$\rho_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}} e^{-|\mathbf{x}|^2/2}. \tag{2.4}$$

In the next sections, we consider special cases of the coupling $\bar{W}$. For each case, we calculate the expected values of the responses $E\{R_p^i\}$, the "correlation"[2] between the stimulus and the spikes of each neuron $E\{\mathbf{X}R_p^i\}$, and the "correlation" between the spikes of the two neurons $E\{R_1^i R_2^{i-k}\}$. Since these statistics can be estimated when one can obtain only the spike times from the neurons, they are readily measurable in neurophysiology experiments. We base our reconstruction of the linear-nonlinear system of (2.2) on these input-output statistics. Most importantly, we will reconstruct the coupling terms $\bar{W}_{pq}^j$.

3. Uncoupled neurons. In this section, we assume that the neurons are uncoupled so that the responses of neurons are independent conditioned on the input.[3] In this case, the response probabilities obey (2.2) and (2.3) with $\bar{W}_{pq}^j = 0$.

The analysis of the single neuron statistics reduces to the case of individual neurons. As detailed in [13], the first two input-output statistics are given by

$$E\{R_p^i\} = \frac{\hat{r}_p}{2}\operatorname{erfc}\left(\frac{\delta_p \bar{T}_p}{\sqrt{2}}\right) \tag{3.1}$$

and

$$E\{\mathbf{X}R_p^i\} = \frac{\delta_p}{\sqrt{2\pi}} e^{-\frac{\delta_p^2 \bar{T}_p^2}{2}} \mathbf{h}_p^i, \tag{3.2}$$

where

$$\delta_p = \frac{1}{\sqrt{1+\epsilon_p^2}} \tag{3.3}$$

[2] We recognize that the statistics $E\{\mathbf{X}R_p^i\}$ and $E\{R_1^i R_2^{i-k}\}$ are not actually correlations. We use the term since these statistics are consistently called correlations in the neuroscience literature. We hope the reader will forgive our loose use of the term. The stimulus-spike correlation $E\{\mathbf{X}R_p^i\}$ can be thought of as the average stimulus that precedes each spike of neuron p.

[3] Meaning $\Pr(R_1^i = 1 \,\&\, R_2^j = 1|\mathbf{X}) = \Pr(R_1^i = 1|\mathbf{X})\Pr(R_2^j = 1|\mathbf{X})$.

and $\operatorname{erfc}(x) = 1 - \operatorname{erf}(x)$. Note that, since both the input and the system are stationary, the results are independent of time index i (except for the temporal index of the linear kernel). Assuming that one knew $\hat{r}_p$, the nonlinearities $g_p(\cdot)$ could be computed by estimating ϵ_p and $\bar{T}_p$ from these statistics [13]. One could also obtain the unit vectors $\mathbf{h}_p^i$ from $E\{\mathbf{X}R_p^i\}/|E\{\mathbf{X}R_p^i\}|$. In this simple case, one does not even need to measure $E\{R_1^i R_2^{i-k}\}$ to reconstruct the system.

Before we calculate an expression for $E\{R_1^i R_2^{i-k}\}$, we define the angles between the linear kernels, which turn out to be the only important geometry of the kernels for white noise input. Let $\bar{\theta}_{pq}^k$ be the angle between kernel q and kernel p shifted k units in time

$$\cos\bar{\theta}_{pq}^k = \mathbf{h}_p^{i-k} \cdot \mathbf{h}_q^i. \tag{3.4}$$

This angle is of course independent of time index i. (The inner product can be represented as a cosine because kernels were normalized to be unit vectors.) Note that $\bar{\theta}_{qp}^{-k} = \bar{\theta}_{pq}^k$. We always define the corresponding sine by $\sin\theta = \sqrt{1-\cos^2\theta}$ so that $\sin\theta \geq 0$.

For a given time shift k, without loss of generality, assume that $\mathbf{h}_1^i$ is the first unit vector $\mathbf{e}_1$ (in stimulus space) and $\mathbf{h}_2^{i-k}$ is a linear combination of the first two unit vectors:[4]

$$\begin{aligned}\mathbf{h}_1^i &= \mathbf{e}_1,\\ \mathbf{h}_2^{i-k} &= \mathbf{e}_1\cos\bar{\theta}_{21}^k + \mathbf{e}_2\sin\bar{\theta}_{21}^k.\end{aligned}$$

Assuming that the nonlinearities satisfy

$$\lim_{x\to-\infty} g_p(x) = 0, \tag{3.5}$$

we compute the correlation between the spikes of neuron 1 and the spikes of neuron 2 by changing variables and integrating by parts twice. In each integration by parts, one boundary term disappears due to (3.5), and the other boundary term is incorporated into the complementary error functions:

$$\begin{aligned}E\{R_1^i R_2^{i-k}\} &= \frac{1}{2\pi}\int g_1(x_1)g_2(x_1\cos\bar{\theta}_{21}^k + x_2\sin\bar{\theta}_{21}^k)e^{-\frac{x_1^2+x_2^2}{2}}\,dx_1\,dx_2\\ &= \frac{1}{2\pi}\int g_1(u)g_2(v)\exp\left(-\frac{u^2}{2} - \frac{(v-u\cos\bar{\theta}_{21}^k)^2}{2\sin^2\bar{\theta}_{21}^k}\right)\frac{du\,dv}{\sin\bar{\theta}_{21}^k}\\ &= \frac{1}{4}\int g_1'(u)g_2'(v)\operatorname{derfc}\left(\frac{u}{\sqrt{2}},\frac{v}{\sqrt{2}},\cos\bar{\theta}_{21}^k\right)du\,dv,\end{aligned} \tag{3.6}$$

where we have defined a double complementary error function

$$\operatorname{derfc}(a,b,c) = \frac{4}{\pi}\int_a^\infty dy\,e^{-y^2}\int_{\frac{b-cy}{\sqrt{1-c^2}}}^\infty dz\,e^{-z^2}. \tag{3.7}$$

The function derfc is a two dimensional analogue of the complementary error function. The integral is taken over the intersection of the two half-planes $\mathbf{x}\cdot\mathbf{u} > a$

[4]Since the stimulus is rotationally invariant, we can rotate the axis so that $\mathbf{h}_1^i$ is parallel to the first axis and $\mathbf{h}_2^{i-k}$ lies in the span of the first two axes. Recall that $|\mathbf{h}_p^i| = 1$.

and $\mathbf{x} \cdot \mathbf{v} > b$, where $\mathbf{u}$ and $\mathbf{v}$ are two unit vectors with inner product $\mathbf{u} \cdot \mathbf{v} = c$. (Here, $\mathbf{x}$ represents a generic vector.) Note that $\operatorname{derfc}(a, b, 0) = \operatorname{erfc}(a)\operatorname{erfc}(b)$ and $\operatorname{derfc}(a, b, c) = \operatorname{derfc}(b, a, c)$.

When the nonlinearity is an error function (i.e., (2.3)), we substitute into (3.6) and use formula (B.8) to obtain

$$E\{R_1^i R_2^{i-k}\} = \frac{\hat{r}_1 \hat{r}_2}{4} \operatorname{derfc}\left(\frac{\delta_1 \bar{T}_1}{\sqrt{2}}, \frac{\delta_2 \bar{T}_2}{\sqrt{2}}, \delta_1 \delta_2 \cos \bar{\theta}_{21}^k\right). \tag{3.8}$$

Equations (3.1), (3.2), and (3.8) are the expressions for the input-output statistics for the simple case of uncoupled neurons.

4. Unidirectional coupling. Let the coupling from neuron 2 to neuron 1 ($\bar{W}_{21}^j$) be nonzero, but keep $\bar{W}_{12}^j = 0$. Then the probability of a spike of neuron 1 at time k is dependent not only on the input but also on the spikes of neuron 2 for times before k, as given by (2.2). The probability of neuron 2's spiking remains the same as in section 3, and thus the input-output statistics $E\{R_2^i\}$ and $E\{\mathbf{X}R_2^i\}$ are unchanged.

In what follows, we calculate expressions for the remaining input-output statistics. We first show that effective parameters of the system can be calculated from $E\{R_p^i\}$ and $E\{\mathbf{X}R_p^i\}$. We next show how the coupling $\bar{W}_{21}^j$ can be calculated from $E\{R_1^i R_2^{i-k}\}$.

We assume that $\bar{W}_{21}^j$ is small and compute a first order approximation by dropping terms that are second order or higher in $\bar{W}_{21}^j$. Since from now on all equalities will be within $O(\bar{W}^2)$, we will, for simplicity, use $=$ to mean equal within $O(\bar{W}^2)$.

4.1. Mean rate of neuron 1. In this section, we show that the mean rate of neuron 1 is nearly identical to the uncoupled case of (3.1), only with the original threshold $\bar{T}_1$ replaced with an effective threshold T_1 to be defined below.

The general expression for the mean rate of neuron 1, calculated in Appendix A.3, is

$$E\{R_1^i\} = \frac{1}{\sqrt{2\pi}} \int g_1(u) e^{-\frac{u^2}{2}} du + \sum_{j \geq 0} \frac{\bar{W}_{21}^j}{2\sqrt{2\pi}} \int g_1'(u) g_2'(v) e^{-\frac{u^2}{2}} \operatorname{erfc}\left(\frac{v - u \cos \bar{\theta}_{21}^k}{\sqrt{2} \sin \bar{\theta}_{21}^k}\right) du\, dv. \tag{4.1}$$

Note that the mean rate $E\{R_1^i\}$ is independent of i (as it must be).

When the nonlinearities are error functions (e.g., (2.3)), the first term is the uncoupled mean rate (i.e., (3.1)). We use formula (B.5) to simplify the $\bar{W}_{21}^j$ term so that the mean rate of neuron 1 is

$$E\{R_1^i\} = \frac{\hat{r}_1}{2} \operatorname{erfc}\left(\frac{\delta_1 \bar{T}_1}{\sqrt{2}}\right) + \frac{\hat{r}_1 \hat{r}_2 \delta_1}{2\sqrt{2\pi}} e^{-\frac{\delta_1^2 \bar{T}_1^2}{2}} \sum_{j \geq 0} \bar{W}_{21}^j \operatorname{erfc}\left(\frac{\delta_2 \bar{T}_2 - \delta_1^2 \delta_2 \bar{T}_1 \cos \bar{\theta}_{21}^j}{\sqrt{1(1 - \delta_1^2 \delta_2^2 \cos^2 \bar{\theta}_{21}^j)}}\right).$$

Using the Taylor series for $\operatorname{erfc}(\frac{\delta_1 \bar{T}_1 + x}{\sqrt{2}})$, we pull the second term into the error function (making only an $O(\bar{W}^2)$ error), obtaining

$$E\{R_1^i\} = \frac{\hat{r}_1}{2} \operatorname{erfc}\left(\frac{\delta_1 T_1}{\sqrt{2}}\right), \tag{4.2}$$

where we let $T_2 = \bar{T}_2$ and have defined the effective threshold for neuron 1:

$$T_1 = \bar{T}_1 - \sum_{j\geq 0} \frac{\hat{r}_2 \bar{W}_{21}^j}{2} \operatorname{erfc}\left(\frac{\delta_2 T_2 - \delta_1^2 \delta_2 \bar{T}_1 \cos\bar{\theta}_{21}^j}{\sqrt{2(1-\delta_1^2\delta_2^2\cos^2\bar{\theta}_{21}^j)}} \right). \tag{4.3}$$

The mean rate of neuron 1 is identical to that of an uncoupled neuron with the effective threshold T_1.

4.2. Correlation of spikes of neuron 1 with the stimulus. We calculate the general expression for the correlation between the spikes of neuron 1 with the stimulus in Appendix A.4, obtaining

$$\begin{aligned} E\{\mathbf{X}R_1^i\} = \frac{1}{\sqrt{2\pi}} \Bigg[& \int g_1'(u) e^{-\frac{u^2}{2}} du \\ & + \sum_{j\geq 0} \frac{\bar{W}_{21}^j}{2} \int g_1'(u) g_2'(v) u e^{-\frac{u^2}{2}} \operatorname{erfc}\left(\frac{v - u\cos\bar{\theta}_{21}^k}{\sqrt{2}\sin\bar{\theta}_{21}^k} \right) du\, dv \Bigg] \mathbf{h}_1^i \\ & + \sum_{j\geq 0} \frac{\bar{W}_{21}^j}{2\pi} \int g_1'(u) g_2'(v) \exp\left(- \frac{u^2 - 2\cos\bar{\theta}_{21}^j uv + v^2}{2\sin^2\bar{\theta}_{21}^j} \right) du\, dv\, \overset{\perp ji}{\mathbf{h}}_{21}, \end{aligned} \tag{4.4}$$

where we define $\overset{\perp ji}{\mathbf{h}}_{21}$ as the component of $\mathbf{h}_2^{i-j}$ that is perpendicular to $\mathbf{h}_1^i$,

$$\overset{\perp ji}{\mathbf{h}}_{21} = \frac{\mathbf{h}_2^{i-j} - \cos\bar{\theta}_{21}^j \mathbf{h}_1^i}{\sin\bar{\theta}_{21}^j}. \tag{4.5}$$

Because of the coupling, $E\{\mathbf{X}R_1^i\}$ is no longer parallel to the linear kernel $\mathbf{h}_1^i$. Each term in the last sum of (4.4) indicates how the coupling $\bar{W}_{21}^j$ leads to a component of $E\{\mathbf{X}R_1^i\}$ that is perpendicular to $\mathbf{h}_1^i$.

When the nonlinearities are error functions (e.g., (2.3)), we use (4.5) and formulas (B.1), (B.6), and (B.2) to obtain the following expression for the correlation between the stimulus and the spikes of neuron 1:

$$\begin{aligned} E\{\mathbf{X}R_1^i\} = \mu_1^0 & \left[1 - \sum_{j\geq 0} \frac{\hat{r}_2 \bar{W}_{21}^j \delta_1^2 \delta_2 \cos\bar{\theta}_{21}^j}{\sqrt{2\pi(1-\delta_1^2\delta_2^2\cos^2\bar{\theta}_{21}^j)}} \exp\left(-\frac{[\delta_2 T_2 - \delta_1^2\delta_2 T_1 \cos\bar{\theta}_{21}^j]^2}{2(1-\delta_1^2\delta_2^2\cos^2\bar{\theta}_{21}^j)} \right) \right] \mathbf{h}_1^i \\ & + \mu_1^0 \sum_{j\geq 0} \frac{\hat{r}_2 \bar{W}_{21}^j \delta_2}{\sqrt{2\pi(1-\delta_1^2\delta_2^2\cos^2\bar{\theta}_{21}^j)}} \exp\left(-\frac{[\delta_2 T_2 - \delta_1^2\delta_2 T_1 \cos\bar{\theta}_{21}^j]^2}{2(1-\delta_1^2\delta_2^2\cos^2\bar{\theta}_{21}^j)} \right) \mathbf{h}_2^{i-j}, \end{aligned} \tag{4.6}$$

where

$$\mu_p^0 = \frac{\hat{r}_p \delta_p}{\sqrt{2\pi}} e^{-\frac{\delta_p^2 T_p^2}{2}}. \tag{4.7}$$

One key to obtaining (4.6) was using the exponential's Taylor series to bring the effective threshold T_1 (4.3) into the exponential of the first term. We let $T_2 = \bar{T}_2$ and simply replaced $\bar{T}_1$ with T_1 in all other terms (making an $O(\bar{W}^2)$ error).

4.3. Reconstruction from the mean rate and stimulus-spike correlations. Equations (4.2) and (4.6) give expressions for the first two input-output statistics of the linear-nonlinear system (2.2) with unidirectional coupling. These equations

show that the coupling has both changed the effective threshold and altered the direction of $E\{\mathbf{X}R_1^i\}$ so that it is no longer parallel to the kernel $\mathbf{h}_1^i$.

Because of these modifications, we can no longer recover $\bar{T}_1$ or $\mathbf{h}_1^i$ (or $\cos\bar{\theta}_{21}^j$) from $E\{R_1^i\}$ and $E\{\mathbf{X}R_1^i\}$ as outlined in section 3. Nonetheless, subject to one more assumption, one can recover the effective threshold T_1, the original δ_1, and an effective angle between the kernels. As shown below, one simply views the neurons as uncoupled and reconstructs the neuron parameters as in section 3. This procedure does not use $E\{R_1^iR_2^{i-k}\}$. We will be able use this last statistic to determine the coupling $\bar{W}_{21}^j$.

4.3.1. Effective angle between kernels. When the neurons were uncoupled, the linear kernel $\mathbf{h}_1^i$ could be determined by the normalized stimulus-spike correlation $E\{\mathbf{X}R_1^i\}/|E\{\mathbf{X}R_1^i\}|$. Although this measurement no longer yields the kernel, we can treat it as an effective kernel and define the effective angle between kernels by

$$\cos\theta_{pq}^k = \frac{E\{\mathbf{X}R_p^{i-k}\}}{\left|E\{\mathbf{X}R_p^{i-k}\}\right|}\cdot\frac{E\{\mathbf{X}R_q^i\}}{\left|E\{\mathbf{X}R_q^i\}\right|}. \tag{4.8}$$

In this case of unidirectional coupling, neuron 2 is unaffected, and the effective angle between neurons 1 and 2 is $\cos\theta_{21}^k = \mathbf{h}_2^{i-k}\cdot E\{\mathbf{X}R_1^i\}/|E\{\mathbf{X}R_1^i\}|$.

We rewrite (4.2) and (4.6) in terms of the measurable effective angle as follows. The magnitude of the stimulus-spike correlation, within $O(\bar{W}^2)$, is

$$\left|E\{\mathbf{X}R_1^i\}\right| = \mu_1^0\left[1+\sum_{j\geq 0}\frac{\hat{r}_2\bar{W}_{21}^j(1-\delta_1^2)\delta_2\cos\bar{\theta}_{21}^j}{\sqrt{2\pi(1-\delta_1^2\delta_2^2\cos^2\bar{\theta}_{21}^j)}}\exp\left(-\frac{[\delta_2T_2-\delta_1^2\delta_2T_1\cos\bar{\theta}_{21}^j]^2}{2(1-\delta_1^2\delta_2^2\cos^2\bar{\theta}_{21}^j)}\right)\right] \tag{4.9}$$

so that

$$\cos\theta_{21}^k = \cos\bar{\theta}_{21}^k + \sum_{j\geq 0}\frac{\hat{r}_2\bar{W}_{21}^j\delta_2(\cos\theta_{22}^{k-j}-\cos\bar{\theta}_{21}^j\cos\bar{\theta}_{21}^k)}{\sqrt{2\pi(1-\delta_1^2\delta_2^2\cos^2\bar{\theta}_{21}^j)}}\exp\left(-\frac{[\delta_2T_2-\delta_1^2\delta_2T_1\cos\bar{\theta}_{21}^j]^2}{2(1-\delta_1^2\delta_2^2\cos^2\bar{\theta}_{21}^j)}\right).$$

Since $\cos\theta_{21}^k$ is within $O(\bar{W})$ of $\cos\bar{\theta}_{21}^k$, we can replace $\cos\bar{\theta}_{21}^k$ by $\cos\theta_{21}^k$ in the last terms (making only an $O(\bar{W}^2)$ error), and write $\cos\bar{\theta}_{21}^k$ in terms of $\cos\theta_{21}^k$:

$$\cos\bar{\theta}_{21}^k = \cos\theta_{21}^k - \sum_{j\geq 0}\bar{W}_{21}^jC_{21}^{jk}, \tag{4.10}$$

where

$$C_{pq}^{jk} = (\cos\theta_{pp}^{k-j} - \cos\theta_{pq}^j\cos\theta_{pq}^k)\mu_{pq}^j, \tag{4.11}$$

$$\mu_{pq}^k = \frac{\hat{r}_p\delta_p\exp\left(-\frac{1}{2}[\lambda_{pq}^k]^2\right)}{\sqrt{2\pi(1-\delta_p^2\delta_q^2\cos^2\theta_{pq}^k)}}, \tag{4.12}$$

and

$$\lambda_{pq}^k = \frac{\delta_pT_p-\delta_p\delta_q^2T_q\cos\theta_{pq}^k}{\sqrt{1-\delta_p^2\delta_q^2\cos^2\theta_{pq}^k}}. \tag{4.13}$$

Note that $\mu_p^0\mu_{pq}^k = \mu_q^0\mu_{qp}^{-k}$.

We now make an $O(\bar{W}^2)$ error by replacing $\cos\bar{\theta}_{21}^k$ with $\cos\theta_{21}^k$ in the stimulus-spike correlation

$$E\{\mathbf{X}R_1^i\} = \mu_1^0\left[\left(1 - \sum_{j\geq 0}\bar{W}_{21}^j\delta_1^2\cos\theta_{21}^j\mu_{21}^j\right)\mathbf{h}_1^i + \sum_{j\geq 0}\bar{W}_{21}^j\mu_{21}^j\mathbf{h}_2^{i-j}\right] \tag{4.14}$$

and in the expression for the effective threshold (4.3),

$$\bar{T}_1 = T_1 + \sum_{j\geq 0}\bar{W}_{21}^j\eta_{21}^j, \tag{4.15}$$

where

$$\eta_{pq}^k = \frac{\hat{r}_p}{2}\operatorname{erfc}\left(\frac{\lambda_{pq}^k}{\sqrt{2}}\right). \tag{4.16}$$

4.3.2. Effective nonlinearity parameters. As shown above, $\cos\theta_{21}^k$, not the original $\cos\bar{\theta}_{21}^k$, is the measurable inner product between the kernels. We next show that, with one additional mild assumption, the parameters T_1 and δ_1 are the nonlinearity parameters measured from $E\{\mathbf{X}R_1^i\}$ and $E\{R_1^i\}$ when treating neuron 1 as an independent neuron as in section 3.

The magnitude of $E\{\mathbf{X}R_1^i\}$ is

$$\left|E\{\mathbf{X}R_1^i\}\right| = \mu_1^0\left(1 + \sum_{j\geq 0}\bar{W}_{21}^j(1-\delta_1^2)\cos\theta_{21}^j\mu_{21}^j\right). \tag{4.17}$$

This expression simplifies to μ_1^0 if we assume that $\bar{W}_{21}^j\delta_1\delta_2(1-\delta_1^2)\cos\theta_{21}^j$ is small enough to ignore. Since we have already assumed that $\bar{W}_{21}^j$ is small, we simply need to assume that $\delta_2(1-\delta_1^2)\cos\theta_{21}^j$ is small to have an expression that is second order in a small parameter. This expression is the product of three factors that are each less than one. It will be small if the nonlinearities are not very sharp or if the kernels of the neurons are not nearly aligned.

With this approximation, the stimulus-spike correlation is

$$\left|E\{\mathbf{X}R_1^i\}\right| \approx \mu_1^0 = \frac{\hat{r}_1\delta_1}{\sqrt{2\pi}}e^{-\frac{\delta_1^2T_1^2}{2}}. \tag{4.18}$$

Recall that the mean rate of neuron 1 (see (4.2)) is

$$E\{R_1^i\} = \frac{\hat{r}_1}{2}\operatorname{erfc}\left(\frac{\delta_1T_1}{\sqrt{2}}\right).$$

These results are the same as (3.1) and (3.2) for an uncoupled neuron with nonlinearity parameters δ_1 and T_1. One can determine δ_1 and T_1 from these equations (assuming that $\hat{r}_1$ is known).

Using only $E\{R_p^i\}$ and $E\{\mathbf{X}R_p^i\}$ in this manner, one can calculate effective nonlinearity parameters of both neuron 1 and neuron 2, as well as the effective angle between the linear kernels. We next show how the connectivity $\bar{W}_{21}$ can be determined from the remaining input-output statistic $E\{R_1^iR_2^{i-k}\}$.

4.4. Correlation between spikes of neurons 1 and 2. We calculate the general expression for the correlation between the spikes of neurons 1 and 2 in Appendix A.5, obtaining the complicated expression

$$\begin{aligned} E\{R_1^i R_2^{i-k}\} &= \frac{1}{4}\int g_1'(u_1)g_2'(u_2)\,\mathrm{derfc}\left(\frac{u_1}{\sqrt{2}},\frac{u_2}{\sqrt{2}},\cos\bar\theta_{21}^k\right)du_1du_2 \\ &\quad+\frac{\bar W_{21}^k}{2\sqrt{2\pi}}\int g_1'(u_1)g_2'(u_2)e^{-\frac{u_1^2}{2}}\,\mathrm{erfc}\left(\frac{u_2-u_1\cos\bar\theta_{21}^k}{\sqrt{2}\sin\bar\theta_{21}^k}\right)du_1du_2 \\ &\quad+\sum_{j\ge0,\,j\ne k}\frac{\bar W_{21}^j}{4\sqrt{2\pi}}\int du_1du_2du_3\,g_1'(u_1)g_2'(u_2)g_2'(u_3)e^{-\frac{u_1^2}{2}} \\ &\quad\times\mathrm{derfc}\left(\frac{u_2-u_1\cos\bar\theta_{21}^k}{\sqrt{2}\sin\bar\theta_{21}^k},\frac{u_3-u_1\cos\bar\theta_{21}^j}{\sqrt{2}\sin\bar\theta_{21}^j},\frac{\cos\theta_{22}^{k-j}-\cos\bar\theta_{21}^j\cos\bar\theta_{21}^k}{\sin\bar\theta_{21}^j\sin\bar\theta_{21}^k}\right). \end{aligned} \tag{4.19}$$

When the nonlinearities are error functions (e.g., (2.3)), we simplify this expression using three formulas ((B.8), (B.9), and (B.5)) and (4.10), (4.15), (4.7), (4.16), and (4.11). We use the following Taylor series expansions of $\mathrm{derfc}(a,b,c)$,

$$\begin{aligned} \mathrm{derfc}(a+x,b,c) &= \mathrm{derfc}(a,b,c)-\frac{2x}{\sqrt{\pi}}e^{-a^2}\,\mathrm{erfc}\left(\frac{b-ca}{\sqrt{1-c^2}}\right)+O(x^2), \\ \mathrm{derfc}(a,b,c+x) &= \mathrm{derfc}(a,b,c)+\frac{2x}{\pi\sqrt{1-c^2}}e^{-\frac{a^2-2abc+b^2}{1-c^2}}+O(x^2), \end{aligned}$$

to pull terms for the effective threshold T_1 and effective kernel inner product $\cos\theta_{21}^j$ into the first term. All other terms are $O(\bar W)$, so we can simply drop the bars from $\bar T_1$ and $\cos\bar\theta_{21}^j$.

Defining

$$\nu_{pq}^k=\frac{\hat r_p\hat r_q}{4}\,\mathrm{derfc}\left(\frac{\delta_pT_p}{\sqrt{2}},\frac{\delta_qT_q}{\sqrt{2}},\delta_p\delta_q\cos\theta_{pq}^k\right), \tag{4.20}$$

$$\tilde\nu_{pq}^{kj}=\begin{cases}\eta_{pq}^k & \text{for } j=k,\\ \frac{(\hat r_p)^2}{4}\,\mathrm{derfc}\left(\frac{\lambda_{pq}^k}{\sqrt{2}},\frac{\lambda_{pq}^j}{\sqrt{2}},\xi_{pq}^{kj}\right) & \text{otherwise,}\end{cases} \tag{4.21}$$

and

$$\xi_{pq}^{kj}=\frac{\delta_p^2\cos\theta_{pp}^{k-j}-\delta_p^2\delta_q^2\cos\theta_{pq}^j\cos\theta_{pq}^k}{\sqrt{(1-\delta_p^2\delta_q^2\cos^2\theta_{pq}^j)(1-\delta_p^2\delta_q^2\cos^2\theta_{pq}^k)}}, \tag{4.22}$$

the correlation between the spikes of neurons 1 and 2 becomes

$$E\{R_1^iR_2^{i-k}\}=\nu_{21}^k+\sum_{j\ge0}A_{21}^{kj}\bar W_{21}^j, \tag{4.23}$$

where

$$A_{pq}^{kj}=\mu_q^0\big[\tilde\nu_{pq}^{kj}-\eta_{pq}^k\eta_{pq}^j+(\cos\theta_{pq}^k\cos\theta_{pq}^j-\cos\theta_{pp}^{k-j})\mu_{pq}^k\mu_{pq}^j\big] \tag{4.24}$$

and μ_p^0, μ_{pq}^k, λ_{pq}^k, and η_{pq}^k are defined in (4.7), (4.12), (4.13), and (4.16), respectively. The term ν_{21}^k in (4.23) is analogous to the correlation observed in the uncoupled case (3.8), and the sum represents additional correlations due to the coupling terms $\bar W_{21}^j$. A discussion of some properties of A_{pq}^{kj} is given in the next section.

The important fact to note about (4.23) is that, with the exception of the $\bar{W}_{21}^j$, every factor on the right-hand side can be calculated from the mean rates $E\{R_p^i\}$ and stimulus-spike correlations $E\{\mathbf{X}R_p^i\}$. Equation (4.23) can then be solved to determine the $\bar{W}_{21}^j$.

5. Mutual coupling. Let both $\bar{W}_{21}^j$ and $\bar{W}_{12}^j$ be nonzero so that the neurons are mutually coupled. Then the probability of a spike of neuron p at time k depends not only on the input but also on the spikes of neuron q for times before k, as given by (2.2).

Since we assume that $\bar{W}_{pq}^j$ is small and compute a first order approximation, the mutual interaction results are identical to the unidirectional results of section 4 applied in both directions. The effect of neuron p on neuron q is $O(\bar{W})$, so the effect of neuron p on itself through neuron q is $O(\bar{W}^2)$ and can be ignored. We can ignore second (and higher) order interactions.

The mutual coupling case involves no more work beyond that for the unidirectional case. The statistics for neuron 1 are unchanged, and the statistics for neuron 2 become analogous to those for neuron 1. The expression for $E\{R_1^i R_2^{i-k}\}$ simply adds a sum in terms of the $\bar{W}_{12}^j$ coupling.

We summarize the model and resulting equations. We are given the system

$$\Pr\bigl(R_p^i = 1 \bigm| \mathbf{X} = \mathbf{x}, \mathbf{R}_q = \mathbf{r}_q\bigr) = g_p\biggl(\mathbf{h}_p^i \cdot \mathbf{x} + \sum_{j\geq 0} \bar{W}_{qp}^j r_q^{i-j}\biggr) \tag{5.1}$$

for $p, q \in \{1, 2\}$, $q \neq p$, with

$$g_p(x) = \frac{\hat{r}_p}{2}\biggl[1 + \operatorname{erf}\biggl(\frac{x - \bar{T}_p}{\epsilon_p\sqrt{2}}\biggr)\biggr]. \tag{5.2}$$

We assume we know $\hat{r}_p$. We can reconstruct the system from the following input-output statistics: $E\{R_p^i\}$, $E\{\mathbf{X}R_p^i\}$, and $E\{R_1^i R_2^{i-k}\}$.

First, we calculate $\delta_p = 1/\sqrt{1+\epsilon_p^2}$ and an effective threshold T_p from $E\{R_p^i\}$ and $|E\{\mathbf{X}R_p^i\}|$ using the equations[5]

$$E\{R_p^i\} = \frac{\hat{r}_p}{2}\operatorname{erfc}\biggl(\frac{\delta_p T_p}{\sqrt{2}}\biggr) \tag{5.3}$$

and

$$\bigl|E\{\mathbf{X}R_p^i\}\bigr| \approx \mu_p^0 = \frac{\hat{r}_p\delta_p}{\sqrt{2\pi}}\exp\biggl(-\frac{\delta_p^2 T_p^2}{2}\biggr). \tag{5.4}$$

Then, we calculate the effective angle between the kernels by

$$\cos\theta_{pq}^k = \frac{E\{\mathbf{X}R_p^{i-k}\}\cdot E\{\mathbf{X}R_q^i\}}{\bigl|E\{\mathbf{X}R_p^{i-k}\}\bigr|\bigl|E\{\mathbf{X}R_q^i\}\bigr|}. \tag{5.5}$$

The last step is to calculate the coupling $\bar{W}$ from the spike correlations with delays $k = -N, \ldots, N$,

$$E\{R_1^i R_2^{i-k}\} = \nu_{21}^k + \sum_{j\geq 0} A_{21}^{kj}\bar{W}_{21}^j + \sum_{j\geq 0} A_{12}^{-kj}\bar{W}_{12}^j, \tag{5.6}$$

[5]The fact that T_p and $E\{\mathbf{X}R_p^i\}$ are given by equations analogous to (4.15) and (4.14) is not needed for the reconstruction.

where

$$A_{pq}^{kj} = \mu_q^0\big[\tilde{\nu}_{pq}^{kj} - \eta_{pq}^k\eta_{pq}^j + (\cos\theta_{pq}^k\cos\theta_{pq}^j - \cos\theta_{pp}^{k-j})\mu_{pq}^k\mu_{pq}^j\big],$$

$$\nu_{pq}^k = \frac{\hat{r}_p\hat{r}_q}{4}\operatorname{derfc}\left(\frac{\delta_pT_p}{\sqrt{2}}, \frac{\delta_qT_q}{\sqrt{2}}, \delta_p\delta_q\cos\theta_{pq}^k\right),$$

$$\tilde{\nu}_{pq}^{kj} = \begin{cases} \eta_{pq}^k & \text{for } j = k, \\ \frac{(\hat{r}_p)^2}{4}\operatorname{derfc}\left(\frac{\lambda_{pq}^k}{\sqrt{2}}, \frac{\lambda_{pq}^j}{\sqrt{2}}, \xi_{pq}^{kj}\right) & \text{otherwise,} \end{cases}$$

$$\eta_{pq}^k = \frac{\hat{r}_p}{2}\operatorname{erfc}\left(\frac{\lambda_{pq}^k}{\sqrt{2}}\right),$$

$$\mu_{pq}^k = \frac{\hat{r}_p\delta_p\exp\left(-\frac{1}{2}[\lambda_{pq}^k]^2\right)}{\sqrt{2\pi(1-\delta_p^2\delta_q^2\cos^2\theta_{pq}^k)}},$$

$$\lambda_{pq}^k = \frac{\delta_pT_p - \delta_p\delta_q^2T_q\cos\theta_{pq}^k}{\sqrt{1-\delta_p^2\delta_q^2\cos^2\theta_{pq}^k}},$$

and

$$\xi_{pq}^{kj} = \frac{\delta_p^2\cos\theta_{pp}^{k-j} - \delta_p^2\delta_q^2\cos\theta_{pq}^j\cos\theta_{pq}^k}{\sqrt{(1-\delta_p^2\delta_q^2\cos^2\theta_{pq}^j)(1-\delta_p^2\delta_q^2\cos^2\theta_{pq}^k)}}.$$

We assume that we have chosen the number of delays (given by $k = -N, \dots, N$) so that W_{21}^j and W_{12}^j for $j = 0, \dots, N$ are all the nonzero connectivity terms of the system. Unfortunately, even though the $\bar{W}$ are the only unknowns left in the system (5.6), we still have $2N+2$ unknowns with only $2N+1$ equations.

To reduce the number of unknowns, we simply do not attempt to distinguish $\bar{W}_{21}^0$ from $\bar{W}_{12}^0$. Although there is no reason these should be identical, the best we can do is calculate their sum. To solve the equations, we define a new $\bar{W}^j$ by

$$\bar{W}^j = \begin{cases} \bar{W}_{12}^{-j} & \text{for } j < 0, \\ \bar{W}_{12}^0 + \bar{W}_{21}^0 & \text{for } j = 0, \\ \bar{W}_{21}^j & \text{for } j > 0. \end{cases} \tag{5.7}$$

Our new equation for the $\bar{W}$ is then

$$E\{R_1^iR_2^{i-k}\} = \nu_{21}^k + \sum_j \tilde{A}^{kj}\bar{W}^j, \tag{5.8}$$

where

$$\tilde{A}^{kj} = \begin{cases} A_{12}^{-k,-j} & \text{for } j < 0, \\ \frac{1}{2}(A_{12}^{-k0} + A_{21}^{k0}) & \text{for } j = 0, \\ A_{21}^{kj} & \text{for } j > 0. \end{cases} \tag{5.9}$$

If we let $\mathcal{S}^k = E\{R_1^i R_2^{i-k}\} - \nu_{21}^k$, we can write the solution of (5.8) for $\bar{W}^j$ in matrix-vector notation as $\bar{W} = \tilde{A}^{-1}\mathcal{S}$, where $\tilde{A}^{-1}$ denotes the matrix inverse of $\tilde{A}$. This solution of (5.8) for $\bar{W}^j$ modifies the correlations in $E\{R_1^i R_2^{i-k}\}$ in two ways. First, the subtraction of ν_{21}^k removes correlations due solely to the fact that neurons are responding to the same stimulus. (See [12] for a detailed discussion.) Second, inverting the matrix $\tilde{A}$ eliminates the filtering of $\bar{W}$ by the temporal structure of $\mathbf{h}_1^i$ and $\mathbf{h}_2^i$.

The relevant temporal structure of the kernels is captured by $\cos\bar{\theta}_{pq}^k = \mathbf{h}_p^{i-k} \cdot \mathbf{h}_q^i$. Clearly, $\cos\bar{\theta}_{pp}^0 = |\mathbf{h}_p^i|^2 = 1$. If, with this exception, $\cos\bar{\theta}_{pq}^k = 0$ (so that the kernels are orthogonal to each other and temporal shifts of themselves), then the effects of $\bar{W}$ are not filtered by the kernels. $\hat{A}$ is a diagonal matrix, and inverting $\tilde{A}$ simply scales the measured correlations. (Too see this fact, recall that $\operatorname{derfc}(a, b, 0) = \operatorname{erfc}(a)\operatorname{erfc}(b)$ and that we can interchange $\cos\bar{\theta}_{pq}^k$ and $\cos\theta_{pq}^k$ in expressions defining A since it appears in $O(\bar{W})$ terms.)

As the inner products $\cos\bar{\theta}_{pq}^k$ increase, the off-diagonal elements of $\tilde{A}$ grow. In fact, the inner products of the kernels with themselves ($\cos\bar{\theta}_{pp}^k$) will be close to 1 for k near 0 if the structure of the kernels changes slowly with time. Typically, the off-diagonal elements of $\tilde{A}$ will still be substantially less than the diagonal elements even with large $\cos\bar{\theta}_{pp}^k$, and inversion of $\tilde{A}$ will be stable. However, close examination of equations defining $\tilde{A}$ reveals that off-diagonals could become equal to the diagonal in the extreme case of very sharp nonlinearities and other parameter limits. (Parameters needed are $\epsilon_1 = \epsilon_2 = 0$ so that $\delta_1 = \delta_2 = 1$, as well as $\hat{r}_1 = \hat{r}_2 = 1$, $\cos\bar{\theta}_{pp}^k = 1$ for $k \neq 0$, and $\cos\bar{\theta}_{pq}^j = 0$ for $p \neq q$.)[6] In this case, the matrix $\tilde{A}$ could become almost singular, and its inversion would not be stable.

Outside this extreme case, the matrix $\tilde{A}$ is well conditioned, and solving (5.8) for $\bar{W}$ removes the filtering caused by the temporal structure of the kernels. Subject to the validity of the model (5.1), the result will faithfully reconstruct the underlying connectivity.

6. Results. To demonstrate the reconstruction procedure, we simulate a pair of coupled linear-nonlinear neurons (see (5.1)) responding to white noise input and use the above method to estimate the parameters. We assume that the maximum output rates $\hat{r}_p$ are known using alternative methods such as those described in [13]. Then, from the responses R_p^i and the discrete white noise input X, one can estimate $E\{R_p^i\}$, $E\{\mathbf{X}R_p^i\}$, and $E\{R_1^i R_2^{i-k}\}$ for $p = 1, 2$ and $k = -N, \ldots, N$. The maximum delay parameter N must be chosen large enough so that the $E\{R_1^i R_2^{i-k}\}$ capture the effects of the $\bar{W}^j$. In the examples, we set $N = 30$.

The calculations depend on estimating the inner products $E\{\mathbf{X}R_p^i\} \cdot E\{\mathbf{X}R_q^{i-k}\}$. We estimate each correlation by $E\{\mathbf{X}R_p^i\} \approx \langle \mathbf{X}R_p^i \rangle$, where $\langle\cdot\rangle$ represents averaging over a data set. A naive estimate of the inner product by $E\{\mathbf{X}R_p^i\} \cdot E\{\mathbf{X}R_q^{i-k}\} \approx \langle \mathbf{X}R_p^i \rangle \cdot \langle \mathbf{X}R_q^{i-k} \rangle$ will be highly biased, especially when the dimension of the kernels $\mathbf{h}_p^i$ and $\mathbf{h}_q^i$ is large. To reduce the bias, we estimate the covariance between the factors of each term defining $\langle \mathbf{X}R_p^i \rangle \cdot \langle \mathbf{X}R_q^{i-k} \rangle$ and subtract it from the estimate.[7]

For our simulations, we used kernels $\mathbf{h}_p^i$ that mimic linear kernels of neurons in

[6] Note that $\lim_{c\to 1} \operatorname{derfc}(a, a, c) = 2\operatorname{erfc}(a)$.

[7] This bias reduction is equivalent to estimating the product of expected values of two random variables Y and Z using the formula for covariance $E\{YZ\} - \operatorname{cov}(Y, Z) = E\{Y\}E\{Z\}$. For more details on bias reduction of inner products, see Appendix B of [13].

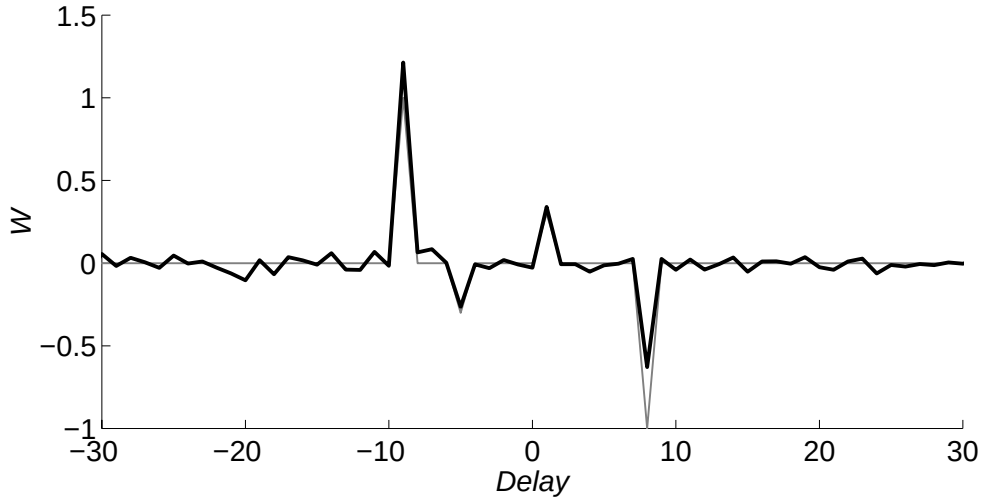


FIG. 1. *Estimated connectivity $\mathcal{W}$ (thick black line) when the nonlinearities are error functions. For comparison, the simulated connectivity $\bar{W}$ is shown with a thin gray line. $\mathcal{W}$ agrees quantitatively with $\bar{W}$, though the magnitudes of the large peaks differ. Delay is in units of time and is the spike time of neuron 1 minus the spike time of neuron 2.*

visual cortex [10]. We used the spatio-temporal linear kernels of the form

$$h_p(\mathbf{j},t) = \begin{cases} te^{-t/5}\exp\left(-\frac{|\mathbf{j}|^2}{50}\right)\sin(0.5(j_1\cos\phi_p + j_2\sin\phi_p)) & \text{for } t>0, \\ 0 & \text{otherwise,} \end{cases} \tag{6.1}$$

where $\mathbf{j} = (j_1, j_2)$ is the spatial grid point and t is time. We set the spatial axis parameters to be $\phi_1 = 0$ and $\phi_2 = \pi/4$. We sampled $h_p(\mathbf{j},t)$ on a $32 \times 32 \times 32$ grid and normalized it to form the unit vector $\mathbf{h}_p^i$. All units are in grid points. The detailed structure of the kernels is insignificant as the only relevant parameters from the kernels are their inner products $\cos\bar{\theta}_{pq}^k$.

In the first example, we set the parameters of the error function nonlinearity (i.e., (5.2)) to $\hat{r}_1 = \hat{r}_2 = 0.5$, $\bar{T}_1 = 1.5$, $\bar{T}_2 = 2.0$, $\epsilon_1 = 0.5$, and $\epsilon_2 = 1.0$. The precise parameter values are arbitrary; we chose them so that the neuron firing rates would be low as observed in white noise experiments. The results are not sensitive to these parameter choices. Just to illustrate the method, we set an artificial coupling of $\bar{W}_{21}^1 = 0.3$, $\bar{W}_{21}^8 = -1.0$, $\bar{W}_{12}^5 = -0.3$, and $\bar{W}_{12}^9 = 1.0$. All other coupling terms were set to zero. We simulated the system for 250,000 units of time, obtaining about 10,000 spikes from each neuron, a realistic number of spikes in white noise experiments [17].

To analyze the results, we assumed that we knew that $\hat{r}_p = 0.5$, and calculated all other parameters from the input-output statistics using the proposed method. We focus on the estimate of the simulated connectivity $\bar{W}$, denoting by $\mathcal{W}$ our estimate of the connectivity. As shown in Figure 1, the estimate $\mathcal{W}$ captures all the qualitative features of $\bar{W}$. For the lower magnitude coupling (with $|\bar{W}| = 0.3$), $\mathcal{W}$ also estimates the magnitudes accurately. However, when $|\bar{W}| = 1.0$, the first order approximation breaks down enough to cause $\mathcal{W}$ to overestimate the positive coupling by 20% and underestimate the magnitude of the negative coupling by nearly 40%. (The asymmetry between positive and negative coupling is most likely due to the low average firing rates of 0.04 spikes per unit time; cf. [14].)

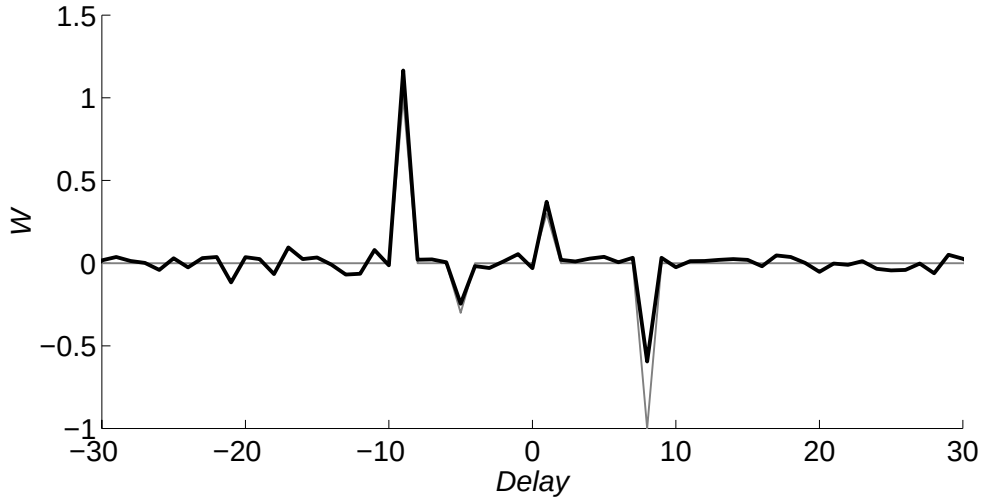


FIG. 2. *Estimated connectivity $\mathcal{W}$ (thick black line) when simulated power law nonlinearities are analyzed as error functions. $\mathcal{W}$ agrees with the simulated connectivity $\bar{W}$ (thin gray line) just as well as in the error function case of Figure* 1.

Since the stimulus standard deviation is assumed to be one, we have effectively scaled $\mathbf{X}$, and likewise $\bar{W}$, $\bar{T}_p$, and ϵ_p, by the stimulus standard deviation. When $|\bar{W}^j| = 1$, it is equal in magnitude to the standard deviation of $\mathbf{h}_p^i \cdot \mathbf{X}$. Since in this case the contribution of $\bar{W}^j$ in (5.1) is the same order of magnitude as the contribution of $\mathbf{h}_p^i \cdot \mathbf{X}$, one cannot expect the first order approximation to be valid. Not only are estimation errors, such as those shown in Figure 1, possible when the coupling magnitude is sufficiently large, but $\mathcal{W}$ can also show additional peaks due to the second order interactions that we ignored in section 5 (not shown).

For a second example, we demonstrate the robustness of the analysis to deviations in the form of the nonlinearities g_p. We repeat the first example, but rather than using an error function nonlinearity, we use a power law nonlinearity,

$$g_p(y) = \begin{cases} A_p y^{\beta_p} & \text{if } y > 0, \\ 0 & \text{otherwise,} \end{cases}$$

with $A_1 = 0.07$, $A_2 = 0.04$, $\beta_1 = 2.5$, and $\beta_2 = 2.0$ (we truncate so that $g_p(x) \leq 1$). Using the same $\bar{W}$ as above, we simulated the system for 250,000 units of time, obtaining approximately 10,000 spikes from neuron 1 and 5,000 spikes from neuron 2.

We analyze the output of the system identically to the first example. We assume that each nonlinearity was an error function nonlinearity with $\hat{r}_p = 1$ and calculate the error function parameters from $E\{R_P^i\}$ and $|E\{\mathbf{X}R_p^i\}|$. The resulting error function parameters (which include the effects from the connectivity) were $\epsilon_1 = 0.76$, $\epsilon_2 = 1.1$, $T_1 = 2.2$, and $T_2 = 3.0$. As shown in Figure 2, the method estimated the connectivity just as well as when the simulated nonlinearity really was an error function. The results were not sensitive to the selection of the maximum firing rate parameters, as the calculated $\mathcal{W}$ was virtually identical if we set $\hat{r}_p = 0.5$ or $\hat{r}_p = 2$ and repeated the analysis.

We repeated this test for simulations based on a wide variety of power law parameters A_p and β_p. We were unable to find an example for which the calculation of $\mathcal{W}$

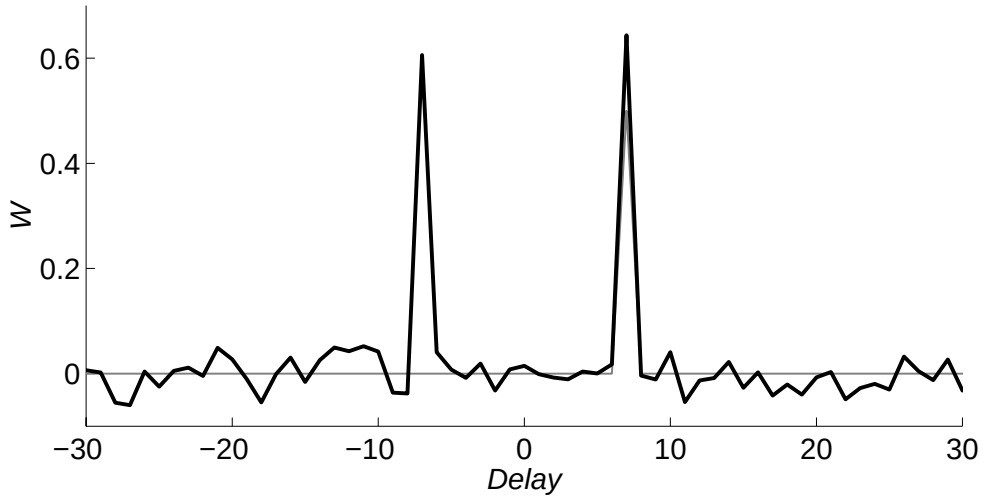


FIG. 3. *Estimated connectivity $\mathcal{W}$ (thick black line) when the two neurons receive common input from a third neuron. The peak at a delay of 7 units of time is due to the simulated connectivity $\bar{W}$ (thin gray line). However, the peak of $\mathcal{W}$ at a delay of -7 is due not to connectivity between the two neurons ($\bar{W} = 0$) but rather to the common input from the third neuron.*

was significantly worse than in Figure 2. Even with $\beta_p < 1$ so that the derivative of $g_p(y)$ was infinite at $y = 0$, the results were similar. The method simply is not sensitive to the detailed form of the nonlinearity.

The measure $\mathcal{W}$ cannot distinguish between correlations caused by the connectivity assumed in (5.1) and correlations caused by other mechanisms. For example, if the two neurons received common input from a third, unmeasured, neuron, that connectivity would appear in the calculation of $\mathcal{W}$.

To demonstrate, we simulated three coupled linear-nonlinear neurons analogous to (5.1). We used (6.1) with $\phi_3 = \pi/2$ for the linear kernel of the third neuron. All three neurons had error function nonlinearities with $\bar{T}_1 = 2$, $\bar{T}_2 = 2.5$, $\bar{T}_3 = 2$, $\epsilon_1 = 0.5$, $\epsilon_2 = 1$, and $\epsilon_2 = 0.7$. We created a connection from neuron 3 to both neurons 1 and 2, as well as a connection from neuron 2 to neuron 1. (We set $\bar{W}_{31}^1 = 1.5$, $\bar{W}_{32}^8 = 1.5$, and $\bar{W}_{21}^7 = 0.5$, leaving the other connectivity terms at zero.) We simulated the system for 250,000 units of time, obtaining approximately 12,000–13,000 spikes per neuron, and then analyzed the system as above by ignoring the output of neuron 3.

As shown in Figure 3, $\mathcal{W}$ has a peak at the delay of 7 corresponding to the connection from neuron 2 to neuron 1 ($\bar{W}_{21}^7$). However, $\mathcal{W}$ also has a peak at a delay of -7. This second peak does not correspond to any direct connection between neuron 1 and neuron 2 ($\bar{W}_{12}^7 = 0$). Instead, the peak is created because the connection from neuron 3 to neuron 2 is 7 units of time delayed compared to the connection from neuron 3 to neuron 1. Since $\mathcal{W}$ cannot distinguish between direct connections and common input, it must be interpreted with care. It cannot be viewed as representing the connectivity between the two measured neurons unless one could somehow rule out a mutual connection from any unmeasured neurons.

7. Discussion. We derived a method for analyzing a pair of coupled linear-nonlinear systems driven by white noise. Through measuring first and second order input-output statistics, one can characterize the systems. In particular, one can

reconstruct the coupling between the systems if the coupling is assumed to be of a particular form (see (5.1)).

We demonstrated that the method is robust to variations in the detailed form of the nonlinearity. We believe this robustness is due to the smoothing by the white noise input. Each input-output statistic depends on the nonlinearities $g_p(\cdot)$ only through expected values over the white noise. The effect of this smoothing is most clearly seen in the initial expression for each statistic in Appendix A. The $g_p(\cdot)$ appear in the integrals as either $g_p(\mathbf{h}_p^i \cdot \mathbf{x})$ or $g_p'(\mathbf{h}_p^i \cdot \mathbf{x})$. Since the kernels are unit vectors, the arguments of the nonlinearity are standard normals. Only the integrals of the $g_p(\cdot)$ over the probability density function of standard normals, not pointwise evaluation of the $g_p(\cdot)$, affect the input-output statistics. These integrals smooth out minor differences between nonlinearity shapes.

Since the method is a first order approximation in the coupling magnitude, measurements of large $\mathcal{W}$ (on the order of the standard deviation of an input component) must be viewed cautiously. According to our simulation results, the breakdown of the first order approximation typically leads only to deviations in the magnitude of the estimated connectivity. However, in extreme cases, large connectivity could lead to the emergence of second order effects in the form of additional peaks in $\mathcal{W}$ that do not reflect the connectivity $\bar{W}$.

More importantly, the method cannot distinguish between the assumed mutual coupling of the model and other mechanisms for creating correlations between the responses, such as common input from outside sources. Measurements of $\mathcal{W}$ would be evidence of mutual coupling only if other mechanisms for correlations could be ruled out. Nonetheless, even if the source of $\mathcal{W}$ cannot be definitively determined, measurement of $\mathcal{W}$ still could provide evidence about the time scale and magnitudes of the interactions in the underlying neural network.

The proposed method was developed to analyze multielectrode recordings of neurons in response to a white noise stimulus. However, the linear-nonlinear model assumed by the analysis is only a crude, phenomenological approximation to the biology. To better interpret the results of the method, one must be able to assign significance to nonzero measurements of $\mathcal{W}$. One future challenge is to develop methods for identifying cases in which nonzero measurements of $\mathcal{W}$ are due simply to deviations from the linear-nonlinear model.

Appendix A. Details of derivation for unidirectional coupling.

A.1. Probability of a spike in neuron 1. Under the first order approximation in $\bar{W}$, we can simplify (2.2) for neuron 1 to

$$\begin{aligned}\Pr\bigl(R_1^i = 1 \bigm| \mathbf{X} = \mathbf{x}, \mathbf{R}_2 = \mathbf{r}_2\bigr) &= g_1\biggl(\mathbf{h}_1^i \cdot \mathbf{x} + \sum_{j \geq 0} \bar{W}_{21}^j r_2^{i-j}\biggr) \\ &= g_1(\mathbf{h}_1^i \cdot \mathbf{x}) + g_1'(\mathbf{h}_1^i \cdot \mathbf{x}) \sum_{j \geq 0} \bar{W}_{21}^j r_2^{i-j}. \end{aligned} \tag{A.1}$$

The probability of a spike in neuron 1 is then

$$\begin{aligned}\Pr\bigl(R_1^i = 1 \bigm| \mathbf{X} = \mathbf{x}\bigr) &= \sum_{\mathbf{r}_2} \Pr\bigl(R_1^i = 1 \bigm| \mathbf{X} = \mathbf{x}, \mathbf{R}_2 = \mathbf{r}_2\bigr) \prod_{\tilde{\jmath}} \Pr\bigl(R_2^{i-\tilde{\jmath}} = r_2^{i-\tilde{\jmath}} \bigm| \mathbf{X} = \mathbf{x}\bigr) \\ &= g_1(\mathbf{h}_1^i \cdot \mathbf{x}) \sum_{\mathbf{r}_2} \prod_{\tilde{\jmath}} \Pr\bigl(R_2^{i-\tilde{\jmath}} = r_2^{i-\tilde{\jmath}} \bigm| \mathbf{X} = \mathbf{x}\bigr) \\ &\quad + g_1'(\mathbf{h}_1^i \cdot \mathbf{x}) \sum_{j \geq 0} \bar{W}_{21}^j \sum_{\mathbf{r}_2} r_2^{i-j} \prod_{\tilde{\jmath}} \Pr\bigl(R_2^{i-\tilde{\jmath}} = r_2^{i-\tilde{\jmath}} \bigm| \mathbf{X} = \mathbf{x}\bigr). \end{aligned} \tag{A.2}$$

The sum is over all values of $\mathbf{r}_2$, where each component $r_2^{\tilde{\jmath}}$ can be either one or zero; i.e., this sum is over every possible spike combination of neuron 2. The product reflects the assumption that, since $\bar{W}_{12} = 0$, the responses of neuron 2, when conditioned on the stimulus, are independent.

The total probability of any spike combination of neuron 2 must equal one,

$$\sum_{\mathbf{r}_2} \prod_{\tilde{\jmath}} \Pr\bigl(R_2^{i-\tilde{\jmath}} = r_2^{i-\tilde{\jmath}} \big| \mathbf{X} = \mathbf{x}\bigr) = 1. \tag{A.3}$$

Moreover, since $r_2^{i-j} \in \{0,1\}$, only terms where $r_2^{i-j} = 1$ make a contribution in the coefficient of $\bar{W}_{21}^j$:

$$\begin{aligned}\sum_{\mathbf{r}_2} r_2^{i-j} \prod_{\tilde{\jmath}} \Pr\bigl(R_2^{i-\tilde{\jmath}} = r_2^{i-\tilde{\jmath}} \big| \mathbf{X} = \mathbf{x}\bigr) \\ = \Pr\bigl(R_2^{i-j} = 1 \big| \mathbf{X} = \mathbf{x}\bigr) \sum_{\mathbf{r}_2 \text{ except } r_2^{i-j}} \prod_{\tilde{\jmath} \neq j} \Pr\bigl(R_2^{i-\tilde{\jmath}} = r_2^{i-\tilde{\jmath}} \big| \mathbf{X} = \mathbf{x}\bigr) \\ = g_2(\mathbf{h}_2^{i-j} \cdot \mathbf{x}). \qquad \text{(A.4)}\end{aligned}$$

In the last step, we used a generalization of (A.3) excluding the $i - j$ time interval.

Combining (A.2), (A.3), and (A.4), the probability of a spike in neuron 1 is

$$\Pr\bigl(R_1^i = 1 \big| \mathbf{X} = \mathbf{x}\bigr) = g_1(\mathbf{h}_1^i \cdot \mathbf{x}) + \sum_{j \geq 0} \bar{W}_{21}^j g_1'(\mathbf{h}_1^i \cdot \mathbf{x}) g_2(\mathbf{h}_2^{i-j} \cdot \mathbf{x}), \tag{A.5}$$

where $=$ indicates equality within $O(\bar{W}^2)$.

A.2. Probability of spike pairs. In the case of unidirectional coupling, the probability of a spike pair is

$$\begin{aligned}\Pr\bigl(R_1^i = 1 \,\&\, R_2^{i-k} = 1 \big| \mathbf{X} = \mathbf{x}, \mathbf{R}_2 = \mathbf{r}_2\bigr) &= g_1\left(\mathbf{h}_1^i \cdot \mathbf{x} + \sum_{j \geq 0} \bar{W}_{21}^j r_2^{i-j}\right) r_2^{i-k} \qquad \text{(A.6)} \\ &= g_1(\mathbf{h}_1^i \cdot \mathbf{x}) r_2^{i-k} + g_1'(\mathbf{h}_1^i \cdot \mathbf{x}) r_2^{i-k} \bar{W}_{21}^k + g_1'(\mathbf{h}_1^i \cdot \mathbf{x}) r_2^{i-k} \sum_{\substack{j \geq 0 \\ j \neq k}} \bar{W}_{21}^j r_2^{i-j}.\end{aligned}$$

Note that $(r_2^{i-k})^2 = r_2^{i-k}$ since $r_2^{i-k} \in \{0,1\}$.

If we repeat the same procedure as in the previous section,

$$\begin{aligned}\Pr\bigl(R_1^i = 1 \,\&\, R_2^{i-k} = 1 \big| \mathbf{X} = \mathbf{x}\bigr) \qquad \text{(A.7)} \\ = \sum_{\mathbf{r}_2} \Pr\bigl(R_1^i = 1 \,\&\, R_2^{i-k} = 1 \big| \mathbf{X} = \mathbf{x}, \mathbf{R}_2 = \mathbf{r}_2\bigr) \prod_{\tilde{\jmath}} \Pr\bigl(R_2^{i-\tilde{\jmath}} = r_2^{i-\tilde{\jmath}} \big| \mathbf{X} = \mathbf{x}\bigr),\end{aligned}$$

the only new term will be

$$\sum_{\mathbf{r}_2} r_2^{i-j} r_2^{i-k} \prod_{\tilde{\jmath}} \Pr\bigl(R_2^{i-\tilde{\jmath}} = r_2^{i-\tilde{\jmath}} \big| \mathbf{X} = \mathbf{x}\bigr) = g_2(\mathbf{h}_2^{i-k} \cdot \mathbf{x}) g_2(\mathbf{h}_2^{i-j} \cdot \mathbf{x}). \tag{A.8}$$

Therefore,

$$\begin{aligned}\Pr\bigl(R_1^i = 1 \,\&\, R_2^{i-k} = 1 \big| \mathbf{X} = \mathbf{x}\bigr) = g_1(\mathbf{h}_1^i \cdot \mathbf{x}) g_2(\mathbf{h}_2^{i-k} \cdot \mathbf{x}) \\ + \bar{W}_{21}^k g_1'(\mathbf{h}_1^i \cdot \mathbf{x}) g_2(\mathbf{h}_2^{i-k} \cdot \mathbf{x}) \\ + \sum_{\substack{j \geq 0 \\ j \neq k}} \bar{W}_{21}^j g_1'(\mathbf{h}_1^i \cdot \mathbf{x}) g_2(\mathbf{h}_2^{i-k} \cdot \mathbf{x}) g_2(\mathbf{h}_2^{i-j} \cdot \mathbf{x}). \qquad \text{(A.9)}\end{aligned}$$

A.3. Mean rate of neuron 1. The mean rate of neuron 1 (see (A.5)) is given by

$$\begin{aligned} E\{R_1^i\} &= \frac{1}{(2\pi)^{n/2}} \int \Pr\bigl(R_1^i = 1 \bigm| \mathbf{X} = \mathbf{x}\bigr) e^{-|\mathbf{x}|^2/2} d\mathbf{x} \\ &= \frac{1}{(2\pi)^{n/2}} \int g_1(\mathbf{h}_1^i \cdot \mathbf{x}) e^{-|\mathbf{x}|^2/2} d\mathbf{x} \\ &\quad + \sum_{j \ge 0} \frac{\bar{W}_{21}^j}{(2\pi)^{n/2}} \int g_1'(\mathbf{h}_1^i \cdot \mathbf{x}) g_2(\mathbf{h}_2^{i-j} \cdot \mathbf{x}) e^{-|\mathbf{x}|^2/2} d\mathbf{x}. \end{aligned} \tag{A.10}$$

The first term is identical to the uncoupled case. For the rest of the terms, we use a different coordinate system for each j. The first unit vector is $\mathbf{e}_1 = \mathbf{h}_1^i$, and the second unit vector is the component of $\mathbf{h}_2^{i-j}$ that is perpendicular to $\mathbf{h}_1^i$, so that $\mathbf{h}_2^{i-j} = \mathbf{e}_1 \cos\bar{\theta}_{21}^j + \mathbf{e}_2 \sin\bar{\theta}_{21}^j$.

We change variables and integrate by parts (assuming (3.5)) to simplify the jth term:

$$\begin{aligned} &\frac{\bar{W}_{21}^j}{2\pi} \int g_1'(x_1) g_2(x_1 \cos\bar{\theta}_{21}^j + x_2 \sin\bar{\theta}_{21}^j) e^{-\frac{x_1^2+x_2^2}{2}} dx_1 dx_2 \\ &\quad = \frac{\bar{W}_{21}^j}{2\pi} \int g_1'(u) g_2(v) \exp\biggl(-\frac{u^2}{2} - \frac{(v - u\cos\bar{\theta}_{21}^j)^2}{2\sin^2\bar{\theta}_{21}^j}\biggr) \frac{du\,dv}{\sin\bar{\theta}_{21}^j} \\ &\quad = \frac{\bar{W}_{21}^j}{2\sqrt{2\pi}} \int g_1'(u) g_2'(v) e^{-\frac{u^2}{2}} \operatorname{erfc}\biggl(\frac{v - u\cos\bar{\theta}_{21}^j}{\sqrt{2}\sin\bar{\theta}_{21}^j}\biggr) du\,dv. \end{aligned} \tag{A.11}$$

The mean rate of neuron 1 is thus

$$\begin{aligned} E\{R_1^i\} = \frac{1}{\sqrt{2\pi}} &\int g_1(u) e^{-\frac{u^2}{2}} du \\ &+ \sum_{j \ge 0} \frac{\bar{W}_{21}^j}{2\sqrt{2\pi}} \int g_1'(u) g_2'(v) e^{-\frac{u^2}{2}} \operatorname{erfc}\biggl(\frac{v - u\cos\bar{\theta}_{21}^j}{\sqrt{2}\sin\bar{\theta}_{21}^j}\biggr) du\,dv. \end{aligned} \tag{A.12}$$

A.4. Correlation of spikes of neuron 1 with the stimulus. The stimulus-spike correlation of neuron 1 (see (A.5)) is

$$\begin{aligned} E\{\mathbf{X}R_1^i\} &= \frac{1}{(2\pi)^{n/2}} \int \mathbf{x} \Pr\bigl(R_1^i = 1 \bigm| \mathbf{X} = \mathbf{x}\bigr) e^{-|\mathbf{x}|^2/2} d\mathbf{x} \\ &= \frac{1}{(2\pi)^{n/2}} \int \mathbf{x} g_1(\mathbf{h}_1^i \cdot \mathbf{x}) e^{-|\mathbf{x}|^2/2} d\mathbf{x} \\ &\quad + \sum_{j \ge 0} \frac{\bar{W}_{21}^j}{(2\pi)^{n/2}} \int \mathbf{x} g_1'(\mathbf{h}_1^i \cdot \mathbf{x}) g_2(\mathbf{h}_2^{i-j} \cdot \mathbf{x}) e^{-|\mathbf{x}|^2/2} d\mathbf{x}. \end{aligned} \tag{A.13}$$

The first term is identical to the uncoupled case, becoming

$$\frac{1}{\sqrt{2\pi}} \int g_1'(u) e^{-\frac{u^2}{2}} du\, \mathbf{h}_1^i$$

with an integration by parts. For the rest of the terms, just as in the previous section, we will use a different coordinate system for each j, with $\mathbf{e}_1 = \mathbf{h}_1^i$ and with $\mathbf{e}_2$ being

the component of $\mathbf{h}_2^{i-j}$ that is perpendicular to $\mathbf{h}_1^i$. We will denote this second unit vector by

$$\overset{\perp ji}{\mathbf{h}}_{21} = \frac{\mathbf{h}_2^{i-j} - \cos\bar{\theta}_{21}^j \mathbf{h}_1^i}{\sin\bar{\theta}_{21}^j}. \tag{A.14}$$

Note that $\mathbf{h}_2^{i-j} = \mathbf{h}_1^i \cos\bar{\theta}_{21}^j + \overset{\perp ji}{\mathbf{h}}_{21} \sin\bar{\theta}_{21}^j$. The jth term thus has two nonzero components,

$$\begin{aligned}
&\frac{\bar{W}_{21}^j}{2\pi} \int (x_1 \mathbf{h}_1^i + x_2 \overset{\perp ji}{\mathbf{h}}_{21}) g_1'(x_1) g_2(x_1 \cos\bar{\theta}_{21}^j + x_2 \sin\bar{\theta}_{21}^j) e^{-\frac{x_1^2+x_2^2}{2}} dx_1 dx_2 \\
&\qquad = \bar{W}_{21}^j (I_{j,1} \mathbf{h}_1^i + I_{j,2} \overset{\perp ji}{\mathbf{h}}_{21}),
\end{aligned} \tag{A.15}$$

where the above defines $I_{j,1}$ and $I_{j,2}$. We change variables and integrate by parts (assuming (3.5)) to simplify the first component:

$$\begin{aligned}
I_{j,1} &= \frac{1}{2\pi} \int u g_1'(u) g_2(v) \exp\left(-\frac{u^2}{2} - \frac{(v - u\cos\bar{\theta}_{21}^j)^2}{2\sin^2\bar{\theta}_{21}^j} \right) \frac{du\,dv}{\sin\bar{\theta}_{21}^j} \\
&= \frac{1}{2\sqrt{2\pi}} \int g_1'(u) g_2'(v) u e^{-\frac{u^2}{2}} \operatorname{erfc}\left(\frac{v - u\cos\bar{\theta}_{21}^j}{\sqrt{2}\sin\bar{\theta}_{21}^j} \right) du\,dv.
\end{aligned} \tag{A.16}$$

To simplify $I_{j,2}$, we first integrate by parts in the x_2 variable, then change variables:

$$\begin{aligned}
I_{j,2} &= \frac{1}{2\pi} \int g_1'(x_1) g_2'(x_1 \cos\bar{\theta}_{21}^j + x_2 \sin\bar{\theta}_{21}^j) \sin\bar{\theta}_{21}^j e^{-\frac{x_1^2+x_2^2}{2}} dx_1 dx_2 \\
&= \frac{1}{2\pi} \int g_1'(u) g_2'(v) \exp\left(-\frac{u^2 - 2\cos\bar{\theta}_{21}^j uv + v^2}{2\sin^2\bar{\theta}_{21}^j} \right) du\,dv.
\end{aligned} \tag{A.17}$$

Combining these results, the stimulus-spike correlation of neuron 1 is

$$\begin{aligned}
E\{\mathbf{X} R_1^i\} = \frac{1}{\sqrt{2\pi}} \Bigg[&\int g_1'(u) e^{-\frac{u^2}{2}} du \\
&+ \sum_{j\ge 0} \frac{\bar{W}_{21}^j}{2} \int g_1'(u) g_2'(v) u e^{-\frac{u^2}{2}} \operatorname{erfc}\left(\frac{v - u\cos\bar{\theta}_{21}^j}{\sqrt{2}\sin\bar{\theta}_{21}^j} \right) du\,dv \Bigg] \mathbf{h}_1^i \\
&+ \sum_{j\ge 0} \frac{\bar{W}_{21}^j}{2\pi} \int g_1'(u) g_2'(v) \exp\left(-\frac{u^2 - 2\cos\bar{\theta}_{21}^j uv + v^2}{2\sin^2\bar{\theta}_{21}^j} \right) du\,dv\, \overset{\perp ji}{\mathbf{h}}_{21}.
\end{aligned} \tag{A.18}$$

A.5. Correlation between spikes of neurons 1 and 2. The correlation between spikes of neuron 1 and neuron 2 is (see (A.9))

$$\begin{aligned}
E\{R_1^i R_2^{i-k}\} &= \frac{1}{(2\pi)^{n/2}} \int \Pr\big(R_1^i = 1 \,\&\, R_2^{i-k} = 1 \big| \mathbf{X} = \mathbf{x}\big) e^{-\frac{|\mathbf{x}|^2}{2}} d\mathbf{x} \\
&= \frac{1}{(2\pi)^{n/2}} \int \Bigg[g_1(\mathbf{h}_1^i \cdot \mathbf{x}) g_2(\mathbf{h}_2^{i-k} \cdot \mathbf{x}) \\
&\qquad + g_1'(\mathbf{h}_1^i \cdot \mathbf{x}) g_2(\mathbf{h}_2^{i-k} \cdot \mathbf{x}) \left(\bar{W}_{21}^k + \sum_{j\ge 0,\, j\ne k} \bar{W}_{21}^j g_2(\mathbf{h}_2^{i-j} \cdot \mathbf{x}) \right) \Bigg] e^{-\frac{|\mathbf{x}|^2}{2}} d\mathbf{x}.
\end{aligned} \tag{A.19}$$

The first term is identical to the uncoupled case (i.e., (3.6)). The $\bar{W}_{21}^k$ term is identical to (A.11).

For the $\bar{W}_{21}^j$ terms with $j \neq k$, we let $\mathbf{e}_1 = \mathbf{h}_1^i$ and $\mathbf{e}_2 = \overset{\perp ki}{\mathbf{h}}_{21}$, and let the third unit vector be the component of $\mathbf{h}_2^{i-j}$ perpendicular to both $\mathbf{e}_1$ and $\mathbf{e}_2$ so that

$$\mathbf{h}_2^{i-j} = \mathbf{e}_1 \cos\bar{\theta}_{21}^j + \mathbf{e}_2 c_{21}^{kj} \sin\bar{\theta}_{21}^j + \mathbf{e}_3 \sin\bar{\theta}_{21}^j \sqrt{1-(c_{21}^{kj})^2},$$

where

$$c_{21}^{kj} = \overset{\perp ki}{\mathbf{h}}_{21} \cdot \overset{\perp ji}{\mathbf{h}}_{21} = \frac{\cos\theta_{22}^{k-j} - \cos\bar{\theta}_{21}^k \cos\bar{\theta}_{21}^j}{\sin\bar{\theta}_{21}^k \sin\bar{\theta}_{21}^j}.$$

Denoting the $\bar{W}_{21}^j$ terms in (A.19) by $\bar{W}_{21}^j I_{kj}$ and changing variables, we compute

$$\begin{aligned} I_{kj} &= \frac{1}{(2\pi)^{3/2}} \int g_1'(x_1) g_2(x_1\cos\bar{\theta}_{21}^k + x_2 \sin\bar{\theta}_{21}^k) \\ &\qquad \times g_2\Big(x_1\cos\bar{\theta}_{21}^j + x_2 c_{21}^{kj}\sin\bar{\theta}_{21}^j + x_3 \sin\bar{\theta}_{21}^j\sqrt{1-(c_{21}^{kj})^2}\Big) e^{-\frac{x_1^2+x_2^2+x_3^2}{2}}\, dx_1 dx_2 dx_3 \\ &= \frac{1}{(2\pi)^{3/2}} \int \frac{du_1 du_2 du_3\, g_1'(u_1) g_2(u_2) g_2(u_3)}{\sin\bar{\theta}_{21}^k \sin\bar{\theta}_{21}^j \sqrt{1-(c_{21}^{kj})^2}} \\ &\qquad \times \exp\left(-\frac{u_1^2}{2} - \frac{(u_2 - u_1\cos\bar{\theta}_{21}^k)^2}{2\sin^2\bar{\theta}_{21}^k} - \frac{\left[\frac{u_3 - u_1\cos\bar{\theta}_{21}^j}{\sin\bar{\theta}_{21}^j} - c_{21}^{kj}\frac{u_2 - u_1\cos\bar{\theta}_{21}^k}{\sin\bar{\theta}_{21}^k}\right]^2}{2[1-(c_{21}^{kj})^2]} \right). \end{aligned}$$

Using (3.5) and integrating by parts twice as in the derivation of (3.6), we simplify this expression to

$$\begin{aligned} I_{kj} = \frac{1}{4\sqrt{2\pi}} &\int g_1'(u_1) g_2'(u_2) g_2'(u_3) e^{-\frac{u_1^2}{2}} \\ &\times \operatorname{derfc}\left(\frac{u_2 - u_1\cos\bar{\theta}_{21}^k}{\sqrt{2}\sin\bar{\theta}_{21}^k}, \frac{u_3 - u_1\cos\bar{\theta}_{21}^j}{\sqrt{2}\sin\bar{\theta}_{21}^j}, c_{21}^{kj} \right) du_1 du_2 du_3. \end{aligned} \tag{A.20}$$

The correlation between spikes of neuron 1 and neuron 2 is therefore

$$\begin{aligned} E\{R_1^i R_2^{i-k}\} &= \frac{1}{4}\int g_1'(u_1) g_2'(u_2) \operatorname{derfc}\left(\frac{u_1}{\sqrt{2}}, \frac{u_2}{\sqrt{2}}, \cos\bar{\theta}_{21}^k\right) du_1 du_2 \\ &\quad + \frac{\bar{W}_{21}^k}{2\sqrt{2\pi}} \int g_1'(u_1) g_2'(u_2) e^{-\frac{u_1^2}{2}} \operatorname{erfc}\left(\frac{u_2 - u_1\cos\bar{\theta}_{21}^k}{\sqrt{2}\sin\bar{\theta}_{21}^k}\right) du_1 du_2 \\ &\quad + \sum_{j\geq 0,\, j\neq k} \frac{\bar{W}_{21}^j}{4\sqrt{2\pi}} \int du_1 du_2 du_3\, g_1'(u_1) g_2'(u_2) g_2'(u_3) e^{-\frac{u_1^2}{2}} \\ &\quad \times \operatorname{derfc}\left(\frac{u_2 - u_1\cos\bar{\theta}_{21}^k}{\sqrt{2}\sin\bar{\theta}_{21}^k}, \frac{u_3 - u_1\cos\bar{\theta}_{21}^j}{\sqrt{2}\sin\bar{\theta}_{21}^j}, \frac{\cos\theta_{22}^{k-j} - \cos\bar{\theta}_{21}^j\cos\bar{\theta}_{21}^k}{\sin\bar{\theta}_{21}^j\sin\bar{\theta}_{21}^k} \right). \end{aligned} \tag{A.21}$$

Appendix B. Formulas used in derivations. In all formulas, each sine is assumed to be positive.

The formulas

$$\frac{1}{\epsilon_p\sqrt{2\pi}}\iint \exp\left(-\frac{(x-T_p)^2}{2\epsilon_p^2}-\frac{x^2}{2}\right)dx = \delta_p e^{-\frac{\delta_p^2 T_p^2}{2}} \tag{B.1}$$

and

$$\begin{aligned}\frac{1}{2\pi\epsilon_p\epsilon_q}\iint \exp\left(-\frac{(x-T_p)^2}{2\epsilon_p^2}-\frac{(y-T_q)^2}{2\epsilon_q^2}-\frac{x^2-2xy\cos\theta+y^2}{2\sin^2\theta}\right)dx\,dy \\ = \frac{\delta_p\delta_q\sin\theta}{\sqrt{1-\delta_p^2\delta_q^2\cos^2\theta}}\exp\left(-\frac{\delta_p^2T_p^2-2\delta_p^2\delta_q^2T_pT_q\cos\theta+\delta_q^2T_q^2}{2(1-\delta_p^2\delta_q^2\cos^2\theta)}\right),\end{aligned} \tag{B.2}$$

where $\delta_q = 1/\sqrt{1+\epsilon_q^2}$, follow from the application of

$$\int \exp(-[ax^2+bx+c])\,dx = \sqrt{\frac{\pi}{a}}\exp\left(\frac{b^2}{4a}-c\right)$$

for $a>0$.

For the following two formulas, change variables in the double integral so that one of the new variables is parallel to the line $u = dx+f$ (where u is the integration variable of the $\operatorname{erfc}(\cdot)$). By completing the square in the resulting integrands, one can derive both

$$\int \exp(-[ax^2+bx+c])\operatorname{erfc}(dx+f)\,dx = \sqrt{\frac{\pi}{a}}\exp\left(\frac{b^2}{4a}-c\right)\operatorname{erfc}\left(\frac{2af-bd}{2\sqrt{a(a+d^2)}}\right) \tag{B.3}$$

and

$$\begin{aligned}&\int x\exp(-[ax^2+bx+c])\operatorname{erfc}(dx+f)\,dx \\ &\quad = -\frac{1}{a}\exp\left(\frac{b^2}{4a}-c\right)\left[\frac{d\exp\left(-\frac{(2af-bd)^2}{4a(a+d^2)}\right)}{\sqrt{a+d^2}}+\frac{b\sqrt{\pi}}{2\sqrt{a}}\operatorname{erfc}\left(\frac{2af-bd}{2\sqrt{a(a+d^2)}}\right)\right]\end{aligned} \tag{B.4}$$

for $a>0$. By applying (B.3) twice, one can show that

$$\begin{aligned}\frac{1}{2\pi\epsilon_p\epsilon_q}\iint \exp\left(-\frac{(x-T_p)^2}{2\epsilon_p^2}-\frac{(y-T_q)^2}{2\epsilon_q^2}-\frac{x^2}{2}\right)\operatorname{erfc}\left(\frac{y-x\cos\theta}{\sqrt{2}\sin\theta}\right)dx\,dy \\ = \delta_p e^{-\frac{\delta_p^2T_p^2}{2}}\operatorname{erfc}\left(\frac{\delta_qT_q-\delta_p^2\delta_qT_p\cos\theta}{\sqrt{2(1-\delta_p^2\delta_q^2\cos^2\theta)}}\right),\end{aligned} \tag{B.5}$$

and by applying both (B.3) and (B.4), one can show that

$$\begin{aligned}&\frac{1}{2\pi\epsilon_p\epsilon_q}\iint x\exp\left(-\frac{(x-T_p)^2}{2\epsilon_p^2}-\frac{(y-T_q)^2}{2\epsilon_q^2}-\frac{x^2}{2}\right)\operatorname{erfc}\left(\frac{y-x\cos\theta}{\sqrt{2}\sin\theta}\right)dx\,dy \\ &\quad = \delta_p^3T_pe^{-\frac{\delta_p^2T_p^2}{2}}\operatorname{erfc}\left(\frac{\delta_qT_q-\delta_p^2\delta_qT_p\cos\theta}{\sqrt{2(1-\delta_p^2\delta_q^2\cos^2\theta)}}\right) \\ &\qquad + \frac{2\delta_p\delta_q(1-\delta_p^2)\cos\theta}{\sqrt{2\pi(1-\delta_p^2\delta_q^2\cos^2\theta)}}\exp\left(-\frac{\delta_p^2T_p^2-2\delta_p^2\delta_q^2T_pT_q\cos\theta+\delta_q^2T_q^2}{2(1-\delta_p^2\delta_q^2\cos^2\theta)}\right).\end{aligned} \tag{B.6}$$

For the following formula, let u be the integration variable of the $\operatorname{erfc}(\cdot)$. Then change variables in the triple integral so that the first variable is parallel to the line $y = dx + f$, and a linear combination of the first and second variables is parallel to the line $u = gx + h - ky$. By repeatedly completing the square in the integrand, one can derive that

$$\int dx \exp(-[ax^2 + bx + c]) \int_{dx+f}^{\infty} dy\, e^{-y^2} \operatorname{erfc}(gx + h - ky)$$

(B.7)

$$= \sqrt{\frac{\pi}{a}} e^{(\frac{b^2}{4a} - c)} \int_{\frac{2af - bd}{2\sqrt{a(a+d^2)}}}^{\infty} e^{-u^2} \operatorname{erfc}\left(\frac{(2ha - bg)\sqrt{a + d^2} - 2\sqrt{a}(gd + ka)u}{2a\sqrt{(kd - g)^2 + a + d^2}} \right) du$$

for $a > 0$. Repeated application of (B.7), combined with extensive algebra, yields

$$\text{(B.8)} \quad \frac{1}{2\pi\epsilon_p\epsilon_q} \iint \exp\left(-\frac{(x - T_p)^2}{2\epsilon_p^2} - \frac{(y - T_q)^2}{2\epsilon_q^2} \right) \operatorname{derfc}\left(\frac{x}{\sqrt{2}}, \frac{y}{\sqrt{2}}, \cos\theta \right) dx\, dy$$
$$= \operatorname{derfc}\left(\frac{\delta_p T_p}{\sqrt{2}}, \frac{\delta_q T_q}{\sqrt{2}}, \delta_p \delta_q \cos\theta \right)$$

and

$$\frac{1}{(2\pi)^{3/2}\epsilon_p\epsilon_q^2} \iiint \exp\left(-\frac{(x - T_p)^2}{2\epsilon_p^2} - \frac{(y - T_q)^2}{2\epsilon_q^2} - \frac{(z - T_q)^2}{2\epsilon_q^2} - \frac{x^2}{2} \right)$$
$$\times \operatorname{derfc}\left(\frac{z - x\cos\theta}{\sqrt{2}\sin\theta}, \frac{y - x\cos\phi}{\sqrt{2}\sin\phi}, \frac{\cos\psi - \cos\theta\cos\phi}{\sin\theta\sin\phi} \right) dx\, dy\, dz$$

(B.9)

$$= \delta_p e^{-\frac{\delta_p^2 T_p^2}{2}} \operatorname{derfc}\left(\frac{\delta_q T_q - \delta_p^2 \delta_q T_p \cos\theta}{\sqrt{2(1 - \delta_p^2\delta_q^2\cos^2\theta)}}, \frac{\delta_q T_q - \delta_p^2 \delta_q T_p \cos\phi}{\sqrt{2(1 - \delta_p^2\delta_q^2\cos^2\phi)}}, \frac{\delta_q^2\cos\psi - \delta_p^2\delta_q^2\cos\theta\cos\phi}{\sqrt{(1 - \delta_p^2\delta_q^2\cos^2\theta)(1 - \delta_p^2\delta_q^2\cos^2\phi)}} \right),$$

where $\operatorname{derfc}(\cdot)$ is defined by (3.7).

Acknowledgments. The author thanks Dario Ringach for numerous helpful discussions throughout the development of this research and Charlie Peskin and Dan Tranchina for constructive criticism on an early version of these ideas.

REFERENCES

[1] Y. Dan, J.-M. Alonso, W. M. Usrey, and R. C. Reid, *Coding of visual information by precisely correlated spikes in the lateral geniculate nucleus*, Nature Neurosci., 1 (1998), pp. 501–507.

[2] G. C. DeAngelis, I. Ohzawa, and R. D. Freeman, *Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex.* I. *General characteristics and postnatal development*, J. Neurophysiol., 69 (1993), pp. 1091–1117.

[3] G. C. DeAngelis, I. Ohzawa, and R. D. Freeman, *Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex.* II. *Linearity of temporal and spatial summation*, J. Neurophysiol., 69 (1993), pp. 1118–1135.

[4] E. DeBoer and P. Kuyper, *Triggered correlation*, IEEE Trans. Biomed. Eng., 15 (1968), pp. 169–179.

[5] R. C. deCharms, D. T. Blake, and M. M. Merzenich, *Optimizing sound features for cortical neurons*, Science, 280 (1998), pp. 1439–1443.

[6] J. J. DiCarlo and K. O. Johnson, *Velocity invariance of receptive field structure in somatosensory cortical area 3b of the alert monkey*, J. Neurosci., 19 (1999), pp. 401–419.

[7] J. J. DiCarlo, K. O. Johnson, and S. S. Hsiao, *Structure of receptive fields in area 3b of primary somatosensory cortex in the alert monkey*, J. Neurosci., 18 (1998), pp. 2626–2645.
[8] R. L. Jenison, J. W. H. Schnupp, R. A. Reale, and J. F. Brugge, *Auditory space-time receptive field dynamics revealed by spherical white-noise analysis*, J. Neurosci., 21 (2001), pp. 4408–4415.
[9] J. P. Jones and L. A. Palmer, *An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex*, J. Neurophysiol., 58 (1987), pp. 1233–1258.
[10] S. Marcelja, *Mathematical description of the responses of simple cortical cells*, J. Opt. Soc. Amer., 70 (1980), pp. 1297–1300.
[11] P. N. Marmarelis and V. Z. Marmarelis, *Analysis of Physiological Systems: The White Noise Approach*, Plenum Press, New York, 1978.
[12] D. Q. Nykamp, *Spike correlation measures that eliminate stimulus effects in response to white noise*, J. Comp. Neurosci., 14 (2003), pp. 193–209.
[13] D. Q. Nykamp and D. L. Ringach, *Full identification of a linear-nonlinear system via cross-correlation analysis*, J. Vision, 2 (2002), pp. 1–11.
[14] G. Palm, A. M. H. J. Aertsen, and G. L. Gerstein, *On the significance of correlations among neuronal spike trains*, Biol. Cybern., 59 (1988), pp. 1–11.
[15] R. C. Reid and J. M. Alonso, *Specificity of monosynaptic connections from thalamus to visual cortex*, Nature, 378 (1995), pp. 281–284.
[16] R. C. Reid, J. D. Victor, and R. M. Shapley, *The use of m-sequences in the analysis of visual neurons: Linear receptive field properties*, Vis. Neurosci., 14 (1997), pp. 1015–1027.
[17] D. L. Ringach, *personal communication*.
[18] D. L. Ringach, M. J. Hawken, and R. Shapley, *Dynamics of orientation tuning in macaque primary visual cortex*, Nature, 387 (1997), pp. 281–284.
[19] W. M. Usrey, J.-M. Alonso, and R. C. Reid, *Synaptic interactions between thalamic inputs to simple cells in cat visual cortex*, J. Neurosci., 20 (2000), pp. 5461–5467.

SIAM J. APPL. MATH.
Vol. 63, No. 4, pp. 1231–1258

HYPERBOLIC HOMOGENIZED MODELS FOR THERMAL AND SOLUTAL DISPERSION*

VEMURI BALAKOTAIAH† AND HSUEH-CHIA CHANG‡

Abstract. We formulate a general theory, based on a Lyapunov–Schmidt expansion, for averaging thermal and solutal dispersion phenomena in multiphase reactors, with specific attention to the important Taylor mechanism due to transverse intraphase and interphase capacitance-weighted velocity gradients. We show that the classical Taylor dispersion phenomena are better described in terms of low dimensional models that are hyperbolic and contain an effective local time or length scale in place of the traditional Taylor dispersion coefficient. This description eliminates the use of an artificial exit boundary condition associated with parabolic homogenized equations as well as the classical upstream-feedback and infinite propagation speed anomalies. Our approach is also applicable for describing steady dispersion in the presence of reaction and thermal generation or consumption. For two-phase systems, maximum dispersion is found to exist at an optimum fraction ϵ_f of the lower-capacitance phase. For the disparate phase capacities of most reactors, thermal or solutal dispersion is shown to have the scaling $\frac{\epsilon_f p^2}{(1-\epsilon_f)\Gamma}\alpha_f$, where α_f is the thermal diffusivity of the low-capacitance phase, Γ is the capacitance ratio, and p is the transverse Peclet number.

Key words. solutal dispersion, thermal dispersion, averaging, Liapunov–Schmidt reduction, multiphase reactors

AMS subject classifications. 76R05, 34C29, 34K60, 35B27

PII. S0036139901368863

1. Introduction.

A major goal of the discipline of chemical engineering known as reaction engineering is to combine the complex kinetics, flow fields, and geometries of multiphase reacting systems (such as a packed bed) into accurate low dimensional homogenized convection-diffusion-reaction models that contain all the pertinent transport and kinetic effects of the above complications. It was realized very early that flow turbulence, tortuosity of the interstitial streamlines, velocity gradient of the flowing phase, adsorption onto a stationary phase (as in a chromatograph), and accumulation near stagnation points or stagnant dead zones can give rise to anomalously high solutal dispersion, orders of magnitude higher than molecular diffusion, that must somehow be modeled and included in the homogenized model as a dispersion term $D_{eff}\frac{\partial^2 c}{\partial z^2}$ with the dispersion coefficient D_{eff}. That such a term stipulates that two boundary conditions be provided for the parabolic homogenized model has also introduced considerable confusion. The classical Danckwerts boundary conditions and many other inconsistent ones provide one boundary condition at each end of the reactor (Danckwerts (1953), Choi and Perlmutter (1976), Wehner and Wilhelm (1956)) and have recently been "justified" by Roberts (1989) using center manifold theory. There are, however, fundamental difficulties with this parabolic equation and the Danckwerts boundary conditions. This model introduces infinitely fast diffusive spreading of a

*Received by the editors August 1, 2001; accepted for publication (in revised form) October 1, 2002; published electronically April 9, 2003.
http://www.siam.org/journals/siap/63-4/36886.html
†Department of Chemical Engineering, University of Houston, Houston, TX 77004 (bala@uh.edu). The research of this author was supported by the Texas Advanced Technology Program and the Robert A. Welch Foundation.
‡Department of Chemical Engineering, University of Notre Dame, Notre Dame, IN 46556 (Hsueh-Chia.Chang.2@nd.edu). The research of this author was supported by NSF grants CTS-9522277 and CTS-9980745.

localized concentration perturbation and upstream diffusive propagation (Hinduja, Sundaresan, and Jackson (1980), Sundaresan, Anderson, and Aris (1980)). Both phenomena are not observed experimentally as the flow-induced dispersion mechanisms are hyperbolic in nature (Hiby (1962), Chang (1982)). As a result, the homogenized parabolic equation and Danckwerts boundary conditions cannot describe the observed dispersion phenomena in finite length reactors.

Local turbulent dispersion can be estimated using classical homogeneous turbulent mixing theory. Dispersion in periodic as well as random velocity fields has been reviewed by Majda and Kramer (1999). These authors also present an excellent review of homogenization methods for the convective diffusion equation with periodic velocity fields.

In many applications such as chromatographs and reactors involving packed beds, the flow is laminar, and the more important larger scale dispersion effects that occur over several reactor radii are mostly due to a Taylor–Aris dispersion mechanism (Taylor (1953), Aris (1959), Brenner and Edwards (1993)). This mechanism occurs when a macroscopic transverse velocity gradient, like Poiseuille flow in a tube or macroscopic flow nonuniformity in a packed bed, induces longitudinal dispersion as transverse diffusion lands molecules onto streamlines or flow channels of different velocity. Adsorption onto a stationary solid phase also can trigger this effect as the solid phase has a velocity (zero) different from the flowing phase. Adsorption-induced Taylor–Aris dispersion is responsible for the dispersion of chromatograph signals (Balakotaiah and Chang (1995)). It is also a main problem in biochemical assays on chip-scale laboratories and reactors using microfluidics (Culbertson, Jacobson, and Ramsey (1998)).

Several theories have been developed to predict solutal Taylor–Aris dispersion in packed beds. In the limit of extremely high Peclet number p in an unbounded medium when diffusion is unimportant in the bulk of the flowing phase, Koch and Brady used a diffusive boundary layer cutoff to show that D_{eff} scales as $p \ln p$ (Koch and Brady (1985)). Roberts (1989) and Balakotaiah and Chang (1995) used center manifold theory to show that reaction can affect the dispersion coefficient in a long reactor whose length is much longer than its transverse dimension (radius) such that diffusion dominates in the transverse direction.

Another confusion concerning dispersion is whether a homogenized model remains valid at steady state and whether D_{eff} and its underlying dispersion mechanism are still in play at steady state. While Taylor's classical theory (Taylor, 1953), Koch and Brady's high-p dispersion mechanism in an unbounded medium, and Roberts's and Balakotaiah and Chang's reactive dispersion theory in a long reactor are clearly for transient dispersion, it seems physically intuitive that the same transverse gradient in longitudinal velocity can affect steady-state reactor conversion or performance (Chang, 1982). In fact, it is common practice to use the homogenized model for both steady and transient reactors (Westerterp, Dilman, and Kronberg (1995)). Steady dispersion, however, lacks theoretical justification. An apparent steady dispersion will be shown here to exist, but its description is fundamentally different from that of the transient one.

Even more important than solutal dispersion is thermal dispersion, a subject that is only recently being scrutinized in detail. It is well known in the reaction engineering literature (Balakotaiah (1996), Subramanian and Balakotaiah (1996)) that reactor dynamics and steady-state multiplicity are extremely sensitive to thermal dispersion. Empirical studies and recent analyses have shown that reactor ignition, extinction, hot spot formation, and thermal runaways of most important (and difficult to control) reactors for exothermic reactions are also extremely sensitive to thermal dispersion

(Balakotaiah, Kodra, and Nyugen (1995), Leighton and Chang (1995), and Keith, Leighton, and Chang (1999)). To compound the problem, thermal dispersion is more sensitive to packing and flow geometries and is far more difficult to estimate than solutal dispersion. Two major difficulties are that the thermal penetration depth into the solid phase is deeper than the solutal one, and that the stationary and mobile capacitances are more disparate. As a result, interphase dispersion due to discrepancies in the phase-averaged thermal velocities can enhance and even dominate the intraphase dispersion mechanism in the flowing phase due to transverse flow velocity gradient. The addition of interphase dispersion, distinct from Taylor's intraphase dispersion, renders the analysis more difficult. Vortmeyer and Schaefer (1974), Leighton and Chang (1995), and Balakotaiah and Dommeti (1999) obtained interphase dispersion coefficients based on lumped models with heat transfer coefficients. Leighton and Chang (1995) showed that the ignition location and light-off time of a catalytic converter is determined mostly by this thermal dispersion mechanism. Keith, Leighton, and Chang (1999) used metal inserts to enhance thermal dispersion of a reverse-flow reactor to prevent thermal runaway. Without including the intraphase dispersion mechanism due to flow nonuniformity, they found that an optimum void fraction of intermediate value and with maximum dispersion exists when the heat capacity ratio Γ is near unity, but none exists for realistic void fractions for disparate capacities. In fact, for the more common case of disparate capacities, a generic scaling seems to exist. This would be a significant general result as most reactors have disparate capacities and complex flow fields. It would be desirable to obtain general dispersion scalings insensitive to the flow fields. However, the omission of intraphase dispersion will be shown here to be valid only for disparate capacities. Greatly enhanced dispersion still exists near unit Γ and at an optimum flowing-phase fraction, but the actual dispersion coefficient must include intraphase dispersion and a detailed description of the macroscopic flow fields.

A general solutal/thermal Taylor–Aris dispersion theory will be formulated here to clearly delineate the intraphase and interphase contributions. The proper limit when the former can be omitted and the simpler lumped-phase models can be utilized is also defined. The theory also shows that dispersion phenomena are better described in terms of reduced (low dimensional) models that are hyperbolic in the longitudinal coordinate and time, and with an effective transfer or exchange time constant between the master (slowly varying) mode and slave (local) modes, in contrast to the traditional parabolic models with an effective dispersion coefficient. The reduced models derived based on the present theory eliminate the classical problems of upstream diffusion and infinite propagation speed associated with the parabolic-type averaged equations derived in the prior literature. The theory utilizes the Lyapunov–Schmidt reduction technique of classical bifurcation theory and is based on a perturbation expansion near zero eigenvalue(s).

2. Packed-bed heat transfer. To illustrate the concept of interphase dispersion (due to transfer or exchange between the phases) and some key ideas in our approach, we first consider a very simple model of a packed bed in which the solid is stationary and the fluid moves (Figure 1). The classical heat transfer model of this system ignores the (transverse) gradients within each phase as well as conduction in the axial direction in each phase. The model is described by the following pair of hyperbolic equations for the solid and fluid temperatures:

$$\epsilon_f \, (\rho c_p)_f \left[\frac{\partial T_f}{\partial t'} + u_0 \frac{\partial T_f}{\partial z'} \right] = h a_v (T_s - T_f), \tag{2.1a}$$

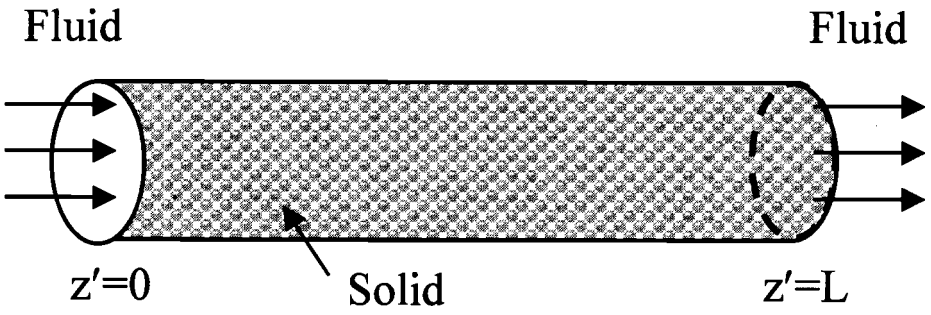


FIG. 1. *Schematic diagram of a packed-bed reactor in which solid and fluid phases interact.*

$$(1-\epsilon_f\ )(\rho c_p)_s \frac{\partial T_s}{\partial t'} = -h a_v (T_s - T_f), \tag{2.1b}$$

with initial and boundary conditions

$$T_f = f_0(t'), \qquad z' = 0,\ t' > 0, \tag{2.1c}$$

$$T_f = T_f^0(z'), \qquad t' = 0,\ z' > 0, \tag{2.1d}$$

$$T_s = T_s^0(z'), \qquad t' = 0,\ z' > 0. \tag{2.1e}$$

Here, u_0 (assumed to be a constant) is the interstitial fluid velocity, ϵ_f is the void fraction of the bed (available for flow), $(\rho c_p)_s$ ($(\rho c_p)_f$) is the solid (fluid) heat capacity per unit volume, h is the interphase heat transfer coefficient, and a_v is the (interphase) transfer area per unit bed volume. Assuming that the bed has a length L, we define

$$z = \frac{z'}{L}, \qquad t = \frac{u_0 t'}{L}, \qquad \Gamma = \frac{(\rho c_p)_s}{(\rho c_p)_f}, \qquad Pe = \frac{(\rho c_p)_f u_0}{L h a_v}, \tag{2.2}$$

and write (2.1a) and (2.1b) in dimensionless form as

$$A \begin{pmatrix} T_f \\ \\ T_s \end{pmatrix} = Pe \begin{pmatrix} \frac{\partial T_f}{\partial t} + \frac{\partial T_f}{\partial z} \\ \\ \frac{\partial T_s}{\partial t} \end{pmatrix}, \tag{2.3a}$$

where the matrix operator A is defined by

$$A = \begin{pmatrix} -\frac{1}{\epsilon_f} & \frac{1}{\epsilon_f} \\ \frac{1}{\Gamma(1-\epsilon_f)} & -\frac{1}{\Gamma(1-\epsilon_f)} \end{pmatrix}. \tag{2.3b}$$

(As shown later, A can be made symmetric by defining an inner product weighted with respect to the relative capacitances of the phases.) We note that the Peclet number Pe is the ratio of interphase transfer time ($\frac{(\rho c_p)_f}{h a_v}$) to the convection time ($\frac{L}{u_0}$), Γ is the ratio of solid to fluid heat capacities, and time is nondimensionalized with respect to the convection time. It is assumed that the Peclet number is small, or equivalently, that the interphase transfer time is much smaller compared to the

convection time. The matrix A is singular with the following null eigenvector and slave eigenvector (with eigenvalue $-\frac{1}{\epsilon_f} - \frac{1}{\Gamma(1-\epsilon_f)}$):

$$\phi_0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad \phi_1 = \begin{pmatrix} \Gamma(1-\epsilon_f) \\ -\epsilon_f \end{pmatrix}. \tag{2.4}$$

Writing

$$\begin{pmatrix} T_f \\ T_s \end{pmatrix} = \begin{pmatrix} 1 & \Gamma(1-\epsilon_f) \\ 1 & -\epsilon_f \end{pmatrix} \begin{pmatrix} T_m \\ T_d \end{pmatrix} \tag{2.5a}$$

or

$$\begin{pmatrix} T_m \\ T_d \end{pmatrix} = \begin{pmatrix} \frac{\epsilon_f}{[\epsilon_f+\Gamma(1-\epsilon_f)]} & \frac{\Gamma(1-\epsilon_f)}{[\epsilon_f+\Gamma(1-\epsilon_f)]} \\ \frac{1}{[\epsilon_f+\Gamma(1-\epsilon_f)]} & \frac{-1}{[\epsilon_f+\Gamma(1-\epsilon_f)]} \end{pmatrix} \begin{pmatrix} T_f \\ T_s \end{pmatrix}, \tag{2.5b}$$

we observe that T_m is the capacitance-weighted (average) temperature, while T_d is the local temperature difference. In terms of these variables, the model may be written as

$$[\epsilon_f + \Gamma(1-\epsilon_f)]\frac{\partial T_m}{\partial t} + \epsilon_f \frac{\partial T_m}{\partial z} = -\epsilon_f(1-\epsilon_f)\Gamma\frac{\partial T_d}{\partial z}, \tag{2.6a}$$

$$T_d = -\frac{Pe\epsilon_f(1-\epsilon_f)\Gamma}{[\epsilon_f+\Gamma(1-\epsilon_f)]^2}\left[\frac{\partial T_m}{\partial z} + \epsilon_f\frac{\partial T_d}{\partial t} + \Gamma(1-\epsilon_f)\left(\frac{\partial T_d}{\partial t} + \frac{\partial T_d}{\partial z}\right)\right]. \tag{2.6b}$$

For $Pe = 0$, the solid and fluid temperatures are in equilibrium, and the average temperature evolves according to (2.6a) with its right-hand side set to zero. For small values of the Peclet number, we have from (2.6b)

$$T_d = -\frac{Pe\epsilon_f(1-\epsilon_f)\Gamma}{[\epsilon_f+\Gamma(1-\epsilon_f)]^2}\left(\frac{\partial T_m}{\partial z}\right) + O(Pe^2). \tag{2.7}$$

Thus, the temperature difference T_d is slaved to the average temperature. Substituting (2.7) into (2.6a) and writing the resulting averaged equation in dimensional form, we obtain

$$\frac{\partial T_m}{\partial t'} + \langle u\rangle \frac{\partial T_m}{\partial z'} = \alpha_{eff}\frac{\partial^2 T_m}{\partial z'^2}, \tag{2.8a}$$

$$\langle u\rangle = u_0\frac{\epsilon_f}{[\epsilon_f+\Gamma(1-\epsilon_f)]}, \qquad \alpha_{eff} = \frac{u_0^2(\rho c_p)_f\ \Gamma^2\ \epsilon_f^2(1-\epsilon_f)^2}{ha_v[\epsilon_f+\Gamma(1-\epsilon_f)]^3}. \tag{2.8b}$$

Here, $\langle u\rangle$ and α_{eff} are the capacitance-weighted velocity and effective thermal diffusivity of the bed, respectively. From the above derivation, it is clear that the reduced model is valid only when $t' \gg \frac{(\rho c_p)_f}{ha_v}$ and $z' \gg u_0\frac{(\rho c_p)_f}{ha_v}$, i.e., when there is a short transient that escapes the effective equation. During this transient, the two phases equilibrate, and the appropriate initial condition for T_m is simply the capacitance-weighted average of the initial conditions of T_s and T_f,

$$T_m(t'=0) = \frac{\epsilon_f\ T_f(t'=0) + \Gamma(1-\epsilon_f)T_s(t'=0)}{[\epsilon_f\ +\Gamma(1-\epsilon_f)]}. \tag{2.8c}$$

This is valid even though $t' = 0$ on the two sides of (2.8c) corresponds to slightly different instants in time. The two boundary conditions for (2.8a) are more problematic, as the original equation only offers one boundary condition at $z' = 0$. Thus, one obvious one (from (2.1b) and (2.1c)) is that

$$T_m(z' = 0, t') = f_0(t') + \frac{(\rho c_p)_s \ \Gamma(1-\epsilon_f)^2}{h a_v [\epsilon_f + \Gamma(1-\epsilon_f)]} \left(\frac{\partial f_0(t')}{\partial t'} \right). \tag{2.8d}$$

In the engineering and the Taylor dispersion theory literature (Danckwerts (1953), Vortmeyer and Schaeffer (1974), Roberts (1992)), the other boundary condition is imposed at the exit $z' = L$, and the often-used exit (Danckwerts) boundary condition is

$$\frac{\partial T_m}{\partial z'}(z' = L, t') = 0. \tag{2.8e}$$

This is clearly not acceptable since the original problem does not possess any boundary condition at the exit. There is also a more fundamental problem associated with the form of the reduced model given by (2.8a). In this form, the reduced model is a parabolic equation, and imposing an artificial exit boundary condition leads to infinite propagation speed for inlet signals. Again, this is certainly not true for the original equations (2.1a) and (2.1b), which may be combined to obtain a single hyperbolic equation for T_f, T_s, or any linear combination of these. For example, without any assumptions on the length or time scales, it is easily seen that the temperature T_i ($i = f, s$, or m) satisfies the hyperbolic equation

$$\frac{\partial T_i}{\partial t'} + \langle u \rangle \frac{\partial T_i}{\partial z'} + \frac{(\rho c_p)_s \ \epsilon_f (1-\epsilon_f)}{h a_v [\epsilon_f + \Gamma(1-\epsilon_f)]} \frac{\partial}{\partial t'} \left[\frac{\partial T_i}{\partial t'} + u_0 \frac{\partial T_i}{\partial z'} \right] = 0. \tag{2.9}$$

For $i = m$, the initial and boundary conditions for (2.9) are the same as those defined by (2.8c) and (2.8d), respectively. Thus, the parabolic form of the reduced equation given by (2.8a) is not preferable as it leads to nonphysical phenomena such as upstream diffusion and infinite speed of propagation. This is certainly not true for the initial model, (2.9), which predicts finite propagation speed for all inlet and initial signals and no upstream diffusion. (This can be seen more clearly by comparing the analytical solutions of the exact and reduced equations for a unit step or impulse inputs. These analytical solutions can be expressed in terms of modified Bessel functions.) The origin of the second spatial derivative term in the reduced equation and the interpretation of the coefficient α_{eff} as an effective (Taylor) diffusivity can be traced back to the paper of Taylor on shear dispersion (Taylor (1953)). We present here an alternate form of the reduced equation (and interpretation of the local coefficients) that eliminates the above-mentioned inconsistencies of the classical Taylor dispersion theory.

We note that, when the interphase transfer time is small, the leading order approximation

$$\frac{\partial T_m}{\partial t'} = - \langle u \rangle \frac{\partial T_m}{\partial z'} + O\left(\frac{(\rho c_p)_f}{h a_v} \right) \tag{2.10}$$

may be used to write the reduced equation in three different forms: as a parabolic equation in z' (i.e., (2.8a)), a parabolic equation in t', or a hyperbolic equation in z' and t'. We also note that the local (2.7) written in terms of either $\frac{\partial T_m}{\partial t'}$ or $\frac{\partial T_m}{\partial z'}$ defines a characteristic time (that is proportional to $\frac{(\rho c_p)_f}{h a_v}$) for heat exchange between the

slowly varying mode T_m and the slave (local) mode T_d. Thus, we write the averaged model as

$$\frac{\partial T_m}{\partial t'} + \langle u \rangle \frac{\partial T_m}{\partial z'} + \langle u \rangle t_H \frac{\partial^2 T_m}{\partial z' \partial t'} = 0, \tag{2.11}$$

where t_H is the characteristic local exchange time (between the two modes) defined by

$$t_H = \frac{(\rho c_p)_s \ (1-\epsilon_f)^2 \Gamma}{h a_v [\epsilon_f + \Gamma(1-\epsilon_f)]}. \tag{2.12}$$

We also define a local length scale as $\ell_H = \langle u \rangle \ t_H$ and note that the reduced model is valid for $z' \gg \ell_H$ and $t' \gg t_H$. In this form, the reduced model defines both the local length and time scales (their ratio being $\langle u \rangle$), and the local effective diffusivity is given by $\alpha_{eff} = \frac{\ell_H^2}{t_H} = \langle u \rangle^2 \ t_H$.

This hyperbolic form of the reduced equation is favored for the following reasons: (i) Since the initial model is hyperbolic, the reduced model should also be hyperbolic; (ii) writing the reduced model as a parabolic equation either in z' or t' requires an artificial boundary or initial condition; (iii) the hyperbolic (2.11) defines a characteristic initial value problem for T_m and hence only T_m needs to be specified along the characteristic curves $z' =$ constant and $t' =$ constant. (In contrast, for the general Cauchy problem, both the function and the normal derivative should be specified along a noncharacteristic curve.) The initial and boundary conditions for (2.11) are the same as those defined by (2.8c) and (2.8d), respectively. Now, no artificial boundary or initial conditions are required, and the reduced model does not lead to any nonphysical phenomena. This hyperbolic form of the reduced model also replaces the concept of an effective (Taylor) diffusivity by that of an effective local time or length scale.

The perturbation expansion can be carried out to higher orders in Pe, and the reduced model (with appropriate initial and boundary conditions) can be expressed in hyperbolic form, but we do not pursue this calculation here. The conditions under which the perturbation expansion converges may also be obtained (for this specific example) in terms of the spatial or time scales appearing in the initial and boundary conditions. We consider it only briefly here and refer to Balakotaiah and Dommeti (1999) for more details.

We note that the local equation (2.6b) may be written as

$$\left[1 + \frac{Pe \ \epsilon_f (1-\epsilon_f)\Gamma}{[\epsilon_f + \Gamma(1-\epsilon_f)]} \left(\frac{\partial}{\partial t}\right)\right] T_d = \frac{Pe \ (1-\epsilon_f)\Gamma}{[\epsilon_f + \Gamma(1-\epsilon_f)]} \left(\frac{\partial T_m}{\partial t}\right). \tag{2.13}$$

Thus, if we consider the special case in which only the inlet conditions are varied, then by taking a Laplace transform, we can reduce the local equation to a linear algebraic equation in terms of the forcing frequency. This equation has a convergent power series expansion in Pe, provided the dimensionless forcing frequency (ω) satisfies the criterion

$$\frac{Pe \ \epsilon_f \ (1-\epsilon_f)\Gamma\omega}{[\epsilon_f + \Gamma(1-\epsilon_f)]} < 1. \tag{2.14}$$

In dimensional terms, (2.14) may be written as

$$\omega' t_H < \frac{(1-\epsilon_f)\Gamma}{\epsilon_f}. \tag{2.15}$$

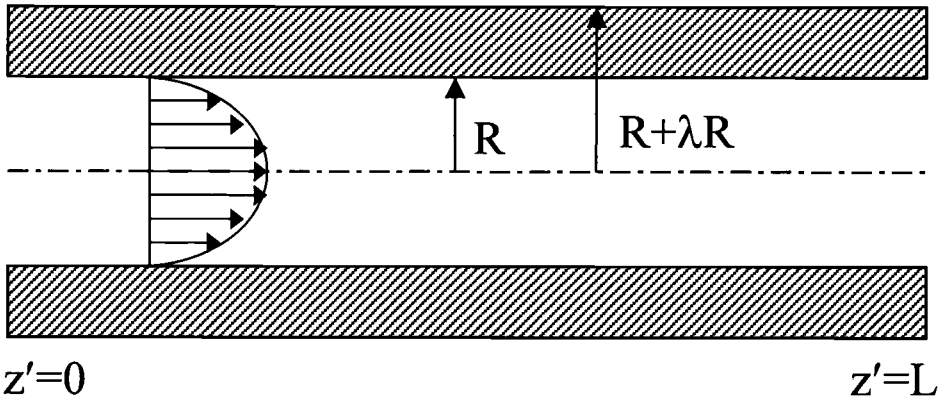


FIG. 2. *Schematic diagram illustrating the classical Taylor solutal dispersion in laminar flow in a circular tube.*

This convergence criterion has a simple physical meaning for the special case in which the volumetric heat capacities of the phases are equal (when the right-hand side of (2.15) is equal to unity): the reduced model exists only if the forcing frequency is less than that defined by the characteristic local exchange time ($\omega' < \frac{1}{t_H}$).

3. Taylor's solutal dispersion theory revisited. In this section, we consider the classical Taylor problem that illustrates intraphase dispersion due to transverse velocity gradients and show that the inconsistencies associated with the parabolic form of the reduced model can be removed by expressing the reduced model in a hyperbolic form. Our approach also shows the similarity between the inter- and intraphase dispersion and the superiority of the hyperbolic models for describing these phenomena.

The dispersion of a nonreactive solute in a circular tube of constant cross section (see Figure 2 for notation) in which the flow is laminar is described by the convective-diffusion equation

$$\frac{\partial C}{\partial t'} + 2\langle u\rangle\left(1-\frac{r^2}{R^2}\right)\frac{\partial C}{\partial z'} = \frac{D}{r}\frac{\partial}{\partial r}\left(r\frac{\partial C}{\partial r}\right),\ 0<r<R,\ z'>0,\ t'>0, \tag{3.1a}$$

$$\frac{\partial C}{\partial r} = 0 @ r = 0, R, \tag{3.1b}$$

$$I.C : C(z', r, 0) = f(z', r), \tag{3.1c}$$

$$B.C : C(0, r, t') = g(r, t'). \tag{3.1d}$$

In writing (3.1a), it is assumed that longitudinal diffusion can be neglected (this assumption is relaxed later on). Here, $\langle u\rangle$ is the average velocity in the pipe, R is the radius, and D is the diffusivity of the species. Defining dimensionless variables

$$z = \frac{z'}{L}, \quad t = \frac{\langle u\rangle t'}{L}, \quad \xi = \frac{r}{R}, \quad Pe = \frac{R^2\langle u\rangle}{LD}, \tag{3.2}$$

we can write (3.1a) and (3.1b) as

$$\pounds C \equiv \frac{1}{\xi}\frac{\partial}{\partial\xi}\left(\xi\frac{\partial C}{\partial\xi}\right) = Pe\left[\frac{\partial C}{\partial t} + 2\left(1-\xi^2\right)\frac{\partial C}{\partial z}\right], \qquad \frac{\partial C}{\partial\xi} = 0 @ \xi = 0, 1. \tag{3.3}$$

We note that the transverse operator $\mathcal{L}$ is symmetric with respect to the inner product

$$(v, w) = \int_0^1 2\xi v(\xi) w(\xi) d\xi.$$

It has a zero eigenvalue with normalized eigenfunction $\phi_0 = 1$. We define the mixing-cup (velocity weighted) and spatial average concentrations by

$$C_m = \int_0^1 4\xi(1-\xi^2) C(\xi, z, t) d\xi, \tag{3.4a}$$

$$\langle C \rangle = \int_0^1 2\xi C(\xi, z, t) d\xi. \tag{3.4b}$$

Transverse averaging of (3.3) gives

$$\frac{\partial \langle C \rangle}{\partial t} + \frac{\partial C_m}{\partial z} = 0. \tag{3.5}$$

We note that when $Pe = 0$, $\langle C \rangle = C_m$, and substitution of this into (3.5) gives the leading order evolution equation for the averaged concentration. Writing

$$C(\xi, z, t) = \langle C \rangle (z, t) + W(\xi, z, t), \qquad W \in \ker\ \mathcal{L}, \tag{3.6}$$

we can solve for the slave variable $W(\xi, z, t)$ in terms of $\langle C \rangle (z, t)$ using a perturbation expansion in Pe (and the Fredholm alternative):

$$\mathcal{L}W - Pe\left[\frac{\partial W}{\partial t} + 2\left(1-\xi^2\right)\frac{\partial W}{\partial z}\right] = Pe\left[\frac{\partial \langle C \rangle}{\partial t} + 2\left(1-\xi^2\right)\frac{\partial \langle C \rangle}{\partial z}\right], \tag{3.7a}$$

$$\frac{\partial W}{\partial \xi} = 0 @ \xi = 0, 1. \tag{3.7b}$$

To leading order, we have

$$W(\xi, z, t) = Pe \frac{\partial \langle C \rangle}{\partial t}\left[\frac{1}{12} - \frac{\xi^2}{4} + \frac{\xi^4}{8}\right] + O(Pe^2). \tag{3.8}$$

Substitution of this into (3.6) and transverse averaging (after multiplying by the velocity profile) gives the local equation relating C_m and $\langle C \rangle$:

$$C_m - \langle C \rangle = \frac{Pe}{48}\frac{\partial \langle C \rangle}{\partial t} + O(Pe^2) = \frac{Pe}{48}\frac{\partial C_m}{\partial t} + O(Pe^2). \tag{3.9}$$

As in the packed-bed problem, this local equation (when written in dimensional form) defines a characteristic transfer time between the slowly evolving mode C_m(or $\langle C \rangle$) and the slave mode $C_m - \langle C \rangle$. Equations (3.5) and (3.9) complete the reduced model to leading order. In this form, the reduced model for intraphase diffusion is similar to the two-mode packed-bed model of interphase diffusion. We can combine the two equations to obtain a single equation for either C_m or $\langle C \rangle$. Since the mixing-cup concentration (which is often measured in experiments) is more relevant in applications, the reduced model in terms of C_m in dimensional form is given by

$$\frac{\partial C_m}{\partial t'} + \langle u \rangle \frac{\partial C_m}{\partial z'} + \langle u \rangle\ t_D \frac{\partial^2 C_m}{\partial z' \partial t'} = 0, \qquad t' \gg t_D,\ z' \gg \ell_D, \tag{3.10}$$

where the local diffusion or mixing time is defined by

$$t_D = \frac{R^2}{48D}. \tag{3.11}$$

The corresponding length scale and local diffusivity are given by $\ell_D = \langle u \rangle\, t_D$, $D_{eff} = \langle u \rangle^2 t_D$. As noted earlier, in the Taylor dispersion literature, (3.10) is written as a parabolic equation with an effective dispersion coefficient D_{eff}, which requires an artificial boundary condition at the exit of the tube (Roberts (1992)). Below we present a solution of (3.10) for general inlet and initial conditions and show that it can describe dispersion for long times as well as the parabolic model. However, unlike the classical parabolic equation over an infinite domain, (3.10) can accommodate an inlet boundary condition. Once again, since (3.10) defines a characteristic initial value problem, to complete the model, we need to specify C_m only along the characteristic curves $z' = 0$ and $t' = 0$. Thus, the initial and boundary conditions for the reduced model are obtained by taking the mixing-cup averages of (13.c) and (13.d):

$$C_m(z', t' = 0) = \int_0^1 4\xi(1-\xi^2)\ f(z', R\xi)d\xi \equiv f_m(z'), \tag{3.12a}$$

$$C_m(z' = 0, t') = \int_0^1 4\xi(1-\xi^2)\ g(R\xi, t')d\xi \equiv g_m(t'). \tag{3.12b}$$

Equations (3.10) and (3.12) complete the hyperbolic model to order Pe. As in the packed-bed example, the perturbation expansion can be carried out to higher orders, and it can be shown that it converges, provided $t_D\omega_t' < 0.858$ and $\ell_D\omega_z' < 0.288$, where $\omega_t'(\omega_z')$ is the temporal (spatial) frequency contained in the inlet or initial conditions. (For details, see Chakraborty and Balakotaiah (2002), Balakotaiah and Chang (1995), and Mercer and Roberts (1990).)

The above analysis can be extended to the general case in which axial diffusion is included in (13.a). In this case, the reduced model may be shown to be

$$\frac{\partial \langle C \rangle}{\partial t'} + \langle u \rangle \frac{\partial C_m}{\partial z'} = D\frac{\partial^2 \langle C \rangle}{\partial z'^2}, \tag{3.13a}$$

$$\langle C \rangle - C_m = -\ t_D \frac{\partial \langle C \rangle}{\partial t'}. \tag{3.13b}$$

We can combine these equations to obtain a single hyperbolic equation for $\langle C \rangle$:

$$\frac{\partial \langle C \rangle}{\partial t'} + \langle u \rangle \frac{\partial \langle C \rangle}{\partial z'} + \langle u \rangle\ t_D \frac{\partial^2 \langle C \rangle}{\partial z' \partial t'} = D\frac{\partial^2 \langle C \rangle}{\partial z'^2}. \tag{3.14}$$

(We note that C_m or any other weighted average concentration also satisfies the same equation (3.14). This is due to the fact that the original conservation equation is linear in the concentration.) We note that when $D \ll \langle u \rangle^2\ t_D$, or equivalently, the radial Peclet number $p = \frac{\langle u \rangle\ R}{D} \gg 6.93$, axial diffusion can be neglected. (Note that the perturbation Peclet number Pe, which is equal to p times the aspect ratio $\left(\frac{R}{L}\right)$, can be small even when $p \gg 6.93$, provided that the aspect ratio is sufficiently small. The conditions $p \gg 6.93$ and $Pe \ll 1$ are usually satisfied for well-designed reactors or chromatographic columns.)

3.1. Solution of the hyperbolic model. In this section, we present the solution of the hyperbolic model defined by (3.10) and (3.12) and compare these solutions to those of the classical parabolic model. We use the local time and length scales to nondimensionalize the variables and write the hyperbolic model in the following form:

$$\frac{\partial C_m}{\partial t} + \frac{\partial C_m}{\partial z} + \frac{\partial^2 C_m}{\partial z \partial t} = 0, \qquad t \gg 1,\ z \gg 1, \tag{3.15a}$$

$$C_m(z, t = 0) = f(z), \tag{3.15b}$$

$$C_m(z = 0, t) = g(t). \tag{3.15c}$$

(With this scaling, the reciprocal of the nondimensional time is the Peclet number.) The substitution

$$C_m = W \exp(-z - t) \tag{3.16a}$$

reduces (3.15a) to the canonical form

$$\frac{\partial^2 W}{\partial z \partial t} - W = 0. \tag{3.16b}$$

The fundamental solution (Riemann function) of (3.16b) is given by (see Garabedian (1964))

$$W_g(z, t, \xi, \eta) = I_0\left(2\sqrt{(z - \xi)(t - \eta)}\right), \tag{3.16c}$$

where I_0 is the modified Bessel function of order zero. Using this fundamental solution, we may express the solution of (3.15) as

$$\begin{aligned} C_m(z, t) = e^{-z-t} \Bigg\{ & c_0 I_0(2\sqrt{zt}) + \int_0^z \frac{dF(\xi)}{d\xi} I_0\left(2\sqrt{t(z - \xi)}\right) d\xi \\ & + \int_0^t \frac{dG(\eta)}{d\eta} I_0\left(2\sqrt{z(t - \eta)}\right) d\eta) \Bigg\}, \end{aligned} \tag{3.17a}$$

where

$$F(z) = e^z f(z), \tag{3.17b}$$

$$G(t) = e^t g(t), \tag{3.17c}$$

$$c_0 = \frac{f(0) + g(0)}{2}. \tag{3.17d}$$

Below, we use this analytical solution to show that the solution of the hyperbolic model, (3.15a)–(3.15c), remains positive for arbitrary but positive inlet and initial conditions. We also compare the solutions of the hyperbolic model with those of the parabolic model for some special cases.

3.2. Positivity of the solutions to the hyperbolic model. We note that (3.15) as well as the general solution given by (3.17) have the permutational symmetry in z and t; i.e., they are invariant to the transformation $(z,t,f) \to (t,z,g)$. Thus, to prove the positivity of the solution of (3.15), it is sufficient to consider the case of $g(t) = 0$ and $f(z) \neq 0$. Now, for the special case of $f(z) = \delta(z - z_0)$ and $g(t) = 0$, the general solution given by (3.17) simplifies to

$$C_{mg}(z,z_0,t) = \begin{cases} I_1\left(2\sqrt{t(z-z_0)}\right) e^{z_0-z-t}\sqrt{\frac{t}{z-z_0}}, & z > z_0, \\ 0, & z < z_0. \end{cases} \tag{3.18}$$

Since this Green's function is positive, the solution given by (3.17) remains positive for all positive inlet and initial conditions. In fact, an alternate form of the analytical solution to (3.15) makes this obvious:

$$C_m(z,t) = \begin{cases} \int_0^z I_1\left(2\sqrt{t(z-z_0)}\right) e^{z_0-z-t}\sqrt{\frac{t}{z-z_0}} f(z_0)dz_0 \\ \qquad + \int_0^t I_1\left(2\sqrt{z(t-t_0)}\right) e^{t_0-t-z}\sqrt{\frac{z}{t-t_0}} g(t_0)dt_0, & z > 0,\ t > 0, \\ 0, & z < 0 \text{ or } t < 0. \end{cases} \tag{3.19}$$

3.3. Comparison of the dispersion curves for parabolic and hyperbolic models. As stated earlier, it is of interest to determine how the solution of the hyperbolic model differs from that of the parabolic models used in the literature to describe solutal dispersion in nonreacting systems. In the engineering literature, the solution of the averaged model to a unit impulse (Delta function) input is known as the dispersion curve. For the parabolic model, this is the standard Gaussian curve given by

$$E_p(z,t) = \frac{1}{\sqrt{4\pi t}} \exp\left\{-\frac{(z-t)^2}{4t}\right\}. \tag{3.20}$$

Thus, the parabolic model predicts a peak in the dispersion curve at $z = t$ and a variance that increases linearly with time. The dispersion curve for the hyperbolic model is obtained by taking $z_0 = 0$ in (3.18):

$$E_h(z,t) = \begin{cases} e^{-z-t}\sqrt{\frac{t}{z}} I_1(2\sqrt{tz}), & z > 0, \\ 0, & z < 0. \end{cases} \tag{3.21}$$

Examination of this curve shows that it has a peak at $z = 0$ for $t \leq 2$. This is consistent with physical observation that, for short times, transverse diffusion has not acted on the initial delta function input and hence the peak should be at the injection point. For $t \gg 1$, (3.21) may be written as

$$E_h(z,t) \approx \left(\frac{t}{z}\right)^{\frac{3}{4}} \frac{1}{\sqrt{4\pi t}} \exp\left\{-\frac{(z-t)^2}{(\sqrt{t}+\sqrt{z})^2}\right\}. \tag{3.22}$$

Thus, the dispersion curves predicted by the two models are close to each other near $t = z$ (and $t \gg 1$), but the hyperbolic model predicts an asymmetric curve with a slightly higher peak at $z = t - \frac{3}{2}$. In addition, as noted earlier, the parabolic model predicts upstream diffusion (since $E_p(z,t)$ is not zero for $z < 0$) and infinite propagation speed. Neither of these nonphysical phenomena is present in the hyperbolic

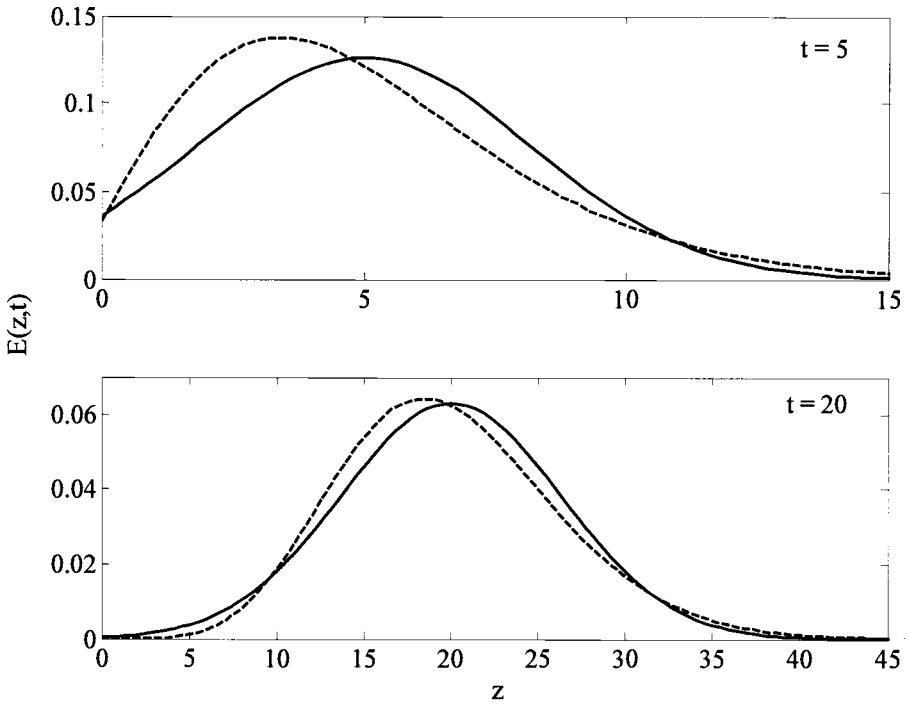


FIG. 3. *Comparison of the dispersion curves predicted by the parabolic and hyperbolic models at* $t = 5$ *and* 20. *The symmetric curve with peak at* $t = z$ *is for the parabolic model, while the asymmetric curve with peak at* $z \approx t - \frac{3}{2}$ *is for the hyperbolic model.*

model. Figure 3 compares the two solutions at two different times, $t = 5$ and $t = 20$. While the two curves are extraordinarily close for large times, they intersect three times (for all $t > 4.84$), and the dispersion curve predicted by the hyperbolic model has a nonzero skewness at all finite times. We note that this skewness can also be predicted by the parabolic-type models as done by Chatwin (1970), but higher order terms (like $\frac{\partial^3 C_m}{\partial z^3}$) have to be included in the perturbation expansion. The hyperbolic model captures the asymmetry at the lowest order.

4. General thermal/solutal dispersion theory. We now extend the theory to the general case of a multiphase system in which the individual phases may be stationary or moving and the capacitance varies with transverse coordinates; i.e., dispersion is due to combined inter- and intraphase mechanisms. We assume a long reactor with weak longitudinal variation of temperature or concentration (due to the small aspect ratio) and strong longitudinal (laminar) flow $w(x, y)$. We retain only transverse molecular solutal or thermal diffusion in x and y. Conversely, only longitudinal convection is appreciable to balance transverse diffusion. The general governing equation for the reactor temperature or concentration in dimensionless form is then

$$F(\theta, Pe) \equiv \frac{1}{\rho c_p} \nabla \cdot k \nabla \theta - Pe \left(\frac{\partial \theta}{\partial t} + w \frac{\partial \theta}{\partial z} \right) = 0, \tag{4.1}$$

where the ∇ operator is only for the transverse direction $\nabla = (\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, 0)^T$ and both the dimensionless conductivity k and the heat or solutal capacity ρc_p are functions of the transverse coordinates x and y within the transverse cross section Ω. The Peclet number in (4.1) is again defined as the ratio of the transverse diffusion or local exchange time ($\frac{R^2}{\alpha_0}$) to the convection time ($\frac{L}{u_0}$). (We note here that the thermal conductivity, heat capacitance per unit volume, and the velocity are nondimensionalized using some reference values k_0, $(\rho c_p)_0$, and u_0, which can be chosen conveniently for each application. Here, $Pe = \frac{R^2 u_0}{L\alpha_0}$, where $\alpha_0 = \frac{k_0}{(\rho c_p)_0}$.) When $\rho c_p = 1$ is uniform throughout Ω, the transport problem (4.1) reduces to mass transport with uniform capacitance. However, the solutal diffusivity can vary from phase to phase and is hence a function of (x, y). Similarly, the transport coefficients w and k do not have to be constant within each phase but can vary continuously with respect to (x, y).

Equation (4.1) must be solved in the transverse direction with continuity in θ and $k\frac{\partial \theta}{\partial n}$ at the phase boundaries. We shall also impose a no-flux boundary condition at the transverse reactor boundary $\partial\Omega$,

$$\left.\frac{\partial \theta}{\partial n}\right|_{\partial\Omega} = 0. \tag{4.2}$$

It is then clear that, at $Pe = 0$, a particular solution to (4.1) is $\theta = \langle\theta\rangle$, the capacitance-weighted transverse average of θ, independent of the transverse coordinates x and y. We then seek correction to $\langle\theta\rangle$ for small Pe. To leading order, one obtains the linear operator in Ω,

$$D_\theta F(\langle\theta\rangle, 0) \cdot v = \frac{1}{\rho c_p}\nabla \cdot k\nabla v \equiv \pounds v, \tag{4.3}$$

with Neumann boundary conditions on the outer boundary $\partial\Omega$, $\frac{\partial v}{\partial n}|_{\partial\Omega} = 0$. This transverse diffusion operator contains a transverse-dependent conductivity k and capacitance ρc_p that change from phase to phase. The operator is self-adjoint with respect to the inner product

$$[u, v] = \frac{1}{\Omega_{cp}} \int_\Omega \rho c_p \; u \; v \; d\Omega, \tag{4.4a}$$

$$\Omega_{cp} = \int_\Omega \rho c_p \; d\Omega. \tag{4.4b}$$

It has a zero eigenvalue with a constant null eigenfunction ϕ_0, which can be chosen as unity. The other eigenfunctions $\phi_n(x, y)$ all have zero integral (transverse average) from a simple application of divergence theorem to $\pounds\phi_n = -\lambda_n\phi_n$ with the Neumann boundary condition to $\partial\Omega$:

$$[\phi_n, \phi_0] = \langle\phi_n\rangle = \frac{1}{\Omega_{cp}} \int_\Omega \rho c_p \phi_n d\Omega = 0. \tag{4.5}$$

(We point out that (4.5) defines the capacitance-weighted transverse average of any function. We also define $\langle\rho c_p\rangle = \Omega_{cp}/|\Omega|$.)

We expand θ in terms of $\{\phi_n\}_{n=0}^{\infty}$,

$$\theta = \langle\theta\rangle + \theta'(x, y, z, t), \tag{4.6}$$

where $\langle\theta\rangle$ represents the null eigenfunction ϕ_0 component without (x, y)-dependence, while $\theta' \perp \ker \pounds$ represents the complement spanned by the other eigenfunctions.

Since $\mathcal{L}$ is self-adjoint with respect to the Hilbert inner product and $\theta' \perp \ker \mathcal{L}$, we have

$$\langle \theta' \rangle = 0. \tag{4.7}$$

In the terminology of bifurcation theory, the equation satisfied by $\langle\theta\rangle$ is the so-called branching equation at a simple zero eigenvalue. (In the engineering literature, this is often referred to as the homogenized equation, reduced model, averaged equation, pseudohomogeneous model, etc.) This reduced equation can be obtained by applying the implicit function theorem to eliminate θ' from the equation $EF(\langle\theta\rangle + \theta', Pe) = 0$. (Here, E is the projection operator onto *range* $\mathcal{L}$.) This Lyapunov–Schmidt reduction can be done by expanding θ' in terms of the small parameter Pe and solving a linear equation at each order by using the Fredholm alternative. Writing

$$\theta' = Pe\theta_1 + Pe^2\theta_2 + \cdots, \tag{4.8}$$

the leading order equation is given by

$$\mathcal{L}\theta_1 - \left(\frac{\partial \langle\theta\rangle}{\partial t} + w \frac{\partial \langle\theta\rangle}{\partial z} \right) = 0. \tag{4.9}$$

Before we solve for θ' in terms of $\langle\theta\rangle$ to obtain the reduced model, we invoke a unique relationship between the capacitance-weighted average $\langle\theta\rangle$ and the mixing-cup average defined by

$$\theta_m = \frac{1}{\Omega_{cp}} \int_\Omega \rho c_p \; w \; \theta \; d\Omega = \frac{[w\theta, \phi_0]}{[w, \phi_0]} = \frac{\langle w\theta \rangle}{\langle w \rangle}, \tag{4.10a}$$

where $\langle w \rangle$ is the capacitance-weighted average dimensionless velocity, i.e.,

$$\langle w \rangle = \frac{1}{\Omega_{cp}} \int_\Omega \rho c_p \; w \; d\Omega. \tag{4.10b}$$

We note that, only in the degenerate case in which w is uniform over Ω, $\theta_m = \langle\theta\rangle$. In all other cases, $\theta_m \neq \langle\theta\rangle$. Integrating (4.1) over the transverse cross-section Ω and invoking the Neumann condition (4.2), one obtains

$$\frac{\partial \langle\theta\rangle}{\partial t} + \langle w \rangle \frac{\partial \theta_m}{\partial z} = 0. \tag{4.11}$$

This relationship is exact and is valid for all orders in Pe. A second relation between $\langle\theta\rangle$ and θ_m may be obtained by multiplying (4.6) by w and taking the inner product with the null eigenfunction ϕ_0:

$$\theta_m = \langle\theta\rangle + Pe \; \frac{[w\theta_1, \phi_0]}{\langle w \rangle} + O(Pe^2). \tag{4.12}$$

Thus, to leading order, the mixing-cup and capacitance-weighted average temperatures are equal, and the evolution equation (4.11) for $\langle\theta\rangle$ reduces to

$$\frac{\partial \langle\theta\rangle}{\partial t} + \langle w \rangle \frac{\partial \langle\theta\rangle}{\partial z} + O(Pe) = 0. \tag{4.13}$$

To obtain the evolution equation to order Pe, we insert (4.13) into (4.9) and define η as

$$\theta_1 \sim \eta \frac{\partial \langle \theta \rangle}{\partial z} \tag{4.14}$$

to reduce (4.9) to the following convenient form:

$$\pounds \eta = w - \langle w \rangle , \tag{4.15a}$$

with no-flux condition at $\partial\Omega$

$$\left. \frac{\partial \eta}{\partial n} \right|_{\partial\Omega} = 0 \tag{4.15b}$$

and the usual continuity of η and $k\frac{d\eta}{dn}$ at the phase boundaries within Ω. Substitution of (4.14) into (4.11) and (4.12) gives the reduced model

$$\frac{\partial \theta_m}{\partial t} + \langle w \rangle \frac{\partial \theta_m}{\partial z} + Pe\Lambda \frac{\partial^2 \theta_m}{\partial z \partial t} = 0, \tag{4.16}$$

where the numerical coefficient Λ is given by

$$\Lambda = - \frac{[w\eta, \phi_0]}{\langle w \rangle} = - \frac{\langle w\eta \rangle}{\langle w \rangle}. \tag{4.17}$$

In dimensional form, (4.16) becomes

$$\frac{\partial \theta_m}{\partial t'} + \langle u \rangle \frac{\partial \theta_m}{\partial z} + \ell_H \frac{\partial^2 \theta_m}{\partial z' \partial t'} = 0, \tag{4.18}$$

where $\langle u \rangle$ is the capacitance-weighted average velocity and the effective local length scale ℓ_H is defined by

$$\ell_H = \Lambda \frac{R^2 u_0}{\alpha_0}. \tag{4.19}$$

The corresponding time scale and the dimensional effective dispersion coefficient are given by $t_H = \frac{\ell_H}{\langle u \rangle}$ and $\alpha_{eff} = \frac{\ell_H^2}{t_H} = \Lambda \frac{R^2 u_0 \langle u \rangle}{\alpha_0}$, respectively. Again, we note that (4.18) is valid only for $t' \gg t_H$ and $z' \gg \ell_H$. Dispersion can still occur for $0 < t' < t_H$ or $0 < z' < \ell_H$ or even when w is uniform in Ω. However, in these higher order cases, the expansion must be carried to higher orders in Pe so that the averaged model can capture this early dispersion. Similarly, as pointed out by Mercer and Roberts (1990) and Young and Jones (1991), early dispersion effects due to point sources or sinks can be captured at order Pe^2 and higher. We shall not discuss these higher order special cases here as they are not very common in reactors and are also specific to each problem.

The initial and boundary conditions for the general equation (4.18) may be derived in the same manner as for the two specific examples illustrated earlier. We now consider the local equation and various special cases of thermal and solutal dispersion.

4.1. The local equation. As noted above, the local equation (4.15) must be solved before we can determine the numerical coefficient Λ and hence the effective local length or time scales. It is clear that the inhomogeneous term on the right-hand side of (4.15a) satisfies the solvability condition for the singular operator. Also, since $\eta \perp \ker \mathcal{L}$,

$$\langle \eta \rangle = 0, \tag{4.20}$$

which can be used to solve for η uniquely.

Multiplying (4.15a) by η, taking the inner product with the null eigenfunction, and using (4.20), one obtains

$$\begin{aligned} \Lambda &= -[\eta w, \phi_0] \\ &= -[\eta \mathcal{L} \eta, \phi_0] \\ &= -\tfrac{1}{\Omega_{cp}} \textstyle\int_\Omega \eta \nabla \cdot k \nabla \eta \, d\Omega. \end{aligned} \tag{4.21}$$

Using (4.15b) and the divergence theorem (integration by parts), (4.21) then yields

$$\Lambda = \frac{1}{\Omega_{cp}} \int_\Omega k \nabla \eta \cdot \nabla \eta \, d\Omega, \tag{4.22}$$

which is always positive. This quadratic form for Λ is more convenient for some of our derivations and could possibly be used in a variational numerical scheme for η.

In the case of single-phase thermal transport or multiphase solutal transport when ρc_p is uniform throughout Ω, (4.22) simplifies to

$$\Lambda = \frac{1}{|\Omega|} \int_\Omega \alpha \nabla \eta \cdot \nabla \eta \, d\Omega, \tag{4.23}$$

where α is the dimensionless local diffusivity. If, in addition, α is independent of the transverse coordinates (x, y) and is normalized to unity, then Λ depends on only the velocity profile and the geometry of Ω. This clearly shows that, for the single-phase limit and for uniform-capacitance solutal transport, dispersion occurs when a transverse gradient in the longitudinal velocity exists.

The interphase dispersion seen in the packed-bed example arises from the different thermal or capacitance-weighted velocities of each phase and hence contributes to dispersion even if the longitudinal velocity w_j is gradientless within each phase j. This limit is particularly interesting, as the local equation has the following algebraic form for each phase Ω_j in Ω:

$$\sum_j A_{ij} \, \eta_j = (w_i - \langle w \rangle), \tag{4.24a}$$

where A is the matrix defining the coupling between the phases. The numerical coefficient Λ takes the simple and explicit form

$$\Lambda = \frac{\sum_j w_j \eta_j \, (\rho c_p)_j \;\; \epsilon_j}{\sum_j \, (\rho c_p)_j \;\; \epsilon_j}, \tag{4.24b}$$

where ϵ_i is the volume fraction of phase i. For the two-phase example of section 2, with A defined by (2.3b) and $w_1 = 1, w_2 = 0$, we have

$$\begin{aligned} \eta &= \frac{\epsilon_f (1 - \epsilon_f) \Gamma}{[\epsilon_f + (1 - \epsilon_f)\Gamma]^2} \begin{pmatrix} -(1 - \epsilon_f)\Gamma \\ \epsilon_f \end{pmatrix}, \\ \Lambda &= \frac{-w_1 \eta_1 \epsilon_f}{[\epsilon_f + (1 - \epsilon_f)\Gamma]} = \frac{\epsilon_f^2 (1 - \epsilon_f)^2 \Gamma^2}{[\epsilon_f + (1 - \epsilon_f)\Gamma]^3}. \end{aligned} \tag{4.25}$$

This interphase dispersion mechanism arises as solutal or heat "molecules" are transported to different phases through transverse thermal random walks. Once arrived, molecules at each phase are caused to propagate longitudinally by different thermal or capacitance-weighted velocities. The proportion of molecules in each phase is determined by the local diffusivity and the size of Ω_i. Hence, the capacitance-weighted longitudinal phase thermal velocities (as defined by (4.24)) simply yield the thermal dispersion.

4.2. Thermal/solutal dispersion with diffusion into the wall. We next examine thermal and solutal dispersion in a cylindrical pipe of radius R with a solid wall of thickness λR ($\lambda > 0$); see Figure 2. The diffusivity in the wall (α_s) is assumed to be distinct from that in the fluid phase. Here, we take the pipe radius (R), the average fluid velocity ($\langle u_f \rangle$), the fluid heat capacitance $(\rho c_p)_f$, and fluid thermal diffusivity (α_f) to nondimensionalize the variables. The velocity field is now

$$w(\xi) = \begin{cases} 2\left[1-\xi^2\right] & \text{in} \quad \Omega_f,\ 0<\xi<1, \\ 0 & \text{in}\ \ \Omega_s,\ 1<\xi<1+\lambda, \end{cases} \tag{4.26}$$

with

$$\langle w \rangle = \frac{\epsilon_f}{\epsilon_f + \Gamma(1-\epsilon_f)},$$

$$\epsilon_f = \frac{1}{(1+\lambda)^2}.$$

The local equation is now

$$\frac{1}{\xi}\frac{\partial}{\partial \xi}\left(\xi \frac{\partial \eta_f}{\partial \xi}\right) = w - \langle w \rangle \quad \text{in } \Omega_f, \tag{4.27a}$$

$$\frac{1}{\mu}\frac{1}{\xi}\frac{\partial}{\partial \xi}\left(\xi \frac{\partial \eta_s}{\partial \xi}\right) = w - \langle w \rangle \quad \text{in } \Omega_s, \tag{4.27b}$$

with boundary conditions

$$\frac{\partial \eta_f}{\partial \xi} = 0, \quad \xi = 0, \tag{4.28a}$$

$$\frac{\partial \eta_s}{\partial \xi} = 0, \quad \xi = (1+\lambda), \tag{4.28b}$$

$$\eta_f = \eta_s, \quad \xi = 1, \tag{4.28c}$$

$$\frac{\Gamma}{\mu}\frac{\partial \eta_s}{\partial \xi} = \frac{\partial \eta_f}{\partial \xi}, \quad \xi = 1, \tag{4.28d}$$

where $\mu = \alpha_f/\alpha_s$ is the ratio between the diffusivities.

Some algebraic manipulation immediately yields

$$\eta_f = \left\{ \frac{1}{2}\xi^2 - \frac{1}{8}\xi^4 - \frac{\langle w \rangle}{4}\xi^2 + s_f \right\}, \tag{4.29a}$$

$$s_f = -\frac{5}{24}\epsilon_f - \frac{3}{8}(1-\epsilon_f) - \langle w \rangle \left\{ (\mu - 1)\frac{(1-\epsilon_f)}{4} - \frac{\epsilon_f}{8} + \frac{\mu}{2} f(\epsilon_f) \right\}, \tag{4.29b}$$

$$f(\epsilon_f) = \frac{\epsilon_f^2 + 2\epsilon_f - 3 - 2ln\epsilon_f}{4\epsilon_f}, \tag{4.29c}$$

$$\eta_s = \left(\frac{\mu}{2}\right) \langle w \rangle \left\{ (1+\lambda)^2 \ln(\xi) - \frac{1}{2}\xi^2 + s_s \right\}, \tag{4.29d}$$

$$s_s = \frac{1}{2} + \frac{s_f + \frac{3}{8} - \frac{1}{4}\langle w \rangle}{\frac{\mu}{2}\langle w \rangle}. \tag{4.29e}$$

Substituting into (4.17), we obtain

$$\Lambda = \frac{\beta_1 \epsilon_f}{\epsilon_f + \Gamma(1-\epsilon_f)} + \frac{(\Gamma - 1)\beta_2 \epsilon_f^2 (1-\epsilon_f)}{[\epsilon_f + \Gamma(1-\epsilon_f)]^2}, \tag{4.30}$$

$$\begin{aligned} \beta_1 = {} & \frac{11 - 8\epsilon_f}{48} + \left(\frac{\epsilon_f}{\epsilon_f + \Gamma(1-\epsilon_f)} \right) \left(\frac{6\epsilon_f - 8}{48} \right) \\ & + \frac{\mu}{4} \left(\frac{\epsilon_f}{\epsilon_f + \Gamma(1-\epsilon_f)} \right) (2f(\epsilon_f) + 1 - \epsilon_f), \end{aligned} \tag{4.31a}$$

$$\beta_2 = \frac{1}{6} + \frac{-\epsilon_f(1-\epsilon_f) + \mu(4\epsilon_f - \epsilon_f^2 - 3 - 2ln\epsilon_f)}{8(1-\epsilon_f)(\epsilon_f + \Gamma(1-\epsilon_f))}. \tag{4.31b}$$

Several limits are of interest. As $\epsilon_f \to 1$, we obtain Taylor's result $\Lambda = \frac{1}{48}$ for solutal dispersion. In the limit of $\Gamma(1-\epsilon_f) \gg \epsilon_f$, for intermediate ϵ_f values away from unity we obtain

$$\Lambda = \frac{\alpha_{eff}}{p^2 \alpha_f} = \frac{11}{48} \frac{\epsilon_f}{\Gamma(1-\epsilon_f)}, \tag{4.32}$$

where $p = \langle u_f \rangle R/\alpha_f$ is the transverse Peclet number. The coefficient 11/48 is also consistent with Leighton and Chang (1995), using the lumped-phase approach. (Note that $\frac{48}{11}$ is the asymptotic Nusselt number, Nu_∞, for laminar flow in a tube with a constant flux boundary condition on the tube wall.) In Figure 4 we plot the product of Γ and the normalized dispersion coefficient Λ given by (4.30) for Γ values of 100 and 1000, with μ varying from 10^{-1} to 10^2. It is obvious that all of them are independent of μ and collapse nicely into the specific scaling of (4.32) except near $\epsilon_f = 1$, where (4.32) is singular. Thus, the high-Γ limit yields a scaling that is insensitive to geometric and

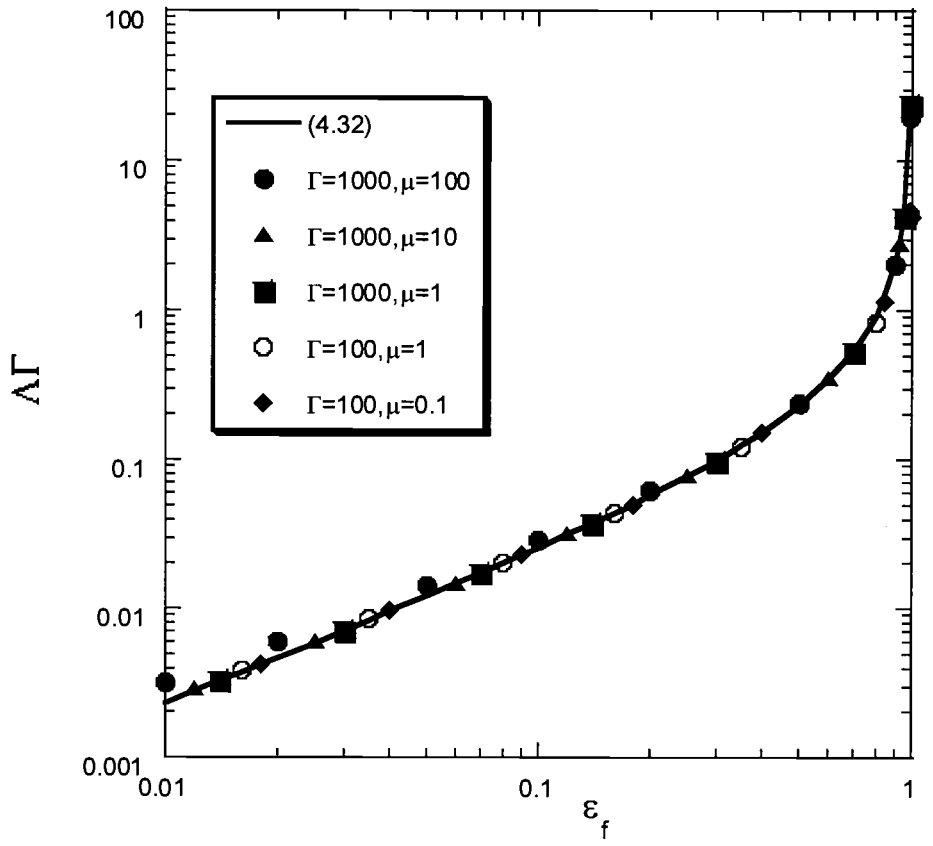


FIG. 4. *High-*Γ *dispersion for various values of diffusivity (*μ*) and capacitance ratio (*Γ*). The curves for various* μ *and* Γ *collapse nicely into the* μ*-independent limit of* (4.32).

flow details and is independent of α_s. The dispersion is small in this limit at $O(\Gamma^{-1})$ and increases sharply with ϵ_f.

For $\epsilon_f \ll 1$ and Γ of unit order, one can deduce from (4.30) that the flowfield-sensitive intraphase term with the β_1 coefficient dominates over the other interphase term. A simple calculation then yields

$$\frac{\alpha_{eff}}{p^2\alpha_f} = \left(\frac{\mu}{4\Gamma^2}\right)\epsilon_f \ln\left(\frac{1}{\epsilon_f}\right), \tag{4.33}$$

with a different Γ scaling from (4.32).

For ϵ_f close to unity, the interphase term becomes equally important as the intraphase term. The high effective dispersion near $\Gamma(1-\epsilon_f) \sim \epsilon_f$ is verified in Table 1, where we have used (4.30) to determine the optimum ϵ_f for various μ and Γ values. Figure 5 shows a plot of the maximum value of the normalized dispersion coefficient that can be obtained for the optimum volume fraction ϵ_f of the low capacitance phase. All these curves approach an asymptotic value for $\Gamma > 10$. This asymptote may be found from (4.30) in the limit of $\Gamma \gg 1$ and $\delta = \Gamma(1-\epsilon_f)$ finite. This simplification

TABLE 1
Optimum fraction of the low capacitance phase at which the dispersion coefficient is maximum.

Γ	$\varepsilon_{\max}$		
	$\mu = 0.1$	$\mu = 1$	$\mu = 10$
1	0.45	0.31	0.16
2	0.63	0.59	0.30
5	0.81	0.81	0.74
10	0.90	0.90	0.89
20	0.95	0.95	0.94
50	0.98	0.98	0.98

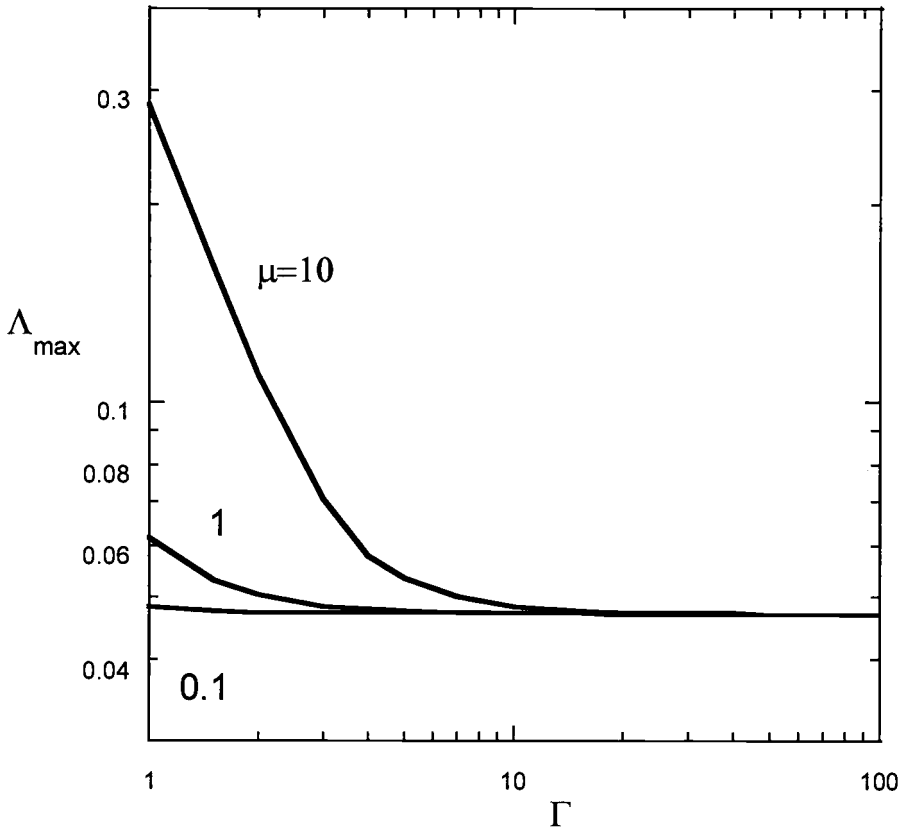


FIG. 5. *Maximum dispersion coefficient obtained at an optimum ϵ_f as a function of Γ for various μ values.*

gives

$$\Lambda = \frac{1 + 6\delta + 11\delta^2}{48(1+\delta)^3}. \tag{4.34}$$

This is the expression derived by Golay (1958) for capillary chromatography with a retentive layer. It follows from (4.34) that $\Lambda_{\max} = 0.047$ at $\delta = 1.15$. As is evident from Figure 5, the optimum normalized dispersion $\Lambda_{\max}$ relative to ϵ_f becomes independent of Γ and μ for Γ beyond 10, as is consistent with (4.30). However, this

optimum dispersion is highest for Γ below 10, near unit order Γ. Hence, the largest dispersion occurs near unit order Γ with a dispersion magnitude that is 3 to 4 times the high Γ limit. Unfortunately, we cannot use the lumped model approach with heat transfer coefficients for these high dispersion reactors. Their dispersion coefficient is highly sensitive to flow distribution/packing, in contrast to the generic limit at high Γ.

Finally, we consider the case of $\Gamma = 1$, which corresponds to the solutal dispersion in a pipe and into the porous wall or particles as in a catalytic monolith or packed-bed reactor, respectively. Here, the phases have equal capacities and, as reasoned above, D_{eff} should be sensitive to details in geometry and flow. For $\Gamma = 1$, (4.30) simplifies to

$$\Lambda = \frac{1}{48} g_1(\epsilon_f) + \frac{\mu}{8} g_2(\epsilon_f), \tag{4.35a}$$

$$\begin{aligned} g_1(\epsilon_f) &= \epsilon_f(6\epsilon_f^2 - 16\epsilon_f + 11), \\ g_2(\epsilon_f) &= \epsilon_f(4\epsilon_f - \epsilon_f^2 - 3 - 2\ln(\epsilon_f)). \end{aligned} \tag{4.35b}$$

We note that the function $g_1(\epsilon_f)$ has a maximum value of 2.26 at $\epsilon_f = 0.465$, while $g_2(\epsilon_f)$ has a maximum value of 0.206 at $\epsilon_f = 0.15$. As μ increases from 0 to ∞, the optimum ϵ_f decreases from 0.465 to 0.206. This is verified in Figure 6, where we have plotted Λ as a function of ϵ_f for different μ values. As expected, Λ is sensitive to both μ and ϵ_f.

5. Reactive and steady dispersion. It is clear from (4.18) that, at steady state, the mixing-cup temperature θ_m remains constant along the reactor—dispersion disappears at steady state. However, steady dispersion can exist under reactive conditions. As transverse diffusion and a longitudinal velocity gradient can produce transient longitudinal dispersion, steady reactive conversion differences across streamlines due to the transverse velocity gradient can also trigger transverse steady diffusive flux. The latter can, in turn, alter the overall conversion and produce an apparent steady dispersion. We shall examine this case here for the simplest scalar case—a single step irreversible reaction for solutal transport or a zeroth order reaction (with excess reactants) for thermal transport valid under thermal ignition conditions (Zeldovich et al. (1985)). Equation (4.1) can then be modified to

$$F(\theta, Pe) \equiv \frac{1}{\rho c_p} \nabla \cdot k \nabla \theta - Pe \left(\frac{\partial \theta}{\partial t} + w \frac{\partial \theta}{\partial z} \right) - PeKf(\theta) = 0, \tag{5.1}$$

where the Damköhler number $K(x, y, z, t)$ reflects the different activity in different phases (e.g., due to varying catalytic activity caused by nonuniform distribution of the catalytic agent and catalyst decay in time), and the nonlinear function $f(\theta)$ captures the temperature or concentration dependence of the reaction rate. (Note also that, unlike the previous cases, $F(\theta, Pe)$ is now nonlinear in θ.) Both $K(x, y, z, t)$ and $f(\theta)$ are of unit order with respect to Pe. The function $f(\theta)$ can be positive or negative, corresponding to solutal/thermal consumption or generation, respectively.

With this scaling, the transverse diffusion operator $\pounds$ remains the dominant linear operator. Moreover, the decomposition into a capacitance-weighted transverse average $\langle\theta\rangle$ and a θ' component, with $\langle\theta'\rangle = 0$, remains valid. However, the overall transverse balance now becomes

$$\frac{\partial \langle\theta\rangle}{\partial t} + \langle w \rangle \frac{\partial \theta_m}{\partial z} + \langle Kf(\theta) \rangle = 0. \tag{5.2}$$

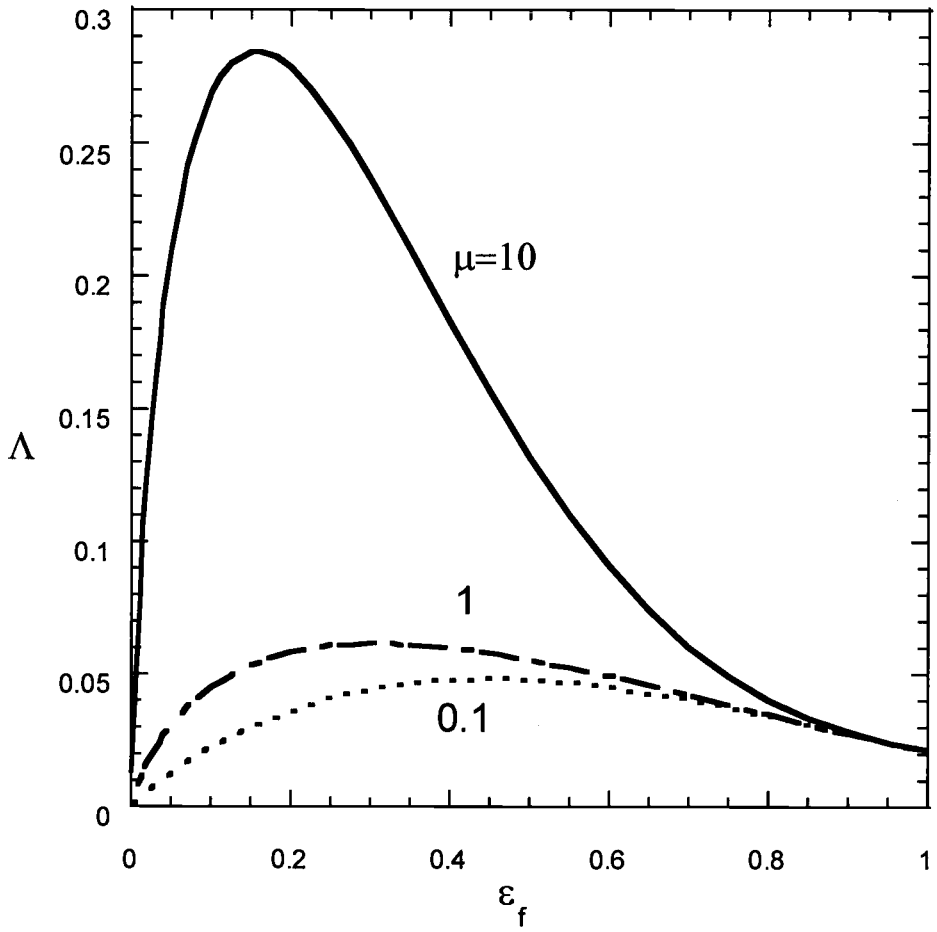


FIG. 6. *Unit order* Γ *dispersion as a function of* ϵ_f *for various* μ *values.*

To leading order, (5.2) reduces to

$$\frac{\partial \langle\theta\rangle}{\partial t} + \langle w\rangle \frac{\partial \langle\theta\rangle}{\partial z} + \langle K\rangle f(\langle\theta\rangle) + O(Pe) = 0. \tag{5.3}$$

The equation for θ' then becomes, to leading order in Pe,

$$\pounds\ \theta' = Pe(w - \langle w\rangle)\frac{\partial \langle\theta\rangle}{\partial z} + Pe(K - \langle K\rangle) f(\langle\theta\rangle). \tag{5.4}$$

We then require the decomposition of θ' into two components

$$\theta' = Pe\left[\eta \frac{\partial \langle\theta\rangle}{\partial z} + \chi f(\langle\theta\rangle)\right] + O(Pe^2); \tag{5.5}$$

the first term containing η is the transient dispersion contribution considered earlier. The reactive (source or sink) contribution is captured by χ as defined by

$$\pounds\ \chi = K - \langle K\rangle\,, \tag{5.6}$$

where $\langle\chi\rangle = 0$ as for η. The other boundary conditions are $\frac{\partial\chi}{\partial n} = 0$ on $\partial\Omega$ and continuity of χ and $k\frac{\partial\chi}{\partial n}$ across the phase boundaries $\partial\Omega_i$.

The homogenized model then becomes

$$\frac{\partial\langle\theta\rangle}{\partial t} + \langle w\rangle\frac{\partial\theta_m}{\partial z} + \langle K\rangle f(\langle\theta\rangle) + Pe\ f'(\langle\theta\rangle)\left(\delta\frac{\partial\langle\theta\rangle}{\partial z} + \kappa f(\langle\theta\rangle)\right) = 0, \tag{5.7a}$$

$$\theta_m = \langle\theta\rangle - Pe\left[\Lambda\frac{\partial\langle\theta\rangle}{\partial z} + \gamma f(\langle\theta\rangle)\right], \tag{5.7b}$$

where Λ is defined earlier by (4.17) and the new constants are

$$\gamma = -\frac{\langle w\chi\rangle}{\langle w\rangle}, \qquad \delta = \langle K\eta\rangle, \qquad \kappa = \langle K\chi\rangle.$$

We can combine (5.7a) and (5.7b) into a single equation for either $\langle\theta\rangle$ or θ_m. However, the coefficients that appear in the resulting equation are no longer constants but depend on the source function. As in the nonreactive case, representing the dispersion terms as second derivatives in z and interpretation of their coefficients as Taylor dispersion coefficients leads to further conceptual difficulties (in addition to the upstream diffusion, infinite propagation speed anomalies, and extra boundary or initial condition needed). Now, the capacitance-weighted average velocity and the dispersion coefficient are no longer constants but depend on the source function and its derivative. (They can be negative and hence lose their physical meaning!) Our approach based on the local length or times scales is still applicable here, the only difference being the additional length or time scales that appear due to the source or sink terms. Thus, in this case, it is preferable to leave the model in the two-mode form, the two modes being the mixing-cup and capacitance-weighted average temperature or concentration. The model is still hyperbolic as in the nonreactive case and reduces to (4.16) when the Damköhler number $K \equiv 0$. The initial and boundary conditions on (5.7) may also be derived in the same manner as in the nonreactive case.

We note that the reduced model now contains four effective local constants that are of the same order of magnitude. It can be seen that the two terms in (5.7b) (containing the constants Λ and γ) represent dispersion effects due to velocity gradients, while the two terms in (5.7a) (containing the constants δ and κ) represent dispersion effects due to a nonuniform reaction rate. The three new terms with coefficients γ, δ, and κ represent reactive dispersion effects (commonly known as mixing effects in the engineering literature), where as the nonreactive term with coefficient Λ may be interpreted as the traditional velocity gradient-induced dispersion. (This term may be interpreted as the so-called micromixing effect; see Chakraborty and Balakotaiah (2002) for more details.) We now examine various special cases.

The first case we consider is that in which $f(\theta)$ is a constant (say, $f(\theta) = -1$). This corresponds to the nonreactive situation, with a source term added to the classical Taylor problem or the thermal dispersion problem. For this case, three of the source-induced dispersion terms vanish, and (5.7) reduces to

$$\frac{\partial\theta_m}{\partial t} + \langle w\rangle\frac{\partial\theta_m}{\partial z} + \Lambda Pe\frac{\partial^2\theta_m}{\partial z\partial t} = \langle K\rangle. \tag{5.8}$$

Thus, the addition of a slowly varying source term simply adds its capacitance-weighted transverse average to the reduced model.

Next, we consider the case in which the Damköhler number K is constant (independent of transverse coordinates and time). Again, the three constants γ, δ, and κ vanish, and the model reduces to

$$\frac{\partial \theta_m}{\partial t} + \langle w \rangle \frac{\partial \theta_m}{\partial z} + \Lambda Pe \frac{\partial^2 \theta_m}{\partial z \partial t} = K\ f\left(\theta_m + \Lambda Pe \frac{\partial \theta_m}{\partial z}\right). \tag{5.9}$$

In this case, in addition to the transient dispersion term, we also have a source correction term, and the reduced model is different from the standard models used in the literature. These earlier literature models were obtained by just adding the source term to the nonreactive reduced model. Such models are clearly invalid as they exclude the correction term which appears in (5.9). (This was first noted by Balakotaiah and Dommeti (1999) using lumped resistance models. This correction term is also missing in the averaged models of Westerterp, Dilman, and Kronberg (1995) and Westerterp et al. (1995) using a heuristic approach.)

The third case we consider is that of a two-phase system in which the low capacitance fluid phase is moving and the solid phase is stationary. We also assume that $K = 0$ in the fluid phase and $K = 1$ in the solid phase. (This is a generalization of the packed-bed model with heat generation in the solid phase.) For this case, it may be seen that $\gamma > 0$, $\kappa > 0$, and $\delta < 0$. If $f'(\langle\theta\rangle) > 0$ (exothermic reaction), this corresponds to a decrease in the capacitance-weighted average velocity and an increase in the source strength in the reduced model.

The last case we consider is that of steady-state dispersion under reactive conditions. Now, since K is independent of t and since the time derivative in (5.2) is zero, we redo the Liapunov–Schmidt reduction. The reduced model is now given by the pair of equations

$$\langle w \rangle \frac{d\theta_m}{dz} + \langle K \rangle f(\langle\theta\rangle) + Pe\ \langle K\chi \rangle f'(\langle\theta\rangle) f(\langle\theta\rangle) = 0, \tag{5.10a}$$

$$\theta_m = \langle\theta\rangle + Pe\ \frac{\langle w\chi \rangle}{\langle w \rangle} f(\langle\theta\rangle), \tag{5.10b}$$

where χ is distinct from the transient version (5.6) and is now defined by

$$\pounds \chi = K - \frac{w}{\langle w \rangle} \langle K \rangle, \tag{5.11a}$$

$$\frac{\partial \chi}{\partial n} = 0 \qquad \text{on } \partial\Omega. \tag{5.11b}$$

The boundary condition to be used on (5.10) is

$$\theta = \theta_{m0} \qquad \text{at } z = 0. \tag{5.12}$$

The reduced model is a differential-algebraic system, and there is no second derivative term. (We note that the reduced model is not an initial value problem in z! For the case of an exothermic reaction, it is an index infinity differential-algebraic system and can have multiple (in fact, an infinite number of) solutions whenever the local equation (5.10b) has multiple solutions. This can happen when the kinetics is autocatalytic. For more details, see Chakraborty and Balakotaiah (2002).) The

steady-state reactive correction/dispersion term $-\langle w\chi\rangle$ is different from the transient reactive and nonreactive dispersion coefficient $-\langle w\eta\rangle$. For the special case of uniform activity (K is independent of transverse coordinates) it is easily seen that

$$\chi = -\frac{K}{\langle w\rangle}\eta, \tag{5.13}$$

where η is as defined in the nonreactive case. For this special case, the steady-state model (5.10) simplifies to

$$\langle w\rangle \frac{d\theta_m}{dz} + K\ f(\langle\theta\rangle) = 0, \tag{5.14a}$$

$$\theta_m = \langle\theta\rangle - \ Pe\frac{K\Lambda}{\langle w\rangle^2} f(\langle\theta\rangle), \tag{5.14b}$$

with initial/boundary condition (5.12). This reduced model is very different from the standard pseudohomogeneous model with Danckwerts boundary conditions. We note that while the numerical coefficient Λ that appears in the above steady-state model is the same as that in the Taylor's transient solutal dispersion problem, there is also a correction to the source term containing the same coefficient (see also (5.9)). In addition, it should be emphasized again that the reduced model is a differential-algebraic system rather than a two-point boundary value problem as in the following classical Danckwerts model:

$$\Lambda\frac{d^2\theta_m}{dz^2} - \langle w\rangle \frac{d\theta_m}{dz} - K\ f(\theta_m) = 0, \tag{5.15a}$$

$$\langle w\rangle\,\theta_{m0} = \langle w\rangle\,\theta_m\ - \Lambda\ \frac{d\theta_m}{dz} \qquad \text{at } z = 0, \tag{5.15b}$$

$$\frac{d\theta_m}{dz} = 0 \qquad \text{at } z = L. \tag{5.15c}$$

In this model, the exit boundary condition (5.15c) is imposed rather than derived from the original two-dimensional model.

6. Discussion. We have shown in this work that dispersion caused by transverse gradients can be described by reduced models that are hyperbolic in the longitudinal coordinate and time and that contain an effective local length or time scale. Our method also overcomes the main deficiencies of the previous approaches to averaging based on the moments method and the center manifold theorem. The former is applicable only to linear problems, while the latter describes the asymptotic behavior close to a fixed point (such as a trivial solution $\theta(x, y, z, t) = 0$). In contrast, our method is based on expansion around a state $\langle\theta\rangle(z, t)$ that is only independent of the transverse coordinates and can be applied to both steady-state and transient problems.

We have presented here the averaged models to only the lowest order in Pe. However, the extension to obtain higher order averaged models is straightforward. For example, for the solutal/thermal dispersion problem, it is easily seen that the reduced model to all orders (in Pe or t_H) is of the form

$$\frac{\partial\langle\theta\rangle}{\partial t'} + \langle u\rangle \frac{\partial\theta_m}{\partial z'} = 0, \tag{6.1a}$$

$$\langle\theta\rangle - \theta_m + \sum_{i=1}^{\infty} \beta_i (t_H)^i \frac{\partial^i \langle\theta\rangle}{\partial t^i} = 0, \tag{6.1b}$$

where t_H is the local time scale and β_i are numerical constants that depend on the velocity profile $w(x, y)$ and the geometry of Ω. Appropriate inlet and initial conditions may also be derived for (6.1).

The present approach may also be extended in many ways. Instead of the no-flux outer wall, we can allow an isothermal wall or a mixed boundary condition with a wall heat transfer coefficient. In such cases, the operator $\mathcal{L}$ is no longer singular and $\langle\theta\rangle$ does not strictly correspond to the null eigenfunction. However, if the transverse gradient remains small (and $\mathcal{L}$ has a discrete spectrum), the invariant manifold approach of Roberts (1989) and Balakotaiah and Chang (1995) can be used to extend the Lyapunov–Schmidt technique presented here.

Acknowledgments. We thank two anonymous referees for many helpful comments.

REFERENCES

R. Aris (1959), *On the dispersion of a solute by diffusion, convection and exchange between phases*, Proc. Roy. Soc. London A, 252, pp. 538–550.

V. Balakotaiah (1996), *Structural stability of nonlinear convection-reaction models*, Chem. Engrg. Edu., Fall, pp. 234–239.

V. Balakotaiah and H.-C. Chang (1995), *Dispersion of chemical solutes in chromatographs and reactors*, Phil. Trans. Roy. Soc. London A, 351, pp. 39–75.

V. Balakotaiah and S.M.S. Dommeti (1999), *Effective models for packed bed catalytic reactors*, Chem. Eng. Sci., 54, pp. 1621–1638.

V. Balakotaiah, D. Kodra, and D. Nguyen (1995), *Runaway limits for homogeneous and catalytic reactors*, Chem. Engrg. Sci., 50, pp. 1149–1171.

H. Brenner and D.A. Edwards (1993), *Macrotransport Processes*, Butterworth-Heinemann, Boston.

S. Chakraborty and V. Balakotaiah (2002), *Low dimensional models for describing micromixing effects in laminar flow tubular reactors*, Chem. Engng. Sci., 55, pp. 2545–2564.

H.-C. Chang (1982), *A non-Fickian model of packed bed reactors*, AIChE J., 28, pp. 208–214.

C. Y. Choi and D.D. Perlmutter (1976), *A unified treatment of the inlet boundary condition for dispersive flow models*, Chem. Eng. Sci., 31, pp. 250–252.

C.T. Culbertson, S.C. Jacobson, and J.M. Ramsey (1998), *Dispersion sources for compact geometries on microchips*, Anal. Chem., 70, pp. 3781–3789.

P.C. Chatwin (1970), *The approach to normality of the concentration distribution of a solute in a solvent flowing along a straight pipe*, J. Fluid Mech., 43, pp. 321–352.

P.V. Danckwerts (1953), *Continuous flow systems—distribution of residence times*, Chem. Eng. Sci., 2, pp. 1–13.

P. Garabedian (1964), *Partial Differential Equations*, Wiley, New York.

M.J.E. Golay (1958), *Theory of chromatography in open and coated tubular columns with round and rectangular cross-sections*, in Gas Chromatography, D.H. Desty, ed., Butterworth, London, pp. 36–49.

J.W. Hiby (1962), *Longitudinal and transverse mixing during single-phase flow through granular beds*, in Proceedings of the Symposium on Interaction between Fluids and Particles, Institution of Chemical Engineers, London, pp. 312–320.

M.J. Hinduja, S. Sundaresan, and R. Jackson (1980), *A crossflow model of dispersion in packed-bed reactors*, AIChE J., 26, pp. 274–281.

J.M. Keith, D.T. Leighton, and H.-C. Chang (1999), *A new design of reverse-flow reactors with enhanced thermal dispersion*, I & EC Res., 38, pp. 667–682.

D.L. Koch and J.F. Brady (1985), *Dispersion in fixed beds*, J. Fluid Mech., 154, pp. 399–427.

D.T. Leighton and H.-C. Chang (1995), *A theory for fast-igniting catalytic converters*, AIChE J., 41, pp. 1898–1914.

A.J. Majda and P.R. Kramer (1999), *Simplified models for turbulent diffusion: Theory, numerical modeling and physical phenomena*, Phys. Rep., 314, pp. 237–574.

G.N. Mercer and A.J. Roberts (1990), *A centre manifold description of contaminant dispersion in channels with varying flow properties*, SIAM J. Appl. Math., 50, pp. 1547–1565.

A.J. Roberts (1989), *The utility of an invariant manifold description of the evolution of a dynamical system*, SIAM J. Math. Anal., 20, pp. 1447–1458.

A.J. Roberts (1992), *Boundary conditions for approximate differential equations*, J. Aust. Math. Soc., B34, pp. 54–80.

S. Subramanian and V. Balakotaiah (1996), *Classification of the steady-state and dynamic behavior of distributed reactor models*, Chem. Engng. Sci., 51, pp. 401–421.

S. Sundaresan, N.R. Amundson, and R. Aris (1980), *Observations on fixed-bed dispersion models*, AIChE J., 26, pp. 529–536.

G.I. Taylor (1953), *Dispersion of soluble matter in solvent flowing slowly through a tube*, Proc. Roy. Soc. London A, 219, pp. 186–203.

D. Vortmeyer and R.J. Schaefer (1974), *Equivalence of one- and two-phase models for heat transfer processes in packed-beds: One dimensional theory*, Chem. Engng. Sci., 29, pp. 485–491.

J.F. Wehner and R.H. Wilhelm (197356), *Boundary conditions for flow reactors*, Chem. Eng. Sci., 28, pp. 89–93.

K.R. Westerterp, V.V. Dilman, and A.E. Kronberg (1995), *Wave model for longitudinal dispersion: Development of model*, AIChE J., 41, pp. 2013–2028.

K.R. Westerterp, V.V. Dilman, A.E. Kronberg, and A.H. Benneker (1995), *Wave model for longitudinal dispersion: Analysis and applications*, AIChE J., 41, pp. 2029–2039.

W.R. Young and S. Jones (1991), *Shear dispersion*, Phys. Fluids, A3, pp. 1087–1101.

Ya.B. Zeldovich, G.I. Barenblatt, V.B. Librovich, and G.M. Makhviladze (1985), *The Mathematical Theory of Combustion and Explosions*, Consultants Bureau, New York.

SIAM J. APPL. MATH.
Vol. 63, No. 4, pp. 1259–1275

EVANS FUNCTION STABILITY OF COMBUSTION WAVES*

V. GUBERNOV†, G. N. MERCER†, H. S. SIDHU†, AND R. O. WEBER†

Abstract. In this paper we investigate the linear stability and properties, such as speed, of the planar travelling combustion front. The speed of the front is estimated both analytically, using the matched asymptotic expansion, and numerically, by means of the shooting and relaxation methods. The Evans function approach extended by the compound matrix method is employed to numerically solve the linear stability problem for the travelling wave solution.

Key words. combustion waves, Evans function, compound matrix, Nyquist plot

AMS subject classifications. 35K57, 80A25

PII. S0036139901400240

1. Introduction.

Problems involving combustion waves are characterized by strong dependence of the reaction rate, which is usually modelled by the Arrhenius law, on the temperature. This sharp dependence of the reaction rate naturally divides the structure of the travelling front into three regions. Ahead of the combustion wave, in a preheat zone, the temperature is low and there is almost no reaction. When the temperature becomes sufficiently high, the reaction rate increases exponentially and the fuel is converted into heat very quickly. This takes place in a narrow region called the reaction zone. Finally, behind the front, in a product zone, all fuel is consumed, no reaction occurs, and the temperature is constant. The temperature and the amount of fuel change rapidly in the reaction zone. The above picture in some sense is close to the boundary layer problem. Therefore similar methods of analysis, like the matched asymptotic expansion (MAE), apply in both cases.

The analysis of steady propagating planar combustion fronts is usually based on MAE. According to this method, in the limit of high activation energy, we seek the travelling wave solution in the form of a series in all three regions; then on the boundaries of these zones the expansions are matched in each order of a small parameter. The asymptotic procedure in principle allows us to find the solution with any desired accuracy and for arbitrary Lewis number. As a rule, only the leading order is considered [1, 2, 3, 4, 5]; however, the higher order approximations can also be obtained [6]. The method of MAE is valid in the limit of large activation energy, and the properties of the steady combustion front, such as speed, can be found only numerically for general values of activation energy. Numerical analysis is mostly focused on solving the system of partial differential equations (PDE) that describe the problem [7, 8]. However, PDE can always be reduced to a system of ordinary differential equations (ODE) for steady propagating planar waves. In this paper we take advantage of the ODE formulation of the problem, which is usually more convenient for numerical analysis. We use shooting and relaxation methods to investigate the dependence of the speed of the front on the parameters of the problem. Besides the benefits of technical implementation, an ODE formulation does not depend on the

*Received by the editors December 28, 2001; accepted for publication (in revised form) October 2, 2002; published electronically April 9, 2003.
http://www.siam.org/journals/siap/63-4/40024.html
†School of Mathematics and Statistics, University of New South Wales at the Australian Defence Force Academy, Canberra, ACT 2600, Australia (vlad@ma.adfa.edu.au, g.mercer@adfa.edu.au, hss@ma.adfa.edu.au, r.weber@adfa.edu.au).

stability of the travelling wave and therefore allows us to continue the solution branch over a broader parameter range.

As we vary the parameters of the problem, a steady propagating planar front can lose stability, giving rise to either pulsating or cellular flames [5, 9]. Analytical investigation of the stability using the MAE leads to the so-called closure problem. In contrast to steady travelling waves, in this case the leading order equations depend on first order corrections, first order equations include second order terms of the asymptotic expansion, etc. In order to find the solution to the leading order problem, an infinite number of equations have to be investigated. One of the ways of overcoming this obstacle is just to truncate the expansion [1, 2, 3, 4, 5, 9]. This yields a closed problem with a replacement of the Arrhenius reaction rate by a delta-function source depending on the temperature at the reaction front. The truncated model has been used extensively for the stability analysis of combustion waves [1, 2, 3, 4]. However, the model with the delta-function source suffers from inconsistencies, as was noted in [5, 9]. The inverse of the small parameter of the expansion appears explicitly in the exponential terms describing the strength of the source. In other words, temperature variations behind the front are considered to be small in the exponential terms and of leading order elsewhere.

An asymptotically consistent approach was proposed in [5] for the system with the Lewis number of the order of unity. In this case the enthalpy does not change at the leading order. A closed problem was derived for the leading order temperature and the first order enthalpy. However, until recently [9], there has not been a consistent approach that treats the model with arbitrary Lewis number.

In [9] a generalization of the MAE method was introduced. The coefficients in the expansions are allowed to depend on the expansion parameter. This enables the correct scaling of the temperature variations ahead of and behind the reaction zone; namely, a restriction is imposed connecting the leading order temperature in the preheat zone and the first two terms of the asymptotic expansion in the product zone. The constraint reflects the fact that small temperature variations behind the front change the leading order terms in the preheat zone. The resulting model includes equations for the leading order variables ahead of the reaction zone and for the first order temperature variations in the product zone, together with matching and boundary conditions. There is no restriction on the range of the Lewis number values.

However, the models describing the propagation of the steady planar combustion waves were derived only in the leading order of the expansion parameter of the asymptotic procedure. In other words the papers mentioned above analyze the linear stability of models which are different from the original one with the Arrhenius kinetics. This is fully justified by the complexity of the problem and reveals the lack of alternative methods to MAE. Fortunately, recent advances in the application of the Evans function [10] to stability analysis of solitary waves and fronts [11, 12, 13, 14, 15] provide us with a powerful tool for the semianalytical investigation of travelling front stability.

The linear stability problem can always be formulated as an eigenvalue problem for some differential operator. The Evans function was first introduced in [10] to study the stability of nerve pulses as an analytical function whose zeros correspond to the isolated eigenvalues of this differential operator. In some specific cases the Evans function can be found explicitly, and the Evans function being zero gives the dispersion relation. In other cases, like solitary waves of a generalized Kordeweg–deVries (KdV) equation [16] and generalized nonlinear Schrödinger equation [17], the

asymptotic behavior of the Evans function can be found using the results obtained in [16], where the derivative of the Evans function was connected to the Melnikov integral [18]. The knowledge of asymptotics allowed the authors to localize zeros of the Evans function and to solve the stability problem analytically.

The asymptotic form of the Evans function for the combustion problem was derived in [15] in the limit of large activation energy and Lewis number of the order of unity. The results of [15] agree with the predictions of the MAE analysis of [5]. Zeros of the Evans function and therefore the stability of the combustion front can be found only numerically for general parameter values. Previously, this problem was solved by direct integration of the governing PDE [7, 8]. We cannot expect this method to be accurate near the critical parameter values, where the rate of instability is weak and a long integration time is needed to detect it. Furthermore, this method is relatively difficult for computational implementation in comparison with the numerical estimation of the Evans function proposed in [19], which is based on ODE integration. However, the latter method is applicable only for problems with a specific type of geometry, such as the linear stability problem for the KdV equation, and fails to work for stiff systems [12], such as in our case.

In the present paper we apply the Evans function method to the stability analysis of the planar combustion front. The linear stability problem associated with the travelling combustion wave is an example of a stiff problem. We extend the conventional algorithm for calculating the Evans function [19] by the compound matrix method [12, 13, 20, 21, 22], which was first employed for analysis of the hydrodynamic stability for the Orr–Sommerfeld equation. This method eliminates the stiffness and makes the linear stability problem numerically tractable.

In this paper we use the combination of shooting-relaxation and the Evans function method (extended by the compound matrix method) as a consistent approach for numerical investigation of both properties and stability of the travelling planar front, based on the ODE formulation of the combustion problem. We show that the method is valid for a wide range of the parameter values.

The paper is organized as follows. The model and governing equations are introduced in section 2. In section 3 we show how MAE can be used to derive travelling wave solutions in the limit of large activation energy, and we compare these solutions with the solutions obtained by shooting and relaxation methods. The linear stability problem is formulated in section 4, whereas the relation to the Evans function is discussed in section 5. In section 6 we quote the asymptotic results of [15] for the Evans function. Numerical stability analysis is carried out in section 7. Finally, concluding remarks can be found in section 8.

2. Model. We consider a premixed fuel in one dimension. The heat loss is neglected. We assume that the rate of exothermic combustion is well described by the Arrhenius law. In nondimensional coordinates, the equations governing this process can be found in [7] and are given as

$$u_t = u_{xx} + ve^{-1/u}, \qquad v_t = \tau v_{xx} - \beta ve^{-1/u}, \tag{2.1}$$

where u and v are the nondimensional temperature and mass fraction of the fuel, respectively; τ is the inverse Lewis number (the ratio of the diffusion rates of mass and heat); and β is the ratio of the activation energy to heat release.

We consider the ambient temperature to be equal to zero. This approximation simplifies the problem, decreasing the number of parameters. As is noted in [7], this is a way to circumvent the "cold-boundary problem" and does not change the behavior of

the system. It is not an appropriate simplification when considering ignition problems. Parameter τ varies from zero, for solid fuel, to unity, for gaseous fuels. The parameter β is of the order of unity or larger.

We consider system (2.1) subject to the following boundary conditions:

$$\begin{array}{llll} u(x,t) \to \beta^{-1}, & v(x,t) \to 0 & \text{as} & x \to -\infty, \\ u(x,t) \to 0, & v(x,t) \to 1 & \text{as} & x \to +\infty. \end{array} \tag{2.2}$$

On the right boundary we have a cold ($u = 0$) and unburned ($v = 1$) state, whereas the opposite limit corresponds to the hot ($u = \beta^{-1}$) and burned ($v = 0$) state.

3. Travelling wave solution. Let us consider the case $\tau \sim O(1)$. We will seek the solution of (2.1) in a form of the front travelling with a constant speed c

$$u(x,t) = u(\xi), \qquad v(x,t) = v(\xi), \tag{3.1}$$

where we have introduced a moving coordinate frame $\xi = x - ct$. After substituting (3.1) into (2.1), it is easy to obtain two second order differential equations

$$u_{\xi\xi} + cu_\xi + ve^{-1/u} = 0, \qquad \tau v_{\xi\xi} + cv_\xi - \beta ve^{-1/u} = 0 \tag{3.2}$$

and boundary conditions

$$\begin{array}{llll} u = \beta^{-1}, & v = 0 & \text{as} & \xi \to -\infty, \\ u = 0, & v = 1 & \text{as} & \xi \to +\infty. \end{array} \tag{3.3}$$

Now let us make the reaction terms in (3.1) symmetric by introducing new variables $\widetilde{u} = \beta u$, $\widetilde{v} = v$ and scale the coordinate $z = c\xi$. This gives us the equations

$$\widetilde{u}_{zz} + \widetilde{u}_z + \beta^2 Q\widetilde{v}e^{\beta(1-1/\widetilde{u})} = 0, \qquad \tau\widetilde{v}_{zz} + \widetilde{v}_z - \beta^2 Q\widetilde{v}e^{\beta(1-1/\widetilde{u})} = 0, \tag{3.4}$$

where $Q = (\beta c^2 e^\beta)^{-1}$ is a flame speed eigenvalue, which has to be found. Boundary conditions are modified as follows:

$$\begin{array}{llll} \widetilde{u} = 1, & \widetilde{v} = 0 & \text{as} & z \to -\infty, \\ \widetilde{u} = 0, & \widetilde{v} = 1 & \text{as} & z \to +\infty. \end{array} \tag{3.5}$$

In the new variables the reaction zone is of the order of β^{-1}. Outside this zone the reaction terms of (3.4) become negligible. Hence in the outer region (3.4) can be written as

$$\widetilde{u}_{zz} + \widetilde{u}_z = 0, \qquad \tau\widetilde{v}_{zz} + \widetilde{v}_z = 0, \tag{3.6}$$

subject to boundary conditions (3.5).

We seek the solution of problem (3.6) in the form of a series, with β^{-1} being a small parameter, and hence we postulate

$$\begin{aligned} \widetilde{u} &= U_0 + \beta^{-1}U_1 + \cdots, \\ \widetilde{v} &= V_0 + \beta^{-1}V_1 + \cdots, \\ Q &= Q_0 + \beta^{-1}Q_1 + \cdots. \end{aligned} \tag{3.7}$$

In the zeroth order, the solution of equations (3.6) is

$$U_0 = \begin{cases} 1, & z < 0, \\ e^{-z}, & z > 0, \end{cases} \quad \text{and} \quad V_0 = \begin{cases} 0, & z < 0, \\ 1 - e^{-z/\tau}, & z > 0. \end{cases} \tag{3.8}$$

In order to find Q_0 we should match the solution (3.8) with the solution of the inner problem.

Let us consider (3.4) in a thin boundary layer where the reaction occurs. We introduce the stretched coordinate $y = \beta z$ and seek solutions of the form

$$\begin{aligned} \widetilde{u} &= 1 + \beta^{-1}u_1 + \beta^{-2}u_2 + \cdots, \\ \widetilde{v} &= \beta^{-1}v_1 + \beta^{-2}v_2 + \cdots, \\ Q &= Q_0 + \beta^{-1}Q_1 + \cdots. \end{aligned} \tag{3.9}$$

After substituting (3.9) into (3.4) and leaving the leading terms of the order $O(\beta)$, we obtain the equations

$$\ddot{u}_1 + Q_0 v_1 e^{u_1} = 0, \qquad \tau\ddot{v}_1 - Q_0 v_1 e^{u_1} = 0, \tag{3.10}$$

with the dot denoting derivative $\partial/\partial y$. In deriving (3.10) we have assumed that $\tau \sim O(1)$ and $Q_0 \sim O(1)$. On the boundaries we can require that the following matching conditions be satisfied:

$$\begin{aligned} &\lim_{y\to\pm\infty}[u_1(y) - U_1(0\pm) - U_{0z}(0\pm)y] = 0, \\ &\lim_{y\to\pm\infty}[v_1(y) - V_1(0\pm) - V_{0z}(0\pm)y] = 0. \end{aligned} \tag{3.11}$$

On the left boundary from (3.8) it follows that $U_{0z}(0-) = V_{0z}(0-) = 0$, and we can rewrite (3.11) in the form

$$\begin{aligned} &u_1 = U_1(0-), \quad v_1 = V_1(0-) \\ &\dot{u}_1 = 0, \qquad\quad \dot{v}_1 = 0 \end{aligned} \qquad \text{as} \quad y \to -\infty. \tag{3.12}$$

System (3.10) has the integral

$$u_1 + \tau v_1 = C_1 + D_1 y. \tag{3.13}$$

Applying the matching condition (3.12), it can be shown that $D_1 = 0$ and $C_1 = U_1(0-) + \tau V_1(0-)$. Using this integral system, (3.10) can be reduced to a single equation

$$\ddot{\theta} - k\theta e^{-\theta} = 0, \tag{3.14}$$

where $\theta = C_1 - u_1$, $k = Q_0/\tau e^{C_1}$. Equation (3.14) can be treated as the equation of motion of a point particle with unit mass in the potential $W(\theta) = k(1+\theta)e^{-\theta}$. The potential $W(\theta)$ has a maximum only for $\theta = 0$. Conditions (3.12) imply that $\dot{\theta}(-\infty) = 0$. This is possible only if the energy $E = \dot{\theta}^2/2 + W(\theta)$ is equal to the maximal value of the potential $W(0) = k$. Therefore, $\theta = 0$ as $y \to -\infty$, and consequently, $C_1 = U_1(0-)$ and $V_1(0-) = 0$. On the other hand, as $y \to +\infty$ the potential $W(y)$ decays exponentially and $\dot{\theta} \to \sqrt{2k}$.

Returning to the original problem, the left boundary conditions can be written as

$$\begin{aligned} &u_1 = U_1(0-), \quad v_1 = 0 \\ &\dot{u}_1 = 0, \qquad\quad \dot{v}_1 = 0 \end{aligned} \qquad \text{as} \quad y \to -\infty. \tag{3.15}$$

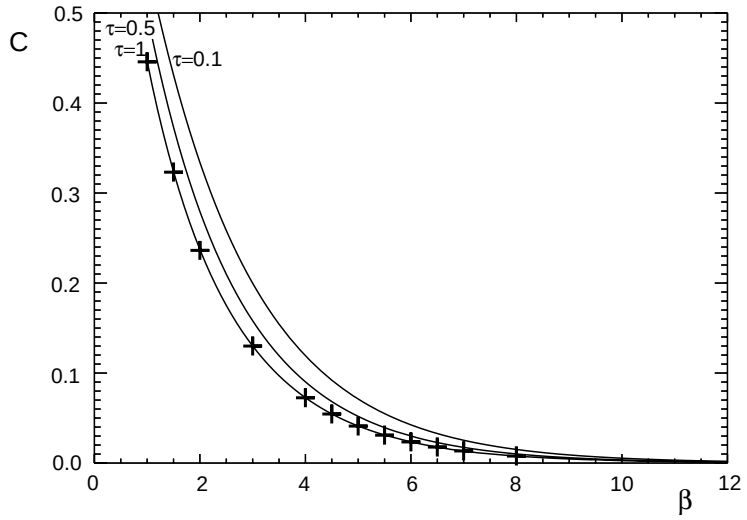


FIG. 1. *Numerically determined speed of the travelling front as a function of β. Solid lines correspond to the results obtained by shooting and relaxation methods for values of parameter $\tau =$ 0.1, 0.5, and 1.0 (right to left). Crosses represent the speed of the front calculated by direct PDE integration of* (2.1) *for* $\tau = 1.0$.

On the right boundary, it follows from (3.11) and (3.13) that

$$\left(\frac{\partial U_0}{\partial z}\right)_{z=0+} = -\tau \left(\frac{\partial V_0}{\partial z}\right)_{z=0+} = -\sqrt{\frac{2Q_0}{\tau e^{U_1(0-)}}}. \tag{3.16}$$

For planar front solution $U_1(0-) = 0$ and from (3.8) we have $(\partial U_0/\partial z)_{z=0+} = -1$. Taking into account the definition of Q, we can obtain the following estimation of the speed:

$$c = \sqrt{2\tau^{-1}\beta^{-1}} e^{-\beta/2}, \tag{3.17}$$

which agrees with the results of [7, 8] in the limit of large β. Note that a similar expression for the front speed was first found in [6].

We also solved (3.2) numerically. As in [23], we used the shooting method to obtain the guess solution, and then the results were corrected with a more accurate method, namely, relaxation. Combination of these methods allowed us to numerically obtain the dependence of the travelling front velocity on the parameters β and τ. In Figure 1 we plot the speed of the front as a function of β for three different values of τ. The results are compared to the predictions obtained in [7] by direct PDE integration of (2.1). The accordance between these two approaches is excellent.

In Figure 2 we compare the prediction of the asymptotic formula (3.17) for the speed of the front with the results obtained numerically. As can be seen, the correspondence is quite good for $\tau = 1$ and large values of β. However, when we decrease the value of τ, the approximation of the speed (3.17), which is valid for $\tau \sim O(1)$, becomes unsatisfactory. For example, when $\tau = 0.1$, the difference between the analytical and numerical results becomes significant.

The stability analysis of the steady propagating combustion front carried out in the following sections is based on how accurately we can approximate the solution of

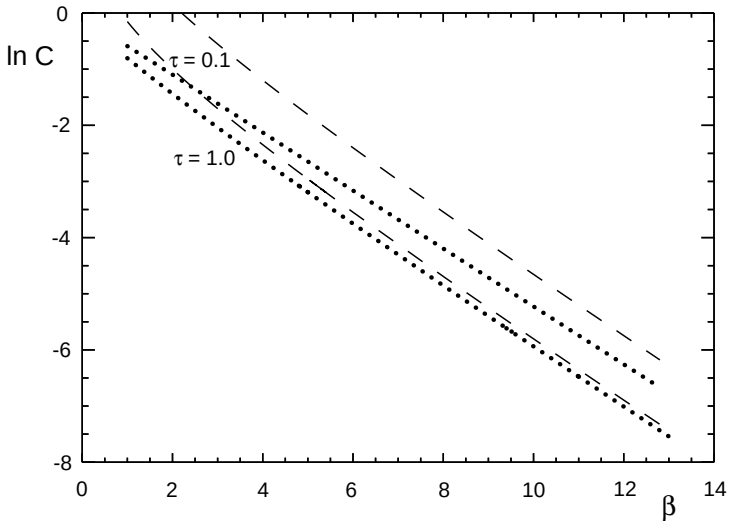


FIG. 2. *Logarithm of the speed of the travelling front as a function of β for values of parameter $\tau = 0.1$ and 1.0. Dots correspond to the numerical results, whereas dashed lines represent the approximation of the speed according to the formula* (3.17). *In each case the upper line is for $\tau = 0.1$ and the lower line for $\tau = 1.0$.*

(3.2). Standard relaxation routine (see [23] and references therein) allows us to control the average local correction made on each iteration step. The solution is considered to be resolved if the correction is less than 10^{-15}. We also tested the accuracy of the method independently by changing the step of the grid and comparing the resulting variations in the values of the front speed. For example, a fourfold mesh refinement changes the value of the front speed in the ninth significant digit. Therefore we assume that the numerical procedure outlined here works reasonably well.

4. Stability of a travelling front. As a first step in the analysis of travelling wave stability, we linearize (2.1) around the front solution (3.1):

$$u(x,t) = u(\xi) + \varphi(\xi,t), \qquad v(x,t) = v(\xi) + \chi(\xi,t), \tag{4.1}$$

where φ and χ are linear perturbation terms. After substitution of (4.1) into (2.1) it is straightforward to derive

$$\begin{pmatrix} \partial\varphi/\partial t \\ \partial\chi/\partial t \end{pmatrix} = \hat{L} \begin{pmatrix} \varphi \\ \chi \end{pmatrix}, \tag{4.2}$$

where

$$\hat{L} = \begin{pmatrix} \partial_\xi^2 + vu^{-2}e^{-1/u} + c\partial_\xi & e^{-1/u} \\ -\beta vu^{-2}e^{-1/u} & \tau\partial_\xi^2 - \beta e^{-1/u} + c\partial_\xi \end{pmatrix}. \tag{4.3}$$

The stability of the travelling front is then defined from the spectra of $\hat{L}$. It is easy to show that the essential spectrum of this operator always lies in the left half-plane and therefore the discrete spectrum is solely responsible for the transition to instability (see [24]). We will seek the solution of (4.2) of the form

$$\varphi(\xi,t) = \varphi(\xi)e^{\lambda t}, \qquad \chi(\xi,t) = \chi(\xi)e^{\lambda t}, \tag{4.4}$$

where λ is a spectral parameter (in combustion literature it is sometimes referred to as the growth rate eigenvalue). Substituting (4.4) into (4.2) and introducing a vector with the components $z_1 = \varphi,\ z_2 = \varphi_\xi,\ z_3 = \chi,\ z_4 = \chi_\xi$, we obtain the system of ODE in the form

$$\dot{\mathbf{z}} = \mathbf{A}\mathbf{z}, \tag{4.5}$$

where

$$\mathbf{A}(\xi,\lambda) = \begin{pmatrix} 0 & 1 & 0 & 0 \\ \lambda - vu^{-2}e^{-1/u} & -c & -e^{-1/u} & 0 \\ 0 & 0 & 0 & 1 \\ \beta\tau^{-1}vu^{-2}e^{-1/u} & 0 & \tau^{-1}(\lambda + \beta e^{-1/u}) & -\tau^{-1}c \end{pmatrix}. \tag{4.6}$$

We use (4.5) to investigate the stability of the travelling front. Following [12], we will say that the travelling front is linearly unstable if, for some fixed complex λ with $Re(\lambda) > 0$, there exists a solution of (4.5) which decays exponentially as $\xi \to \pm\infty$. We will refer to this λ as an eigenvalue and to the corresponding solution as an eigenmode.

5. Evans function. Let us introduce the limit matrices

$$\mathbf{A}_\pm(\lambda) \equiv \lim_{\xi\to\pm\infty} \mathbf{A}(\xi,\lambda). \tag{5.1}$$

The explicit form of $\mathbf{A}_\pm$ can be found from the boundary conditions (3.3). The limit matrices have eigenvalues

$$\begin{aligned} &\mu^-_{1,2}(\lambda) = \tfrac{-c\mp\sqrt{c^2+4\lambda}}{2}, && \mu^+_{1,2} = \mu^-_{1,2}, \\ &\mu^-_{3,4}(\lambda) = \tfrac{-c\mp\sqrt{c^2+4\tau(\lambda+\beta e^{-\beta})}}{2\tau}, && \mu^+_{3,4}(\lambda) = \mu^-_{3,4}(\lambda - \beta e^{-\beta}), \end{aligned} \tag{5.2}$$

with corresponding eigenvectors $\mathbf{k}^\pm_i$ (for $i = 1,\dots,4$). Equations (5.2) imply that $\mathbf{A}_-$ has two eigenvalues $\mu^-_{2,4}$ with positive real parts and two eigenvalues $\mu^-_{1,3}$ with negative real parts. Similarly, for $\mathbf{A}_+$ we have $Re(\mu^+_{2,4}) > 0$ and $Re(\mu^+_{1,3}) < 0$. Therefore, for any value of λ there exist two linearly independent solutions $\mathbf{z}^-_{2,4}(\xi,\lambda)$ of (4.5) corresponding to unstable subspaces of $\mathbf{A}_-$ satisfying the conditions

$$\lim_{\xi\to-\infty} \exp(-\mu^-_i\xi)\mathbf{z}^-_i(\xi,\lambda) = \mathbf{k}^-_i, \qquad i = 2,4, \tag{5.3}$$

and two linearly independent solutions $\mathbf{z}^+_{1,3}(\xi,\lambda)$ of (4.5) corresponding to stable subspaces of $\mathbf{A}_+$ satisfying the conditions

$$\lim_{\xi\to+\infty} \exp(-\mu^+_i\xi)\mathbf{z}^+_i(\xi,\lambda) = \mathbf{k}^+_i, \qquad i = 1,3. \tag{5.4}$$

Now we can consider a space of solutions of (4.5) bounded as $\xi \to -\infty$ and a space of solutions bounded as $\xi \to +\infty$. If these spaces intersect nontrivially for some value λ, then λ is an eigenvalue. We will call the function which measures whether these spaces intersect the Evans function. Geometrically this means that for some value of λ and any value of coordinate ξ the plane defined by the vectors $\mathbf{z}^-_{2,4}$ intersects nontrivially with the plane defined by the vectors $\mathbf{z}^+_{1,3}$. We can also say that λ is an eigenvalue if and only if the solutions $\mathbf{z}^-_{2,4}$ and $\mathbf{z}^+_{1,3}$ are linearly dependent or, equivalently, the Wronskian evaluated on these solutions (a matrix whose columns

are $\mathbf{z}^-_{2,4}(\xi)$ and $\mathbf{z}^+_{1,3}(\xi)$) is equal to zero. One of the Evans function definitions is given in [15] via this Wronskian, which is evaluated for definiteness at $\xi = 0$. Let $\mathbf{e}_i$ be the orthonormal basis in four dimensional space $\mathbf{C^4}$ of system (4.5) solutions. In this basis the vectors $\mathbf{z}^\pm_i$ have coordinates $(z^\pm_{i1}, z^\pm_{i2}, z^\pm_{i3}, z^\pm_{i4})^T$, and the Evans function is defined as

$$D(\lambda) = \begin{vmatrix} z^-_{21}(0,\lambda) & z^-_{41}(0,\lambda) & z^+_{11}(0,\lambda) & z^+_{31}(0,\lambda) \\ z^-_{22}(0,\lambda) & z^-_{42}(0,\lambda) & z^+_{12}(0,\lambda) & z^+_{32}(0,\lambda) \\ z^-_{23}(0,\lambda) & z^-_{43}(0,\lambda) & z^+_{13}(0,\lambda) & z^+_{33}(0,\lambda) \\ z^-_{24}(0,\lambda) & z^-_{44}(0,\lambda) & z^+_{14}(0,\lambda) & z^+_{34}(0,\lambda) \end{vmatrix}. \tag{5.5}$$

In what follows we will also require an alternative definition of the Evans function. Returning to the geometrical picture, we can say that to find the eigenvalues we do not have to seek the solutions $\mathbf{z}^-_{2,4}$ and $\mathbf{z}^+_{1,3}$, but it is sufficient to determine the orientation of the planes, constructed on corresponding pairs of vectors, in space $\mathbf{C^4}$ of system (4.5) solutions.

If we take two linear independent vectors in $\mathbf{C^n}$, then the orientation of a plane containing both vectors can be determined by a wedge product of them. (Two vectors are linear dependent if and only if their wedge product is equal to zero.) If $n = 3$, a wedge product coincides with a vector product, and the result of the operation is a vector belonging to $\mathbf{C^3}$. However, in the general case the result of a wedge product of two vectors is a vector lying in $\bigwedge^2(\mathbf{C^n})$, where $\bigwedge^2(\mathbf{C^n})$ is the second exterior power of $\mathbf{C^n}$. It has dimension $\dim[\bigwedge^2(\mathbf{C^n})] = n!/2!(n-2)!$.

In our case a plane can be defined by a six component vector; for instance, we define

$$\mathbf{V}^- = \mathbf{z}^-_2 \wedge \mathbf{z}^-_4, \qquad \mathbf{V}^+ = \mathbf{z}^+_1 \wedge \mathbf{z}^+_3, \tag{5.6}$$

where $\mathbf{V}^\pm \in \bigwedge^2(\mathbf{C^4})$ and $\wedge$ stands for wedge product. If λ is an eigenvalue, the planes associated with the vectors $\mathbf{V}^+$ and $\mathbf{V}^-$ intersect nontrivially. This means that $\mathbf{V}^+$ and $\mathbf{V}^-$ are linear dependent and a wedge product $\mathbf{V}^+ \wedge \mathbf{V}^-$ equals zero. Now we can make use of the Evans function definition given in [25] as

$$\widetilde{D}(\lambda) = \exp\left[-\int_0^\xi Tr(\mathbf{A}(s,\lambda))\right] \mathbf{V}^+(\xi,\lambda) \wedge \mathbf{V}^-(\xi,\lambda). \tag{5.7}$$

For the sake of further consideration it is convenient to take $\xi = 0$ in the definition (5.7). As was shown in [12], in this case we can rewrite (5.7) as

$$\widetilde{D}(\lambda) = D(\lambda)\Gamma, \tag{5.8}$$

where $\Gamma = \mathbf{e}_1 \wedge \mathbf{e}_2 \wedge \mathbf{e}_3 \wedge \mathbf{e}_4$ is a standard volume in $\mathbf{C^4}$ and

$$D(\lambda) = [\mathbf{V}^+, \Sigma\overline{\mathbf{V}^-}]. \tag{5.9}$$

Here, the overline denotes the complex conjugate, $[\cdot,\cdot]$ is the complex inner product in $\mathbf{C^6}$, and Σ is the operator that in a basis

$$\begin{aligned} &\mathbf{v}_1 = \mathbf{e}_1 \wedge \mathbf{e}_2, \quad \mathbf{v}_2 = \mathbf{e}_1 \wedge \mathbf{e}_3, \quad \mathbf{v}_3 = \mathbf{e}_1 \wedge \mathbf{e}_4, \\ &\mathbf{v}_4 = \mathbf{e}_2 \wedge \mathbf{e}_3, \quad \mathbf{v}_5 = \mathbf{e}_2 \wedge \mathbf{e}_4, \quad \mathbf{v}_6 = \mathbf{e}_3 \wedge \mathbf{e}_4, \end{aligned} \tag{5.10}$$

is represented by the matrix

$$\Sigma = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{5.11}$$

In the basis (5.10) the components of vector $\mathbf{V}^-$ can be shown to be

$$\begin{array}{ll} \mathbf{V}_1^- = z_{21}z_{42} - z_{41}z_{22}, & \mathbf{V}_4^- = z_{22}z_{43} - z_{42}z_{23}, \\ \mathbf{V}_2^- = z_{21}z_{43} - z_{41}z_{23}, & \mathbf{V}_5^- = z_{22}z_{44} - z_{42}z_{24}, \\ \mathbf{V}_3^- = z_{21}z_{44} - z_{41}z_{24}, & \mathbf{V}_6^- = z_{23}z_{44} - z_{43}z_{24}. \end{array} \tag{5.12}$$

A similar relation holds for $\mathbf{V}^+$. Using these expressions for $\mathbf{V}^\pm$, we can show (by means of direct substitution) that definitions (5.5) and (5.9) are identical.

The vectors $\mathbf{V}^\pm$ can be also called the second compounds of the matrices formed from a pair $\mathbf{z}_{2,4}^-$ or $\mathbf{z}_{1,3}^+$, respectively, and were considered in [20, 21, 22] in relation with the Orr–Sommerfeld equation. The definition (5.5) is the simpler of the two, but the importance of the second definition (5.9) will be revealed in section 7.

6. Evans function for $\tau \sim O(1)$ and $\beta \gg 1$. Let us return to the definition (5.5) of the Evans function. In [15] it was shown that in the case of $\tau \sim 1$ and $\beta \gg 1$ the Evans function can be approximated analytically with good accuracy. In this section we assume that $\tau = 1 - \beta^{-1}\ell$, where $\ell \sim 1$.

First, let us rewrite the spectral problem (4.2)–(4.3) for the operator $\hat{L}$ in the symmetric form. By introducing the variables $\widetilde{u} = \beta u$, $\widetilde{v} = v$, $\widetilde{\varphi} = \beta\varphi$, $\widetilde{\chi} = \chi$, and the scaled coordinate $z = c\xi$ defined in section 3, we can obtain

$$\hat{L}\begin{pmatrix} \widetilde{\varphi} \\ \widetilde{\chi} \end{pmatrix} = \widetilde{\lambda}\begin{pmatrix} \widetilde{\varphi} \\ \widetilde{\chi} \end{pmatrix}, \tag{6.1}$$

where $\widetilde{\lambda} = \lambda/c^2$ and

$$\hat{L} = \begin{pmatrix} \partial_z^2 + \partial_z + \beta^3 Q\widetilde{v}\widetilde{u}^{-2}e^{\beta(1-1/\widetilde{u})} & \beta^2 Q e^{\beta(1-1/\widetilde{u})} \\ -\beta^3 Q\widetilde{v}\widetilde{u}^{-2}e^{\beta(1-1/\widetilde{u})} & \tau\partial_z^2 + \partial_z - \beta^2 Q e^{\beta(1-1/\widetilde{u})} \end{pmatrix}. \tag{6.2}$$

A similar type of equation was considered in [15]. The solutions $\widetilde{\mathbf{z}}_{2,4}^-$ and $\widetilde{\mathbf{z}}_{1,3}^+$ of (6.1), having the same meaning as in the previous section, were found. Using the definition (5.5), the Evans function can be shown to be approximated in the limit $\beta \to \infty$ as

$$D(\widetilde{\lambda}, \ell) = 2\Gamma^2(\Gamma - 1) - \ell(2\widetilde{\lambda} - \Gamma + 1), \tag{6.3}$$

where $\Gamma = \sqrt{1 + 4\widetilde{\lambda}}$. As ℓ crosses $\ell_c = 4(1+\sqrt{3}) \approx 10.92$, a pair of complex conjugate eigenvalues $\pm i\widetilde{\lambda}_c$, where $\widetilde{\lambda}_c \approx 0.6356$, passes into the right half of the complex plane, giving rise to a Hopf bifurcation. The same result was obtained in [4, 5] using the matched asymptotic expansion. Now we can estimate the boundary of stability of the planar front as $\beta = \ell_c/(1-\tau)$ and the Hopf frequency $\lambda = c^2\widetilde{\lambda}_c$. Unfortunately, we cannot expect this prediction to be quantitatively accurate, because in the beginning of the consideration it was assumed that $\ell \sim O(1)$, whereas ℓ_c has been found to be much greater.

7. Numerics and the compound matrix method. The method of calculating the Evans function was proposed in [19]. The idea is based on the definition (5.5). According to (5.5) it is necessary to determine the solutions $\mathbf{z}^-_{2,4}$ or $\mathbf{z}^+_{1,3}$ at $\xi = 0$. In order to numerically trace the solution, for example, growing as we integrate forward, the coordinate is exponentially scaled so as to eliminate the maximal rate of exponential growth. However, the numerical algorithm introduced in [19] allows us to find only the solutions which correspond to maximal (minimal) rates of exponential growth (decay) as $\xi \to \pm\infty$.

In our case we need to obtain two pairs of solutions: one pair bounded as ξ tends to $+\infty$, another as $\xi \to -\infty$. Let us consider for definiteness $\xi < 0$. From the analysis developed in section 5 it follows that system (4.5) has two solutions $\mathbf{z}^-_{2,4}$ bounded as $\xi \to -\infty$ and two solutions $\mathbf{z}^-_{1,3}$ unbounded as $\xi \to -\infty$ which are of no interest. In order to neglect $\mathbf{z}^-_{1,3}$ we have to numerically integrate the system (4.5) from $\xi = -l_1$ to $\xi = 0$ (where l_1 is chosen sufficiently large). Integrating forward we can find only $\mathbf{z}^-_4$, because $\mu^-_2 < \mu^-_4$ and the solution $\mathbf{z}^-_2$ is always ruled out due to errors of the numerical scheme. The same obstacles remain in the limit $\xi \to +\infty$. Systems with this kind of behavior are called stiff. The stiffness makes the direct calculation using (5.5) impossible, and some procedure of orthogonalization is required. The compound matrix method, which we employed in order to avoid this type of difficulty, is described below and will be seen to be closely related to the definition (5.9) introduced in section 5.

Let $\mathbf{z}_1$ and $\mathbf{z}_2$ be two solutions of (4.5); then the vector $\mathbf{V} = \mathbf{z_1} \wedge \mathbf{z_2}$ is the solution of the equation

$$\dot{\mathbf{V}} = \mathbf{B}\mathbf{V}, \tag{7.1}$$

where $\mathbf{B}$ is a 6×6 matrix whose elements can be found from the matrix $\mathbf{A}$ (see [13, 20, 21, 22] for details). It can be shown that the eigenvalues $s^\pm_i$ of $\mathbf{B}$ in the limits $\xi = \pm\infty$ are given via eigenvalues $\mu^\pm_i$ of $\mathbf{A}_\pm$ as

$$\begin{aligned} s^\pm_1 &= \mu^\pm_1 + \mu^\pm_2, \quad s^\pm_4 = \mu^\pm_2 + \mu^\pm_3, \\ s^\pm_2 &= \mu^\pm_1 + \mu^\pm_3, \quad s^\pm_5 = \mu^\pm_2 + \mu^\pm_4, \\ s^\pm_3 &= \mu^\pm_1 + \mu^\pm_4, \quad s^\pm_6 = \mu^\pm_3 + \mu^\pm_4. \end{aligned} \tag{7.2}$$

Therefore $\mathbf{V}^-(\xi)$, defined by (5.6), is the solution of (7.1) corresponding to the largest rate of exponential growth s^-_5 as we integrate forward from $\xi = -\infty$ to $\xi = 0$. Similarly $\mathbf{V}^+(\xi)$, defined by (5.6), is the solution of (7.1) corresponding to the largest rate of exponential growth s^+_2 as we integrate backward from $\xi = +\infty$ to $\xi = 0$. These solutions can always be found numerically by means of the method introduced in [19], and then we use definition (5.9) to calculate the Evans function numerically.

The problem of stability of the travelling front of (2.1) then reduces to the search for zeros of the Evans function (5.9) located in the right half-plane. Zeros of $D(\lambda)$ can be calculated using an argument principle. The number of zeros in the right half-plane equals the number of times the image of the imaginary axis under $D(it)$, $t \in R$, winds (wraps) around the origin. Graphs of $D(it), t \in R$, are called Nyquist plots [19].

Figure 3 shows Nyquist plots for $\tau = 0.1$ and $\beta = 6.8,\ 7.026,\ 7.2$. For $\beta = 6.8$ the curve does not encircle the origin, and hence the travelling front is stable. Transition to instability occurs for $\beta = 7.026$, when two complex conjugate eigenvalues cross the imaginary axes and the curve passes through the origin three times. The front

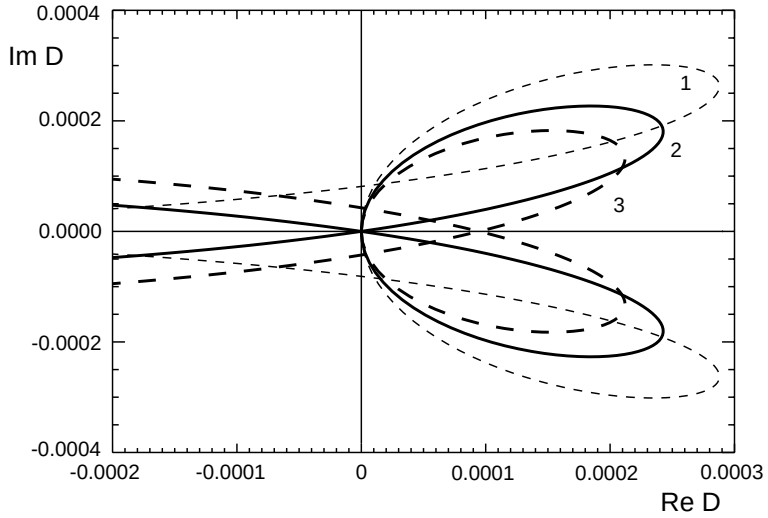


FIG. 3. *Transition to instability. The image of the imaginary axis $D(it)$ near the origin for $\tau = 0.1$ and for $\beta = 6.8$ (curve 1), 7.026 (curve 2), 7.2 (curve 3). Note that curve 1 does not encircle the origin, curve 2 is the transition, and curve 3 encircles the origin two times.*

is clearly unstable for $\beta = 7.2$, when the curve encircles the origin two times, and therefore there are two points of discrete spectra in the right half-plane.

The Nyquist plot technique allows us to obtain the criterion of transition to instability. Using this criterion, we can conclude whether the travelling front is stable or not for some fixed parameter values. However, it is quite difficult to calculate the critical parameter values. More detailed information can be collected by means of the Newton–Raphson method, which we apply to the equation $D(\lambda) = 0$. This allows us to locate the zeros of the Evans function on the complex plane for any given values of β and τ. In Figure 4 it is shown how two complex conjugate eigenvalues move from the left half-plane to the right half-plane, resulting in Hopf bifurcation at the parameter values when λ is purely imaginary. To determine the critical value β_c, when a pair of eigenvalues is located exactly on the imaginary axis, we consider the equation $Re\lambda = 0$ together with $D(\lambda) = 0$ and solve this system using the Newton–Raphson method. We start the iteration process with appropriate guess values for λ and β. We repeat the process until the zero is found with an accuracy of 10^{-12}. This gives us $\beta_c = 7.02609...$ for $\tau = 0.1$. While searching for the zeros of the Evans function, we checked the credibility of the method by decreasing the integration step by a factor of four. This results in the variation of the critical value of β in the ninth significant figure.

It is clear that we can use the procedure described above to find the dependence of the critical value β_c on τ. In Figure 5 the stability boundary is plotted in the parameter plane (β, τ). In section 6 we showed that as we increase τ towards 1, the corresponding value β_c tends to infinity. At the same time, according to (3.17) and the results of section 6, the speed of the front and the Hopf frequency tends to zero. As a result, the travelling front becomes flatter, and we should infinitely increase the interval of integration. For example, when $\tau = 0.6$, the critical values become $\beta_c = 21.087$, $c = 9.509 \times 10^{-6}$, $\lambda_c = 6.969 \times 10^{-11}$, and the interval of integration is about 10^6. In addition, it is numerically difficult to trace such small

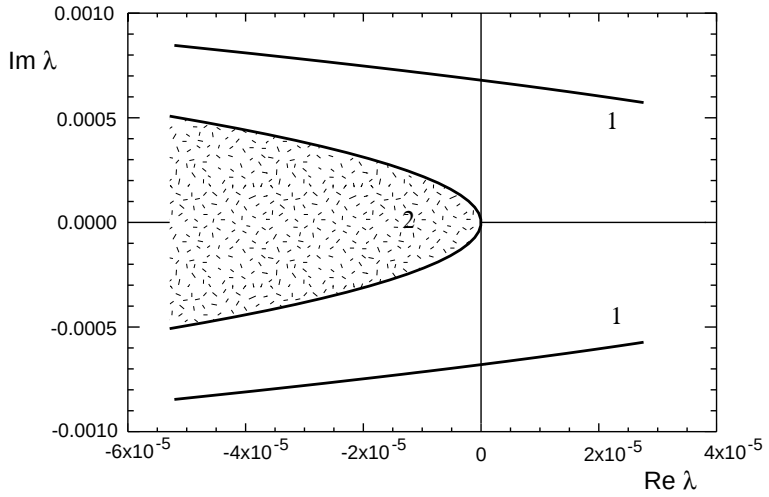


FIG. 4. *The spectra of operator* $\hat{L}$ *for* $\tau = 0.1$ *and* β *in the range* [6.8, 7.2]. *Curves* 1 *denote the location of eigenvalues for different values of* β. *Curve* 2 *is the boundary of the region where the essential spectrum lies. The value of* β *at which curves* 1 *pass through the imaginary axis gives the critical value* β_c.

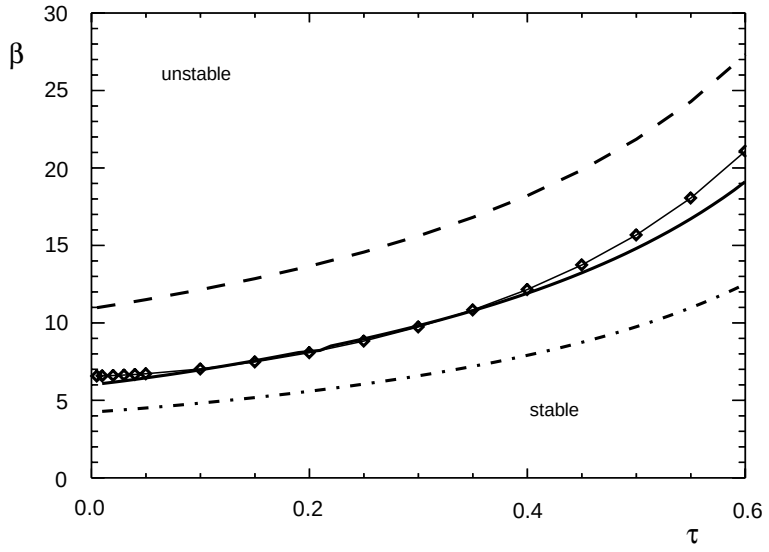


FIG. 5. *Stability boundary* $\beta_c(\tau)$. *Dots connected with a thin solid curve are the results of numerical calculations based on the Evans function. Other lines represent the analytical predictions obtained with* (1) *dash line—Evans function asymptotic from section* 6, (2) *dash-dot line—truncation model,* (3) *thick solid line—generalized MAE.*

values of λ_c. This implies that we can find the boundary of stability only on the interval $\tau \in [0, 0.6]$. As a result, it is extremely difficult to verify the asymptotic formula for $\beta_c(\tau)$ derived in section 6, as this is valid for $\tau \sim 1$. In Figure 5 we also plot the prediction obtained with the truncation model [9]. As can be seen, the discrepancy between the theoretical and numerical results is large for both asymptotic models. The generalized MAE developed in [9] gives the best correspondence with

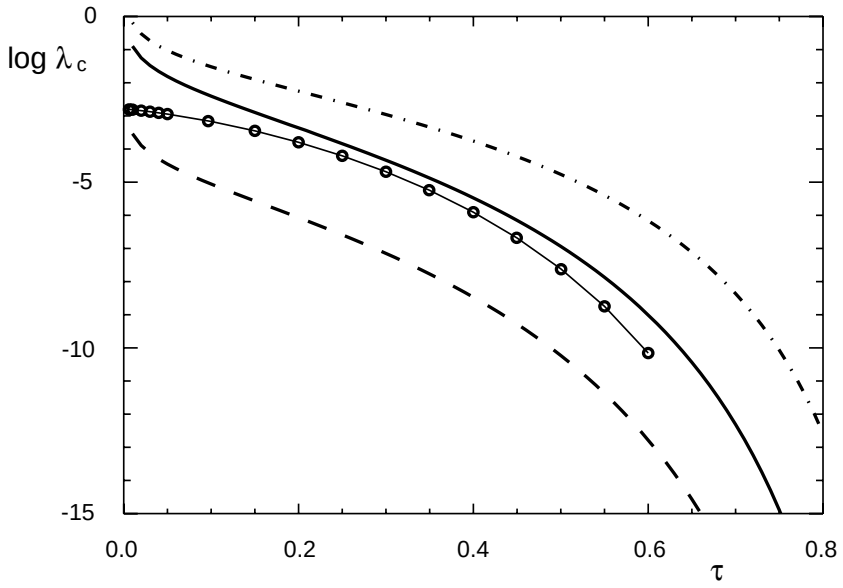


FIG. 6. *The logarithm of the Hopf frequency as a function of τ. Dots connected with a thin solid curve correspond to numerical results. Other lines represent the analytical predictions obtained with* (1) *dash line—Evans function asymptotic from section* 6, (2) *dash-dot line—truncation model,* (3) *thick solid line—generalized MAE.*

the numerical data for $\tau \in [0.1, 0.5]$. In Figure 6 we plot the logarithm of the Hopf frequency as a function of τ for $\beta = \beta_c$. As in the previous figure, the generalized MAE estimation best fits the numerical results.

Finally, using the compound matrix method, we can find the eigenmodes of the system (4.2)–(4.3) by inverting the relations (5.12) for $\mathbf{V}^-$ and similarly for $\mathbf{V}^+$. In Figures 7 and 8 we show the eigenmodes obtained for $\tau = 0.3$ for the critical value $\beta_c = 9.7404$, when two eigenvalues lie on the imaginary axis. These eigenmodes can be used, for example, in perturbation analysis when we want to investigate properties of the solution, bifurcating from the steady propagating front (see [24] for details).

8. Conclusion. The speed of planar combustion fronts was investigated using the shooting and relaxation methods for different values of Lewis number and different values of β (the ratio of the activation energy to heat release). We have compared the numerical results with the asymptotic estimation of the speed of the combustion front. Derivation of the speed estimation is given in section 3 and agrees with the results of [5, 7, 8]. The correspondence between the numerical and analytical results is good for large β and $\tau \sim 1$, whereas for moderate values of β the difference becomes significant. This is expected, as the asymptotic formula for speed of the front was derived in the leading order of the asymptotic expansion with β being a small parameter. It is also important to note that as we decrease the value of τ up to the order of β^{-1} the asymptotic prediction of the front speed becomes unsatisfactory even for sufficiently large values of β. This reveals the fact that, in the asymptotic treatment of the problem, τ was considered to be of the order of units, and therefore the asymptotic approach fails to work for $\tau \sim \beta^{-1}$. It is interesting to note that the dependence of speed on β seems to be correct, whereas the dependence of c on τ is valid only for $\tau \sim 1$. Excellent correspondence of the numerical results obtained with the integration of the

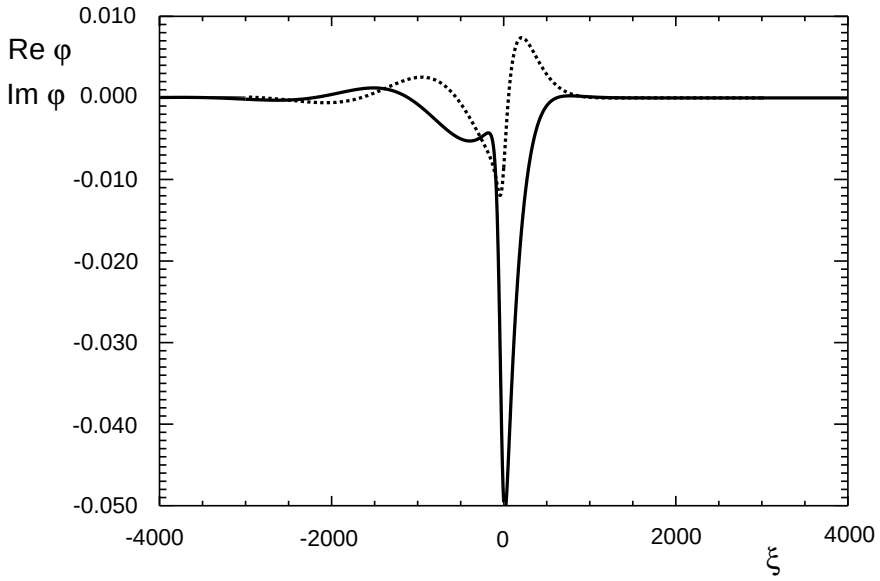


FIG. 7. *The eigenmode $\varphi(\xi)$ of the system* (4.2)–(4.3) *for $\tau = 0.3$, $\beta = \beta_c = 9.7404$, and purely complex eigenvalue $\lambda = i\lambda_c$. The solid line corresponds to Re φ, and the dashed line denotes Im φ.*

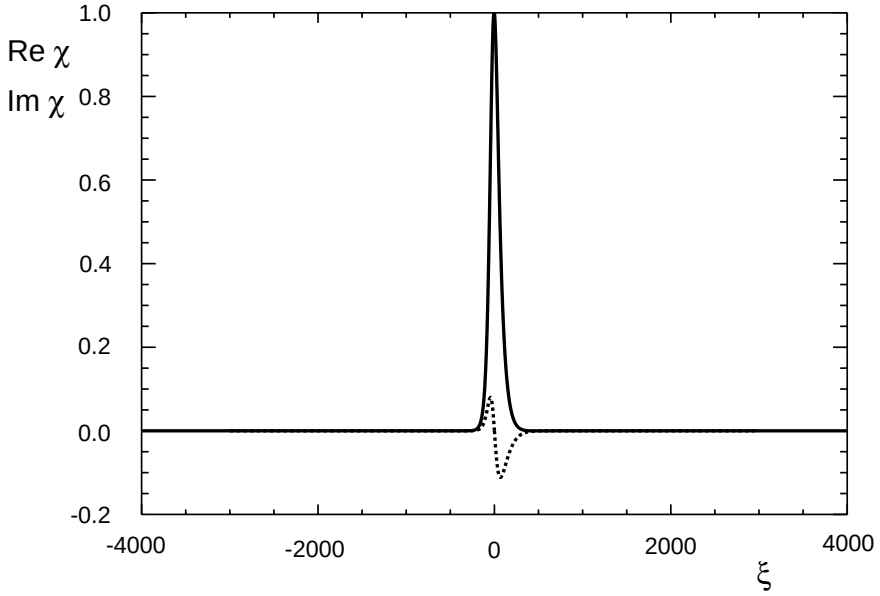


FIG. 8. *The eigenmode $\chi(\xi)$ of the system* (4.2)–(4.3) *for $\tau = 0.3$, $\beta = \beta_c = 9.7404$, and purely complex eigenvalue $\lambda = i\lambda_c$. The solid line corresponds to Re χ, and the dashed line denotes Im χ.*

governing PDE and corresponding ODE supports the credibility of both methods (the difference was found in the third significant digit). At the same time, we believe that the shooting and relaxation methods are preferable, as they require fewer computer resources, are more accurate, and do not depend on the stability of the solution.

The Evans function method was employed to examine the linear stability problem.

In this paper we consider only pulsating instabilities. Methods used in this paper are able to treat cellular instabilities as well (see [14]), and this is the subject of a separate paper. It was shown in section 7 that the conventional method for calculating the Evans function fails in our case. We demonstrated that the compound matrix method significantly expands the applicability of the Evans function approach. The results obtained with the compound matrix method were compared to the predictions of the asymptotic models. It appears that classic asymptotic models derived with truncated series are able to give only qualitative behavior for the front stability. We cannot expect the model derived in [4, 5] to give good quantitative results for $\tau \in [0, 0.6]$ considered in this paper, as it is valid for $\tau \simeq 1$. We would like to mention that it is very difficult to check the results of the asymptotic analysis for $\tau \to 1$ both numerically and experimentally, because the pulsating instability manifests itself only for extremely large values of β in this case. To the best of our knowledge β values of less than 30, considered in this paper while analyzing the stability of the travelling front, cover most of the combustion reactions. It turns out that the results obtained with the generalized matched asymptotic expansion method in [9] best correspond to the numerical data found in the present paper.

The compound matrix method not only allowed us to obtain a simple numerical criterion for the transition to instability for steady propagating solutions, but also provided us with more detailed information about the eigenvalues and eigenmodes. We located the eigenvalues, responsible for transition to instability, on the complex plane and found corresponding eigenmodes for the linear stability problem. This information can be useful, for example, in the analysis of the bifurcating solutions and their stability.

Finally, we would like to say a few words regarding the ambient temperature, which was taken to be equal to zero in this paper. The main disadvantage of this approach is the fact that the ambient temperature is a convenient control parameter used in experiments. This parameter is also ruled out in the asymptotic theories mentioned earlier. Therefore it is of clear interest to investigate the effect of this parameter on the stability of the combustion front. In the future we will be applying the methods described in this paper to models with finite ambient temperature and taking heat loss into consideration.

Acknowledgments. The authors thank S. J. A. Malham and K. Y. Kolossovski for helpful discussions, and the anonymous referees for their useful comments.

REFERENCES

[1] G. I. Sivashinsky, *Structure of Busen flames*, J. Chem. Phys., 62 (1975), pp. 638–643.
[2] G. I. Sivashinsky, *Diffusional-thermal theory of cellular flames*, Combust. Sci. Technol., 15 (1977), pp. 137–146.
[3] B. J. Matkowsky and G. I. Sivashinsky, *Propagation of a pulsating reaction front in solid fuel combustion*, SIAM J. Appl. Math., 35 (1978), pp. 465–478.
[4] B. J. Matkowsky and D. O. Olagunju, *Propagation of a pulsating flame front in a gaseous combustible mixture*, SIAM J. Appl. Math., 39 (1980), pp. 290–300.
[5] S. B. Margolis and B. J. Matkowsky, *Nonlinear stability and bifurcation in the transition from laminar to turbulent flame propagation*, Combust. Sci. Technol., 34 (1983), pp. 45–77.
[6] W. B. Bush and F. E. Fendell, *Asymptotic analysis of laminar flame propagation for general Lewis numbers*, Combust. Sci. Technol., 1 (1970), pp. 421–428.
[7] R. O. Weber, G. N. Mercer, H. S. Sidhu, and B. F. Gray, *Combustion waves for gases ($Le = 1$) and solids ($Le \to \infty$)*, Proc. Roy. Soc. London Ser. A, 453 (1997), pp. 1105–1118.
[8] A. C. McIntosh, R. O. Weber, and G. N. Mercer, *Non-adiabatic combustion waves for general Lewis numbers: Wave speed and extinction conditions*, ANZIAM J., submitted.

[9] D. A. Schult, *Matched asymptotic expansions and the closure problem for combustion waves*, SIAM J. Appl. Math., 60 (1999), pp. 136–155.
[10] J. W. Evans, *Nerve axon equations,* IV: *The stable and unstable impulse*, Indiana Univ. Math. J., 24 (1975), pp. 1169–1190.
[11] T. J. Bridges and G. Derks, *The symplectic Evans matrix, and the instability of solitary waves and fronts with symmetry*, Arch. Ration. Mech. Anal., 156 (2001), pp. 1–87.
[12] A. L. Afendikov and T. J. Bridges, *Instability of the Hocking–Stewartson pulse and its implications for three-dimensional Poiseuille flow*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., 457 (2001), pp. 1–16.
[13] L. A. Allen and T. J. Bridges, *Numerical exterior algebra and the compound matrix method*, Numer. Math., 92 (2002), pp. 197–232.
[14] N. J. Balmforth, R. V. Craster, and S. J. A. Malham, *Unsteady fronts in an autocatalytic system*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., 455 (1999), pp. 1401–1433.
[15] D. Terman, *Stability of planar wave solutions to a combustion model*, SIAM J. Math. Anal., 21 (1990), pp. 1139–1171.
[16] R. L. Pego and M. I. Weinstein, *Evans' function, Melnikov's integral, and solitary wave instabilities*, in Differential Equations with Applications to Mathematical Physics, W. F. Ames, E. M. Harrell II, and J. V. Herod, eds., Academic Press, San Diego, 1993, pp. 273–286.
[17] D. E. Pelinovsky, Y. S. Kivshar, and V.V. Afanasjev, *Internal modes of envelope solitons*, Phys. D, 116 (1998), pp. 121–142.
[18] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer-Verlag, New York, 1983.
[19] R. L. Pego, P. Smereka, and M. I. Weinstein, *Oscillatory instability of traveling waves for a KdV-Burgers equation*, Phys. D, 67 (1993), pp. 45–65.
[20] B. S. Ng and W. H. Reid, *An initial value method for eigenvalue problems using compound matrices*, J. Comput. Phys., 30 (1979), pp. 125–136.
[21] B. S. Ng and W. H. Reid, *The compound matrix method for ordinary differential equation*, J. Comput. Phys., 58 (1985), pp. 209–228.
[22] P. G. Drazin and W. H. Reid, *Hydrodynamic Stability*, Cambridge University Press, London, 1981.
[23] V. Gubernov, G. N. Mercer, H. S. Sidhu, and R. O. Weber, *Numerical methods for the analysis of travelling waves in reaction-diffusion equations*, ANZIAM J(E), to appear.
[24] A. I. Volpert, V. A. Volpert, and V. A. Volpert, *Travelling Wave Solutions of Parabolic Systems*, Transl. Math. Monogr. 140, AMS, Providence, RI, 1994.
[25] J. Alexander, R. Gardner, and C. Jones, *A topological invariant arising in the stability analysis of travelling waves*, J. Reine Angew. Math., 410 (1990), pp. 167–212.

SIAM J. APPL. MATH.
Vol. 63, No. 4, pp. 1276–1291

ANISOTROPIC POLARIZATION TENSORS AND DETECTION OF AN ANISOTROPIC INCLUSION*

HYEONBAE KANG†, EUNJOO KIM†, AND KYOUNGSUN KIM†

Abstract. We introduce the notion of (generalized) anisotropic polarization tensors and prove that they are symmetric and positive-definite. We also estimate the eigenvalues of the anisotropic polarization tensor in terms of the volume of the given domain. We apply these properties to an electrical impedance tomography problem to detect a single anisotropic inhomogeneity of small size. The goal is to detect an unknown inclusion when anisotropic conductivities of the background and the inclusion are known. Three results of computational experiments for the detection of an inclusion are given. One experiment is to assure the validity of the algorithm, while the other two are to see how anisotropy plays a role in the detection procedure.

Key words. anisotropic conductivity, anisotropic polarization tensor, asymptotic expansion, far-field, electrical impedance tomography

AMS subject classification. 35B30

PII. S003613990240619X

1. Introduction.

Let B be a bounded Lipschitz domain in $\mathbb{R}^d$, $d = 2, 3$. Suppose that the conductivity of B is different from that of background, $\mathbb{R}^d \setminus B$; the polarization tensor (PT) describes to the first order changes in the voltage potential due to the presence of the inhomogeneity B. For cases when the conductivities are isotropic, i.e., when they are independent of direction, the concept of PT was introduced by Schiffer and Szegö [18] and Pólya and Szegö [17]. If the conductivity is zero, namely, if B is insulated, the PT tensor is called the virtual mass. This concept of PT was extensively studied by many authors [17], [13], [8], [16], [6] for various purposes. In [6], Cedio-Fengya, Moskow, and Vogelius proved the symmetry and positive-definiteness of PT, and these properties became essential ingredients in recent development on the electrical impedance tomography (EIT) problem for detecting small inhomogeneities; see, for example, Cedio-Fengya, Moskow, and Vogelius [6], Brühl, Hanke, and Vogelius [5], and Ammari and Seo [4]. There are other works on locating inclusions from EIT data, such as [9], [10], [19]. The concept of PT was recently generalized by Ammari and Kang to the higher orders for the purpose of derivation of the complete asymptotic expansion of the voltage potential, and some of the important properties of PT were obtained [1], [2].

In this paper, we introduce the notion of (generalized) *anisotropic polarization tensor* (APT) and establish the symmetry and positive-definiteness of the first order tensor. We note that, in the anisotropic case, polarization occurs due to not only the presence of discontinuity, but also the difference of the anisotropy. These tensors are defined in the same way as the generalized isotropic polarization tensor in [1] and [2]. We also estimate the eigenvalues of the APT in terms of the volume of B.

We then apply these properties of APT to the EIT problem to detect an inhomogeneity inclusion with anisotropic conductivity. Let Ω be a bounded domain in

*Received by the editors April 23, 2002; accepted for publication (in revised form) October 7, 2002; published electronically April 9, 2003. This research was partly supported by KOSEF 98-0701-03-5 and BK21 at the School of Mathematical Sciences of Seoul National University.

http://www.siam.org/journals/siap/63-4/40619.html

†School of Mathematical Sciences, Seoul National University, Seoul 151-747, Korea (hkang@math.snu.ac.kr, kej@math.snu.ac.kr, kgsun@math.snu.ac.kr)

$\mathbb{R}^d$, $d = 2, 3$, with a connected Lipschitz boundary $\partial\Omega$. Suppose that Ω contains a single small inhomogeneity D of the form $D = z + \epsilon B$, where B is a bounded Lipschitz domain in $\mathbb{R}^d$ containing the origin, ϵ is small and approximately the order of magnitude of D, and z is a point in D. We assume that D is well separated from the boundary of Ω, namely, there exists a constant $c_0 > 0$ such that $\text{dist}(z, \partial\Omega) \geq c_0$. We also assume that the background $\Omega \setminus D$ has homogeneous anisotropic conductivity A, and D has homogeneous anisotropic conductivity $\tilde{A}$, where A and $\tilde{A}$ are constant $d \times d$ positive-definite symmetric matrices. Thus the conductivity profile of the body Ω is

$$\gamma(x) = \chi(\Omega \setminus D)A + \chi(D)\tilde{A}, \tag{1.1}$$

where $\chi(D)$ is the characteristic function of D. We also assume that $A - \tilde{A}$ is either positive- or negative-definite. (For this, see the remark immediately following the proof of Lemma 3.2.) The EIT problem we consider is to determine D from a finite number of pairs of applied current and measured voltage on $\partial\Omega$, when A and $\tilde{A}$ are known.

Let N be the unit normal to $\partial\Omega$. For a given $g \in L^2_0(\partial\Omega) := \{f \in L^2(\partial\Omega) : \int_{\partial\Omega} f d\sigma = 0\}$, let u_ϵ denote the steady-state voltage potential in the presence of the conductivity inhomogeneities, i.e., the solution to

$$\begin{cases} \nabla \cdot (\gamma(x)\nabla u(x)) = 0 & \text{in } \Omega, \\ \langle A\nabla u, N\rangle|_{\partial\Omega} = g & \left(\int_{\partial\Omega} u d\sigma = 0\right). \end{cases} \tag{1.2}$$

We first consider asymptotic expansion of u_ϵ as $\epsilon \to 0$ in terms of the background potential and APTs. The background potential U is the steady-state voltage potential in the absence of the conductivity inhomogeneities, i.e., the solution to

$$\begin{cases} \nabla \cdot (A\nabla U) = 0 & \text{in } \Omega, \\ \langle A\nabla U, N\rangle|_{\partial\Omega} = g & \left(\int_{\partial\Omega} U d\sigma = 0\right). \end{cases} \tag{1.3}$$

We follow the lines in [1] to derive the asymptotic formula of u_ϵ on $\partial\Omega$ (Theorem 4.1). Then using a simple formula found in [1] relating the Neumann function and the fundamental solution, we convert this asymptotic formula to that of a function determined by boundary measurements outside Ω to calculate the far-field relation (Theorem 4.3). We then use this relation to derive an algorithm for finding APT, the order of magnitude, and z. In this process, those properties of APT proved in earlier sections will play essential roles. This method is similar to the one proposed in [3] for detecting single isotropic inhomogeneity. We also present a different algorithm for finding the center z based on an idea of [15] and [4].

Using the algorithm, we perform three computational experiments. The first one is computational validation of the reconstruction formula with and without noise. The formula or algorithm of reconstruction is derived under the assumption that $\tilde{A} - A$ is either positive- or negative-definite. However, there are practical situations in which the conductivity is high in one direction and low in another. The second experiment is to see what happens if $\tilde{A} - A$ has eigenvalues of mixed signs. The results of this experiment show that, even in this case, algorithms find the inclusion fairly well. Unlike the isotropic case, there is difficulty in understanding what the

limiting case of extreme conductivity is. The purpose of the third experiment is to see what happens if the condition number of $\tilde{A} - A$ is very large. The result indicates that if the condition number of $\tilde{A} - A$ becomes large, so does the error of the reconstruction. During these experiments, we compare the performance of two algorithms. Theoretically, the first algorithm gives better precision of $O(\epsilon^d)$ in finding the center. However, computational results show that the second algorithm works better under the presence of noise in the data.

This paper is organized as follows. In section 2, we briefly review the layer potentials for the operator $\nabla \cdot A\nabla$. In section 3, APTs are defined, and symmetry, positive-definiteness, and estimation of eigenvalues are proved. In section 4, we derive reconstruction formulas for finding APT, the order of magnitude of the inclusion, and the center. Section 5 is for presentation of the computational experiments.

2. Layer potentials. Let D be a bounded domain in $\mathbb{R}^d, d = 2, 3$. We assume that ∂D is Lipschitz. Let A be a positive-definite symmetric matrix and A_* be the positive-definite symmetric matrix such that $A^{-1} = A_*^2$. Let $\Gamma(x)$ be the fundamental solution of the operator $\nabla \cdot A\nabla$:

$$\Gamma(x) = \begin{cases} \dfrac{1}{2\pi\sqrt{|A|}} \ln \|A_* x\|, & d = 2, \\ -\dfrac{1}{4\pi\sqrt{|A|}\|A_* x\|}, & d = 3, \end{cases} \tag{2.1}$$

where $|A|$ is the determinant of A. The single and double layer potentials associated with A of the density function ϕ on D are defined by

$$\mathcal{S}_D\phi(x) := \int_{\partial D} \Gamma(x - y)\phi(y)d\sigma(y), \quad x \in \mathbb{R}^d, \tag{2.2}$$

$$\mathcal{D}_D\phi(x) := \int_{\partial D} \frac{\partial}{\partial \nu_y}\Gamma(x - y)\phi(y)d\sigma(y), \quad x \in \mathbb{R}^d \setminus \partial D, \tag{2.3}$$

where $\frac{\partial u}{\partial \nu} := \langle AN, \nabla u\rangle$, and N is the outward unit normal to ∂D.

We now suppose that D is compactly contained in a bounded domain Ω. Suppose that the background conductor $\Omega \setminus D$ has anisotropic conductivity A, and D has $\tilde{A}$. We always suppose that $\tilde{A} - A$ is either positive-definite or negative-definite. $\tilde{\mathcal{S}}_D$ and $\mathcal{S}_D$ denote the single layer potentials on D corresponding to $\tilde{A}$ and A, respectively. We also denote $\frac{\partial u}{\partial \tilde{\nu}} := \langle \tilde{A}N, \nabla u\rangle$. The subscript $+$ in the notation $u|_+$ denotes the limit to the boundary from the outside of a given domain, and $-$ indicates the limit from the inside. The following result of Escauriaza and Seo [7] is an essential ingredient of the present paper.

THEOREM 2.1. *For each $(F, G) \in L_1^2(\partial D)\times L^2(\partial D)$, there exists a unique solution $(f, g) \in L^2(\partial D) \times L^2(\partial D)$ of the integral equation*

$$\begin{cases} \tilde{\mathcal{S}}_D f - \mathcal{S}_D g = F \\ \dfrac{\partial}{\partial \tilde{\nu}}\tilde{\mathcal{S}}_D f\Big|_- - \dfrac{\partial}{\partial \nu}\mathcal{S}_D g\Big|_+ = G \end{cases} \quad \text{on } \partial D. \tag{2.4}$$

Moreover, there exists a constant C depending on only the largest and smallest eigenvalues of $\tilde{A}$, A, and $\tilde{A} - A$, and the Lipschitz character of D such that

$$\|f\|_{L^2(\partial D)} + \|g\|_{L^2(\partial D)} \le C(\|F\|_{L_1^2(\partial D)} + \|G\|_{L^2(\partial D)}). \tag{2.5}$$

3. The APT. We will be using the usual notation for multiindices: for multiindex $\alpha = (\alpha_1, \dots, \alpha_d)$, $x^\alpha := x_1^{\alpha_1} \cdots x_d^{\alpha_d}$, $\partial^\alpha := \partial_1^{\alpha_1} \cdots \partial_d^{\alpha_d}$, etc.

As in [1], (generalized) APTs are defined as follows.

DEFINITION 3.1. *Let B be a bounded Lipschitz domain in $\mathbb{R}^d$, and $\tilde{A}$ and A be the anisotropic conductivities of B and $\mathbb{R}^d \setminus \overline{B}$, respectively. For a multiindex α, let $(f_\alpha, g_\alpha) \in L^2(\partial B) \times L^2(\partial B)$ be the unique solution of*

$$\begin{cases} \tilde{\mathcal{S}}_B f_\alpha - \mathcal{S}_B g_\alpha = x^\alpha \\ \dfrac{\partial}{\partial \tilde{\nu}} \tilde{\mathcal{S}}_B f_\alpha \Big|_- - \dfrac{\partial}{\partial \nu} \mathcal{S}_B g_\alpha \Big|_+ = \frac{\partial}{\partial \nu} x^\alpha \end{cases} \quad \text{on } \partial B. \tag{3.1}$$

For a pair of multiindices α, β, define the anisotropic polarization tensors *(APTs) associated with the domain B and anisotropic conductivities $\tilde{A}$ and A by*

$$m_{\alpha\beta} = \int_{\partial B} y^\beta g_\alpha(y) d\sigma. \tag{3.2}$$

The first order APT ($|\alpha| = |\beta| = 1$) is particularly important. In this section, we prove those properties of the first order APT which play significant roles in detection of an inhomogeneity inclusion, such as symmetry, positive-definiteness, and estimation of the eigenvalues. When $|\alpha| = |\beta| = 1$, denote $m_{\alpha\beta}$ by m_{ij}, $i, j = 1, \dots, d$, i.e.,

$$m_{ij} = \int_{\partial B} y_j g_i(y) d\sigma, \tag{3.3}$$

where $f_i = f_\alpha$ and $g_i = g_\alpha$ with $\alpha = e_i$. Here and throughout this section we drop the subscript B from the notation of single layer potentials. Define

$$\phi_i(x) := \begin{cases} \mathcal{S} g_i(x), & x \in \mathbb{R}^d \setminus B, \\ \tilde{\mathcal{S}} f_i(x), & x \in B. \end{cases}$$

Then ϕ_i is the unique solution of the following transmission problem:

$$\begin{cases} \nabla \cdot (A \nabla \phi_i) = 0 & \text{in } \mathbb{R}^d \setminus B, \\ \nabla \cdot (\tilde{A} \nabla \phi_i) = 0 & \text{in } B, \\ \phi_i|_- - \phi_i|_+ = x_i & \text{on } \partial B, \\ \dfrac{\partial \phi_i}{\partial \tilde{\nu}} \Big|_- - \dfrac{\partial \phi_i}{\partial \nu} \Big|_+ = \dfrac{\partial x_i}{\partial \nu} & \text{on } \partial B, \\ \phi_i(x) = O(|x|^{-d+1}) & \text{as } |x| \to \infty. \end{cases}$$

By the jump relation of the single layer potential, $g_i = \frac{\partial (\mathcal{S} g_i)}{\partial \nu}\big|_+ - \frac{\partial (\mathcal{S} g_i)}{\partial \nu}\big|_-$. Thus it

follows from (3.1) that

$$\begin{aligned}\int_{\partial B} y_j g_i d\sigma &= \int_{\partial B} y_j \left[\frac{\partial(\mathcal{S}g_i)}{\partial \nu}\Big|_+ - \frac{\partial(\mathcal{S}g_i)}{\partial \nu}\Big|_- \right] d\sigma \\ &= \int_{\partial B} y_j \left[\frac{\partial(\tilde{\mathcal{S}}f_i)}{\partial \tilde{\nu}}\Big|_- - \frac{\partial y_i}{\partial \nu} - \frac{\partial(\mathcal{S}g_i)}{\partial \nu}\Big|_- \right] d\sigma \\ &= \int_{\partial B} \left[\frac{\partial y_j}{\partial \tilde{\nu}} \tilde{\mathcal{S}} f_i - \frac{\partial y_j}{\partial \nu}\big(y_i + \mathcal{S}g_i\big) \right] d\sigma \\ &= \int_{\partial B} \left[\frac{\partial y_j}{\partial \tilde{\nu}} - \frac{\partial y_j}{\partial \nu} \right] \phi_i|_-(y) d\sigma.\end{aligned}$$

Therefore,

$$m_{ij} = \left\langle (\tilde{A} - A)e_j, \int_B \nabla \phi_i(y) dy \right\rangle, \tag{3.4}$$

or equivalently,

$$m_{ij} = \langle (\tilde{A} - A)e_j, e_i \rangle |B| + \int_{\partial B} \langle (\tilde{A} - A)N, e_j \rangle \phi_i|_+(y) d\sigma. \tag{3.5}$$

In particular, if $\tilde{A}$ and A are isotropic, say $\tilde{A} = \tilde{\mu} I$ and $A = \mu I$, then one can easily check that

$$m_{ij} = \left(1 - \frac{\mu}{\tilde{\mu}}\right) \left[\tilde{\mu} \delta_{ij} |B| + \tilde{\mu} \int_{\partial B} y_j \left. \frac{\partial \phi_i}{\partial \nu} \right|_+ \right],$$

which is the polarization tensor studied in [6].

Let $M = (m_{ij})$. Suppose that $d = 3$ for convenience. The two dimensional case can be treated in the same way without change. Let A_* be the unique positive-definite symmetric matrix such that $A^{-1} = A_*^2$. Then $A_* \tilde{A} A_*$ is also symmetric positive-definite. Let λ_j and v_j, $j = 1, 2, 3$, be eigenvalues and corresponding eigenvectors of $A_* \tilde{A} A_*$, respectively. Let $w_j = A_* v_j$, $j = 1, 2, 3$. Since $A_* \tilde{A} A_* v_j = \lambda_j v_j$, we have $\tilde{A} w_j = \lambda_j A w_j$. Therefore, we get

$$(\tilde{A} - A) w_j = (\lambda_j - 1) A w_j = \frac{\lambda_j - 1}{\lambda_j} \tilde{A} w_j. \tag{3.6}$$

Note that $\lambda_j > 1$ (< 1) if $\tilde{A} - A$ is positive (negative)-definite. Let $f_i^* := w_i \cdot (f_1, f_2, f_3)$ and $g_i^* := w_i \cdot (g_1, g_2, g_3)$, $i = 1, 2, 3$. Then (f_i^*, g_i^*) is the unique solution of

$$\begin{cases} \tilde{\mathcal{S}} f_i^*(x) - \mathcal{S} g_i^*(x) &= w_i \cdot x \\ \left. \dfrac{\partial}{\partial \tilde{\nu}} \tilde{\mathcal{S}} f_i^* \right|_-(x) - \left. \frac{\partial}{\partial \nu} \mathcal{S} g_i^* \right|_+(x) &= \dfrac{\partial}{\partial \nu}(w_i \cdot x) \end{cases} \quad \text{on } \partial B. \tag{3.7}$$

For a domain D and positive-definite symmetric matrix A, define a bilinear form

$$(f, g)_D^A := \int_D \langle A \nabla f, \nabla g \rangle dx,$$

and let Q_D^A denote the corresponding quadratic form.

LEMMA 3.2.

$$\int_{\partial B}(w_j\cdot y)g_i^*d\sigma=\int_{\partial B}(w_i\cdot y)g_j^*d\sigma,\quad i,j=1,2,3, \tag{3.8}$$

and

$$\begin{aligned}\frac{\lambda_j+1}{\lambda_j-1}\int_{\partial B}(w_j\cdot y)g_i^*d\sigma&=(\tilde{\mathcal{S}}f_j^*,\tilde{\mathcal{S}}f_i^*)_B^{\tilde{A}}+(\mathcal{S}g_j^*,\mathcal{S}g_i^*)_{\mathbb{R}^3\setminus B}^A\\&\quad+(w_j\cdot y,w_i\cdot y)_B^A-(\mathcal{S}g_j^*,w_i\cdot y)_B^A+(w_j\cdot y,\mathcal{S}g_i^*)_B^A.\end{aligned} \tag{3.9}$$

Proof. In the same way as before, we can show that

$$\int_{\partial B}(w_j\cdot y)g_i^*d\sigma=\int_{\partial B}\left[\frac{\partial(w_j\cdot y)}{\partial\tilde{\nu}}-\frac{\partial(w_j\cdot y)}{\partial\nu}\right]\tilde{\mathcal{S}}f_i^*d\sigma. \tag{3.10}$$

Note that, by (3.6),

$$\frac{\partial(w_j\cdot y)}{\partial\tilde{\nu}}-\frac{\partial(w_j\cdot y)}{\partial\nu}=\langle N,(\tilde{A}-A)w_j\rangle=\begin{cases}(\lambda_j-1)\dfrac{\partial(w_j\cdot y)}{\partial\nu},\\[2mm]\dfrac{\lambda_j-1}{\lambda_j}\dfrac{\partial(w_j\cdot y)}{\partial\tilde{\nu}}.\end{cases} \tag{3.11}$$

It then follows from (3.7) and the first relation in (3.11) that

$$\begin{aligned}\frac{1}{\lambda_j-1}\int_{\partial B}(w_j\cdot y)g_i^*d\sigma&=\int_{\partial B}\frac{\partial(w_j\cdot y)}{\partial\nu}\tilde{\mathcal{S}}f_i^*d\sigma\\&=\int_{\partial B}\frac{\partial(w_j\cdot y+\mathcal{S}g_j^*)}{\partial\nu}\bigg|_-(w_i\cdot y+\mathcal{S}g_i^*)d\sigma-\int_{\partial B}\frac{\partial\mathcal{S}g_j^*}{\partial\nu}\bigg|_-(w_i\cdot y+\mathcal{S}g_i^*)d\sigma\\&=\int_{\partial B}\frac{\partial(w_j\cdot y+\mathcal{S}g_j^*)}{\partial\nu}\bigg|_-(w_i\cdot y+\mathcal{S}g_i^*)d\sigma-\int_{\partial B}(\mathcal{S}g_j^*)\frac{\partial(w_i\cdot y+\mathcal{S}g_i^*)}{\partial\nu}\bigg|_-d\sigma.\end{aligned} \tag{3.12}$$

On the other hand, by (3.7) and the second relation in (3.11), we get

$$\begin{aligned}\frac{\lambda_j}{\lambda_j-1}\int_{\partial B}(w_j\cdot y)g_i^*d\sigma&=\int_{\partial B}\frac{\partial(w_j\cdot y)}{\partial\tilde{\nu}}\tilde{\mathcal{S}}f_i^*d\sigma=\int_{\partial B}(w_j\cdot y)\,\frac{\partial(\tilde{\mathcal{S}}f_i^*)}{\partial\tilde{\nu}}\bigg|_-d\sigma\\&=\int_{\partial B}\tilde{\mathcal{S}}f_j^*\,\frac{\partial(\tilde{\mathcal{S}}f_i^*)}{\partial\tilde{\nu}}\bigg|_--\int_{\partial B}\mathcal{S}g_j^*\,\frac{\partial\mathcal{S}g_i^*}{\partial\nu}\bigg|_+-\int_{\partial B}\mathcal{S}g_j^*\frac{\partial(w_i\cdot y)}{\partial\nu}.\end{aligned} \tag{3.13}$$

Subtracting (3.12) from (3.13), we get

$$\begin{aligned}\int_{\partial B}(w_j\cdot y)g_i^*d\sigma&=-\int_{\partial B}\frac{\partial(w_j\cdot y+\mathcal{S}g_j^*)}{\partial\nu}\bigg|_-(w_i\cdot y+\mathcal{S}g_i^*)\\&\quad+\int_{\partial B}\tilde{\mathcal{S}}f_j^*\,\frac{\partial(\tilde{\mathcal{S}}f_i^*)}{\partial\tilde{\nu}}\bigg|_--\int_{\partial B}\mathcal{S}g_j^*\left[\frac{\partial\mathcal{S}g_i^*}{\partial\nu}\bigg|_+-\frac{\partial\mathcal{S}g_i^*}{\partial\nu}\bigg|_-\right]\\&=-(w_j\cdot y+\mathcal{S}g_j^*,w_i\cdot y+\mathcal{S}g_i^*)_B^A+(\tilde{\mathcal{S}}f_j^*,\tilde{\mathcal{S}}f_i^*)_B^{\tilde{A}}-\int_{\partial B}\mathcal{S}g_j^*g_i^*d\sigma.\end{aligned}$$

Since $\int_B \mathcal{S}g_j^* g_i^* d\sigma = \int_B g_j^* \mathcal{S}g_i^* d\sigma$, we have (3.8).

To prove (3.9) we write (3.12) in slightly different way: By (3.7), we get

$$\frac{1}{\lambda_j - 1}\int_{\partial B}(w_j \cdot y) g_i^* d\sigma = \int_{\partial B}\frac{\partial (w_j \cdot y)}{\partial \nu}\mathcal{S}g_i^* d\sigma + \int_{\partial B}\frac{\partial (w_j \cdot y)}{\partial \nu}(w_i \cdot y) d\sigma. \tag{3.14}$$

Adding (3.13) with (3.14), we get (3.9). This completes the proof. □

Remark. Let $M := (m_{ij})$. If $\tilde{A} - A$ is positive- or negative-semidefinite, not definite, then some $\lambda_j = 1$. Suppose $\lambda_j = 1$. Then (3.10) and (3.6) show that

$$\langle w_i, M w_j\rangle = \int_{\partial B}(w_j \cdot y) g_i^* d\sigma = 0, \quad i = 1, 2, 3.$$

Since w_1, w_2, w_3 are linearly independent, $Mw_j = 0$. In particular, M is singular. This is the reason why we are assuming that $\tilde{A} - A$ is positive- or negative-definite.

THEOREM 3.3. *The anisotropic polarization tensor M is symmetric, and for each $a \neq 0 \in \mathbb{R}^d$ there exists $\bar{a} \in \mathbb{R}^d$ such that $\langle \bar{a}, Ma\rangle > 0$. In fact, if $a = \sum_i a_i^* w_i$, then*

$$\bar{a} = \sum_i \frac{\lambda_i + 1}{\lambda_i - 1} a_i^* w_i. \tag{3.15}$$

Proof. Let $a, b \in \mathbb{R}^3$, let $a = \sum_i a_i^* w_i$ and $b = \sum_i b_i^* w_i$. Then

$$\langle b, Ma\rangle = \int_{\partial B}(a \cdot y)(b \cdot g)(x) d\sigma = \sum_{i,j=1}^{d} a_j^* b_i^* \int_{\partial B}(w_j \cdot y)(w_i \cdot g)(y) d\sigma,$$

where $g = (g_1, g_2, g_3)$. Thus, by (3.8), we obtain

$$\begin{aligned}\langle b, Ma\rangle &= \sum_{i,j=1}^{d} a_j^* b_i^* \int_{\partial B}(w_j \cdot y)(w_i \cdot g)(y) d\sigma \\ &= \sum_{i,j=1}^{d} a_j^* b_i^* \int_{\partial B}(w_i \cdot y)(w_j \cdot g)(y) d\sigma \\ &= \int_{\partial B}(b \cdot y)(a \cdot g)(y) d\sigma = \langle a, Mb\rangle.\end{aligned}$$

For a given a, define $\bar{a}$ as (3.15). Then

$$\langle \bar{a}, Ma\rangle = \sum_{i,j=1}^{d} \frac{\lambda_j + 1}{\lambda_j - 1} a_i^* a_j^* \int_{\partial B}(w_j \cdot y)(w_i \cdot g)(y) d\sigma.$$

Let $f_a = \sum_i a_i^* f_i^* = a \cdot f$ and $g_a = \sum_i a_i^* g_i^* = a \cdot g$. It then follows from (3.9) that

$$\begin{aligned}\langle \bar{a}, Ma\rangle &= (\tilde{\mathcal{S}} f_a, \tilde{\mathcal{S}} f_a)_B^{\tilde{A}} + (\mathcal{S} g_a, \mathcal{S} g_a)_{\mathbb{R}^3 \setminus B}^{A} \\ &\quad + (a \cdot y, a \cdot y)_B^A - (\mathcal{S} g_a, a \cdot y)_B^A + (a \cdot y, \mathcal{S} g_a)_B^A \\ &= Q_B^{\tilde{A}}(\tilde{\mathcal{S}} f_a) + Q_{\mathbb{R}^d \setminus B}^{A}(\mathcal{S} g_a) + \langle a, Aa\rangle |B|.\end{aligned} \tag{3.16}$$

Thus the proof is complete. □

Without loss of generality, let us suppose that $\tilde{A} - A$ is positive-definite so that $\lambda_j > 1$, $j = 1, 2, 3$. One can immediately observe from (3.16) that

$$C_1 \|a\|^2 |B| \leq \langle \bar{a}, Ma \rangle \tag{3.17}$$

for some constant C_1 depending only on A.

Fix a point $z \in B$. Since $\int_{\partial B} g_i(y) d\sigma = 0$, we get from the definition of APT that

$$\begin{aligned} |m_{ij}| &\leq \left(\int_{\partial B} (y_j - z_j)^2 d\sigma \right)^{1/2} \left(\int_{\partial B} |g_i(y)|^2 d\sigma \right)^{1/2} \\ &\leq \operatorname{diam}(B) |\partial B|^{1/2} \|g_i\|_{L^2(\partial B)}. \end{aligned}$$

It then follows from (2.5) that

$$|m_{ij}| \leq C \operatorname{diam}(B) |\partial B| \leq C|B|,$$

where the dependence of the constant C is the same as that in Theorem 2.1. Thus we obtain

$$\langle \bar{a}, Ma \rangle \leq C|B| \|a\| \|\bar{a}\| \leq C_2 |B| \|a\|^2$$

for some C_2. Therefore we get

$$C_1 \|a\|^2 |B| \leq \langle \bar{a}, Ma \rangle \leq C_2 \|a\|^2 |B|. \tag{3.18}$$

Define constant matrices W and Λ by

$$W := [w_1, w_2, w_3]^T \quad \text{and} \quad \Lambda := \operatorname{diag}[\mu_1, \mu_2, \mu_3], \quad \mu_j := \frac{\lambda_j + 1}{\lambda_j - 1}. \tag{3.19}$$

We also define

$$V := [v_1, v_2, v_3]^T. \tag{3.20}$$

Note that W is nonsingular, $W = A_* V$, and W and Λ are determined by A and $\tilde{A}$. One can easily check that $\bar{a}$ defined by (3.15) can be written as

$$\bar{a} = W^{-1} \Lambda W a. \tag{3.21}$$

Thus (3.18) reads

$$C_1 |B| \|a\|^2 \leq \langle W^{-1} \Lambda W a, Ma \rangle \leq C_2 |B| \|a\|^2.$$

Let $b = Wa$. Then we have

$$C_1 |B| \|b\|^2 \leq \langle W^{-1} \Lambda b, M W^{-1} b \rangle = \langle \Lambda b, (W^{-1})^T M W^{-1} b \rangle \leq C_2 |B| \|b\|^2. \tag{3.22}$$

Here constants C_1 and C_2 are different from previous ones. However, their dependence is the same. Let v be an eigenvector of the symmetric matrix $(W^{-1})^T M W^{-1}$ and μ be the corresponding eigenvalue. Then $\mu > 0$ and

$$\langle \Lambda v, (W^{-1})^T M W^{-1} v \rangle = \langle \Lambda v, \mu v \rangle = \mu \|\Lambda^{1/2} v\|^2.$$

It then follows from (3.22) that

$$C_1|B|\|v\|^2 \le \mu\|\Lambda^{1/2}v\|^2 \le C_2|B|\|v\|^2,$$

and hence

$$C_1|B| \le \mu \le C_2|B|. \tag{3.23}$$

Therefore, $(W^{-1})^T M W^{-1}$ is positive-definite and

$$C_1|B|\|b\|^2 \le \langle b, (W^{-1})^T M W^{-1} b\rangle \le C_2|B|\|b\|^2 \quad \forall b \in \mathbb{R}^d.$$

Hence, by letting $a = W^{-1}b$, we get

$$C_1|B|\|a\|^2 \le \langle a, Ma\rangle \le C_2|B|\|a\|^2 \quad \forall a \in \mathbb{R}^d.$$

As a consequence, we obtain the following estimation of the eigenvalues of M.

THEOREM 3.4. *If $\tilde{A} - A$ is positive(negative)-definite, so is M. If μ is an eigenvalue of M, then*

$$C_1|B| \le |\mu| \le C_2|B| \tag{3.24}$$

for some constants C_1 and C_2, whose dependence is the same as that in Theorem 2.1.

4. Detection of an inclusion. Recall that $D = z + \epsilon B$. Let $N(x,y)$, $x \in \partial\Omega$, $y \in \Omega$, be the Neumann function for $\nabla \cdot A\nabla$ on Ω, and $m_{\alpha\beta}$ the APTs. Following the same lines of derivation used in [1], one can prove the following asymptotic expansion of the solution u_ϵ to (1.2) on $\partial\Omega$.

THEOREM 4.1. *For $x \in \partial\Omega$,*

$$u_\epsilon(x) = U(x) - \epsilon^d \sum_{|\alpha|=1}^{d} \sum_{|\beta|=1}^{d-|\alpha|+1} \frac{\epsilon^{|\alpha|+|\beta|-2}}{\alpha!\beta!} \partial^\alpha U(z) m_{\alpha\beta} \partial_z^\beta N(x,z) + O(\epsilon^{2d}), \tag{4.1}$$

where the remainder $O(\epsilon^{2d})$ is dominated by $C\epsilon^{2d}$ for some C independent of $x \in \partial\Omega$ and z.

In fact, as in [1], one can represent the solution u_ϵ in terms of layer potentials and the Neumann functions, and then obtain the expansion (4.1). We note that if there are well-separated multiple inclusions, then an asymptotic formula can be obtained by adding formula (4.1).

We now define a function $H[g]$ for $g \in L_0^2(\partial\Omega)$ by

$$H[g](x) = -\mathcal{S}_\Omega(g)(x) + \mathcal{D}_\Omega(u_\epsilon|_{\partial\Omega})(x), \quad x \in \mathbb{R}^d \setminus \overline{\Omega}. \tag{4.2}$$

Substituting (4.1) into (4.2), one can see that

$$H[g](x) = -\mathcal{S}_\Omega(g)(x) + \mathcal{D}_\Omega(U|_{\partial\Omega})(x) - \epsilon^d \sum_{|\alpha|=1}^{d} \sum_{|\beta|=1}^{d-|\alpha|+1} \frac{\epsilon^{|\alpha|+|\beta|-2}}{\alpha!\beta!} \partial^\alpha U(z) m_{\alpha\beta} \partial_z^\beta \mathcal{D}_\Omega(N(\cdot,z))(x) + O\left(\frac{\epsilon^{2d}}{|x|^{d-1}}\right).$$

We then observe two simple but important facts. First, observe that $-\mathcal{S}_\Omega(g)(x) + \mathcal{D}_\Omega(U|_{\partial\Omega})(x) = 0$ if $x \in \mathbb{R}^d \setminus \overline{\Omega}$. Second, in the same way as that used for the proof of Lemma 2.3 of [1], we can show that

$$\mathcal{D}_\Omega(N(\cdot,z))|_+(x) = \Gamma(x-z) \quad \text{modulo constants} \quad \forall x \in \partial\Omega,\ \forall z \in \Omega. \tag{4.3}$$

Thus we obtain the following asymptotic expansion of $H[g]$ outside Ω.

THEOREM 4.2. *For $x \in \mathbb{R}^d \setminus \overline{\Omega}$,*

$$H[g](x) = -\epsilon^d \sum_{|\alpha|=1}^{d} \sum_{|\beta|=1}^{d-|\alpha|+1} \frac{\epsilon^{|\alpha|+|\beta|-2}}{\alpha!\beta!} \partial^\alpha U(z) m_{\alpha\beta} \partial_z^\beta \Gamma(x,z) + O\left(\frac{\epsilon^{2d}}{|x|^{d-1}}\right). \tag{4.4}$$

Suppose now that $g = \langle AN, a\rangle$ for a constant vector $a \in \mathbb{R}^d$. Then $U(x) = a \cdot x$ and the formula (4.4) takes the form

$$H[g](x) = -\epsilon^d \sum_{|\alpha|=1} \sum_{|\beta|=1}^{d} \frac{\epsilon^{|\beta|-1}}{\beta!} \partial^\alpha U(z) m_{\alpha\beta} \partial_z^\beta \Gamma(x,z) + O\left(\frac{\epsilon^{2d}}{|x|^{d-1}}\right). \tag{4.5}$$

Then by explicitly computing $\partial_z^\beta \Gamma(x,z)$, one can show that

$$H[g](x) = \frac{1}{\omega_d} \left\langle a, \epsilon^d M A_* \frac{A_*(x-z)}{|A_*(x-z)|^d} \right\rangle + O\left(\frac{\epsilon^d}{|x|^d}\right) + O\left(\frac{\epsilon^{2d}}{|x|^{d-1}}\right), \tag{4.6}$$

where $\omega_d = 2\pi$ if $d = 2$, and $\omega_d = 4\pi$ if $d = 3$, and $M = (m_{ij})$ is the first order APT.

For a general Neumann datum g, we have

$$H[g](x) = \frac{1}{\omega_d} \left\langle \nabla U(z), \epsilon^d M A_* \frac{A_*(x-z)}{|A_*(x-z)|^d} \right\rangle + O\left(\frac{\epsilon^d}{|x|^d}\right) + O\left(\frac{\epsilon^{d+1}}{|x|^{d-1}}\right). \tag{4.7}$$

Since $\frac{A_*(x-z)}{|A_*(x-z)|^d} = \frac{A_* x}{|A_* x|^d} + O(|x|^{-d})$, we obtain from (4.6) and (4.7) the following far-field relations.

THEOREM 4.3. *For $g \in L_0^2(\partial\Omega)$, let U_g be the solution of* (1.3). *Then, for $|x| = O(\epsilon^{-1})$,*

$$\omega_d |A_* x|^{d-1} H[g](x) = \left\langle \nabla U_g(z), \epsilon^d M A_* \frac{A_* x}{|A_* x|} \right\rangle \quad \textit{modulo } O(\epsilon^{d+1}). \tag{4.8}$$

If $g = \langle AN, a\rangle$, then for $|x| = O(\epsilon^{-d})$

$$\omega_d |A_* x|^{d-1} H[g](x) = \left\langle a, \epsilon^d M A_* \frac{A_* x}{|A_* x|} \right\rangle \quad \textit{modulo } O(\epsilon^{2d}). \tag{4.9}$$

We note that (4.8) is a general far-field relation, while (4.9) is a formula with better precision.

Using (4.6), (4.8), and (4.9), we can detect APT, the order of magnitude of D, and z.

Detection of APT. Now let $a = e_i$, or equivalently, $g = \langle AN, e_i\rangle$, and choose $b_j = O(\epsilon^{-d})$ so that

$$A_* \frac{A_* b_j}{|A_* b_j|} = e_j, \quad i, j = 1, \dots, d. \tag{4.10}$$

It then follows from (4.9) that

$$\epsilon^d m_{ij} = \omega_d |A_* b_j|^{d-1} H[g](b_j) \quad \text{modulo } O(\epsilon^{2d}). \tag{4.11}$$

Since ϵ is not known a priori, in actual computations we find unit vectors b_j satisfying (4.10) and then compute $\omega_d |t A_* b_j|^{d-1} H[g](t b_j)$ as $t \to \infty$. Since the first order APT is invariant under translation, as one can easily check, $\epsilon^d M$ is the first order APT for the domain D.

Detection of order of magnitude. Having found $\epsilon^d M$, we proceed to find the order of magnitude ϵ and the center z. Using (3.24), we can determine the order of magnitude of D. Let μ be the smallest (in absolute value) eigenvalue of $\epsilon^d M$. Then, by Theorem 3.4, $\epsilon^d|B| \approx |\mu|$. Considering the ambiguity of representation of $D = z + \epsilon B$, $\epsilon^d|B|$ or $\epsilon|B|^{1/d}$ seems the right notion of the order of magnitude.

Detection of center—Method 1. Let v_j, $j = 1, \dots, d$, be orthonormal eigenvectors of the symmetric matrix $A_*(\epsilon^d M)A_*$ with the corresponding eigenvalue λ_j, and $a_j := A_* v_j$ and $g_j := \langle AN, a_j \rangle$. Let $x(t) := ta_j + O(\epsilon^{-1})a_j^\perp$, where $a_j^\perp$ is a vector perpendicular to a_j. Then $|x(t)| = O(\epsilon^{-1})$, and hence, by (4.6), we get

$$H[g_j](x(t)) = \frac{\lambda_j}{\omega_d} \frac{|a_j|^2 t - a_j \cdot z}{|A_*(x(t) - z)|^d} \quad \text{modulo } O(\epsilon^{2d}). \tag{4.12}$$

Find the unique zero, call it t_j, of $H[g_j](x(t))$ as a function of t for $j = 1, \dots, d$. Let $\bar{z} = t_1 a_1 + \cdots + t_d a_d$. This $\bar{z}$ is the center. In fact, by the same argument as in [4], we can prove that

$$|\bar{z} - z| = O(\epsilon^d).$$

Detection of center—Method 2. Let b_j, $j = 1, \dots, d$, be the unit vector defined by (4.10). Then, from (4.8), we get

$$\omega_d |tA_* b_j|^{d-1} H[g](tb_j) = \langle \nabla U_g(z), \epsilon^d M e_j \rangle \quad \text{modulo } O(\epsilon^{d+1}). \tag{4.13}$$

Let $g = \frac{\partial U}{\partial \nu}$, where U is a second order homogeneous harmonic polynomial. By computing $\omega_d |tA_* b_j|^{d-1} H[g](tb_j)$ as $t \to \infty$, we recover $\langle \nabla U(z), \epsilon^d M e_j \rangle$, $j = 1, \dots, d$. From this we now recover $\nabla U_g(z)$, and hence the center z modulo $O(\epsilon^1)$.

The precision of this method is $O(\epsilon^1)$, which is worse than Method 1. However, numerical experiments in the next section show that this method performs better when there is noise in the data.

5. Computational experiments. This section presents results of numerical experiments of finding the inhomogeneity, $D \subset \Omega \subset \mathbb{R}^2$. In the following, Ω is assumed to be the disk centered at $(0, 0)$, with radius $r = 2$ and the background conductivity $A = I$. We also assume that $D = z + \epsilon B$, where B is the unit disk centered at $(0, 0)$. We note that in the anisotropic case, D being a disk does not provide a special advantage. Moreover, in the process of solving the inverse problem, we don't use any a priori knowledge of D being a disk.

Direct problem. Let u be the solution of (1.2). In order to collect the data $u|_{\partial\Omega}$, we solve the direct problem (1.2) as follows: In the same way as in [12], it can be proved that u is represented by

$$u(x) = \begin{cases} \mathcal{D}_\Omega u(x) - \mathcal{S}_\Omega g(x) + \mathcal{S}_D \phi(x) & \text{in } \Omega \setminus D, \\ \tilde{\mathcal{S}}_D \psi(x) & \text{in } D, \end{cases}$$

where $u|_{\partial\Omega}$, ϕ, and ψ satisfy the following relations:

$$\begin{aligned} & u = \mathcal{D}_\Omega u\big|_- - \mathcal{S}_\Omega g\big|_- + \mathcal{S}_D \phi \qquad \text{on } \partial\Omega, \\ & \mathcal{D}_\Omega u - \mathcal{S}_\Omega g + \mathcal{S}_D \phi\big|_+ = \tilde{\mathcal{S}}_D \varphi\big|_- \qquad \text{on } \partial D, \\ & \frac{\partial}{\partial \nu} \mathcal{D}_\Omega u - \frac{\partial}{\partial \nu} \mathcal{S}_\Omega g + \frac{\partial}{\partial \nu} \mathcal{S}_D \phi \Big|_+ = \frac{\partial}{\partial \tilde{\nu}} \tilde{\mathcal{S}}_D \varphi \Big|_- \qquad \text{on } \partial D. \end{aligned}$$

We solve this integral equation using the collocation method (see [14]) and obtain $u|_{\partial\Omega}$ on $\partial\Omega$ for given data g. We also add some noise to the computed data. Adding $p\%$ noise means that we have

$$u(1 + p \cdot \mathrm{rand}(1))$$

as the measured Dirichlet data. Here rand(1) is the random number in $(-1, 1)$.

Reconstruction Algorithm 1.

Step 1. Obtain Dirichlet data u on $\partial\Omega$ for given Neumann data $g_j = \langle N, e_j \rangle$, $j = 1, 2$.

Step 2. For $i, j = 1, 2$, calculate $\lim_{t\to\infty} \omega_2 t H[g_i](te_j)$ to obtain the matrix $\epsilon^2 M$.

Step 3. Find orthonormal eigenvectors v_1, v_2 and corresponding eigenvalues μ_1, μ_2 of $\epsilon^2 M$. Let μ be the minimum of μ_1, μ_2. The order of magnitude of D is $\epsilon = \sqrt{\mu |B|^{-1}}$.

Step 4. Let $g_j' = \langle N, v_j \rangle$ and $x_j(t) = tv_j + \frac{1}{\epsilon} v_j^{\perp}$, $j = 1, 2$. Find the zero, say t_j, of $H[g_j'](x_j(t)) = v_j \cdot e_1 H[g_1](x_j(t)) + v_j \cdot e_2 H[g_2](x_j(t))$ as a function of t. We obtain the center $\bar{z} = t_1 v_1 + t_2 v_2$.

Reconstruction Algorithm 2. Step 4 in the above algorithm is replaced with

Step 4′. For $g = \frac{\partial(x_1 x_2)}{\partial\nu}$, compute $h_j = \lim_{t\to\infty} \omega_2 t H[g](te_j)$, $j = 1, 2$. Then

$$(z_1, z_2) = (h_1, h_2) \begin{pmatrix} \epsilon^2 m_{12} & \epsilon^2 m_{22} \\ \epsilon^2 m_{11} & \epsilon^2 m_{12} \end{pmatrix}^{-1}.$$

We add the same amount of random noise in this step as well.

We now present the results of computational experiments. The first experiment is when $\tilde{A} - A$ is positive-definite; the second one is when $\tilde{A} - A$ is *not* positive-definite; the third one is to investigate the role of the condition number of $\tilde{A} - A$ in the reconstruction process.

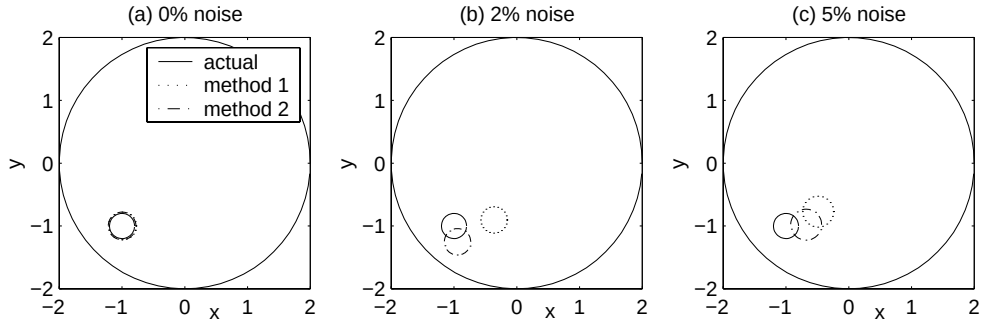


Fig. 1. *First example.*

Table 1

Numerical table for Figure 1. *Here* r *and* $\bar{r}$ *are the actual and computed radii, and* z, $\bar{z}_1$, *and* $\bar{z}_2$ *are the actual radius and those computed by Algorithms* 1 *and* 2, *respectively.*

z	r	noise(%)	$\bar{r}$	$\bar{z_1}$	$\lvert z - \bar{z_1}\rvert$
				$\bar{z_2}$	$\lvert z - \bar{z_2}\rvert$
$(-1, -1)$	0.2	0	0.2204	$(-0.9994, -0.9994)$	8.1154e-004
				$(-0.9994, -0.9994)$	8.4362e-004
		2	0.2101	$(-0.3681, -0.9038)$	0.6391
				$(-0.9445, -1.2519)$	0.2580
		5	0.2463	$(-0.4936, -0.7715)$	0.5556
				$(-0.6787, -0.9790)$	0.3220

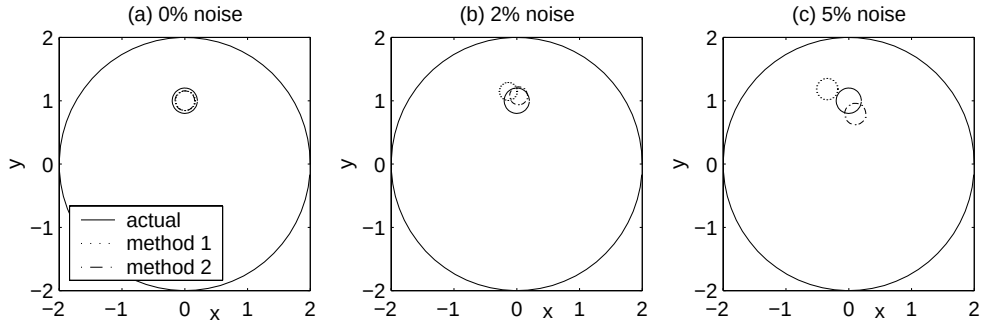


FIG. 2. *Second example.*

TABLE 2
Numerical table for Figure 2. *See explanations for Table* 1.

z	r	noise(%)	$\bar{r}$	$\bar{z_1}$ / $\bar{z_2}$	$\lvert z-\bar{z_1}\rvert$ / $\lvert z-\bar{z_2}\rvert$
$(0,1)$	0.2	0	0.1557	$(-0.0000, 0.9999)$	1.2316e-004
				$(-0.0000, 0.9998)$	1.6690e-004
		2	0.1417	$(-0.1421, 1.1468)$	0.2043
				$(0.0288, 1.0771)$	0.0823
		5	0.1689	$(-0.3499, 1.1841)$	0.3954
				$(0.1036, 0.7906)$	0.2336

Experiment 1. Let $\tilde{A} = \left(\begin{smallmatrix} 10 & 2 \\ 2 & 5 \end{smallmatrix}\right)$ and the actual inhomogeneity $D = (-1,-1) + 0.2B$. Note that $\tilde{A} - A$ is positive-definite. Figure 1 shows the results when there are 0%, 2%, and 5% random noise. Figure 2 is the result when $\tilde{A} = \left(\begin{smallmatrix} 10 & 1 \\ 1 & 2 \end{smallmatrix}\right)$ and $D = (0,1) + 0.2B$.

These results show that both Algorithms 1 and 2 detect the order of magnitude of the inclusion fairly well even with the presence of noise. However, Algorithm 2 performs better than Algorithm 1 in detecting the center when there is noise. A probable cause for this is that the zeros of functions in (4.12), which is already small, are sensitive to the noise.

Figure 3 shows that the location of the unknown inclusions does not affect the performance of the algorithms as long as they are away from $\partial\Omega$.

Experiment 2. This experiment is designed to determine whether the algorithms work in the case when $\tilde{A} - A$ is neither positive- nor negative-definite. Let $\tilde{A} = \left(\begin{smallmatrix} 2 & 0 \\ 2 & 1/2 \end{smallmatrix}\right)$ and $D = (1,0) + 0.2B$. Figure 4 shows the result. The algorithms seem to be working equally well for this case. It would be interesting to prove that the reconstruction formulas in this paper hold even when $\tilde{A} - A$ is neither positive- nor negative-definite. In this example as well, Algorithm 2 performs better in detection of the center.

Experiment 3. This experiment is designed to determine how the condition number of $\tilde{A} - A$ affects the precision of the algorithm. Suppose $A = I$. We first take $\tilde{A} = \left(\begin{smallmatrix} \lambda & 0 \\ 0 & 2 \end{smallmatrix}\right)$ and observe how the relative error $\frac{|z-\bar{z}|}{\epsilon^2}$ changes as λ increases. We then take $\tilde{A} = \left(\begin{smallmatrix} \lambda+1 & 0 \\ 0 & \lambda \end{smallmatrix}\right)$ and make the same observations. Figure 5 compares changes of relative errors of these two cases when $\lambda = 10, 10^2, 10^4, 10^5, 10^6$. It exhibits a clear difference: In the first case, when the condition number of $\tilde{A} - A$ increases as λ increases, the relative error is increasing, while in the second case, when the condition number does not change, the error is stabilized. The second case is somewhat similar to the isotropic case, and this kind of result is expected (see [11] or [2]). It is known

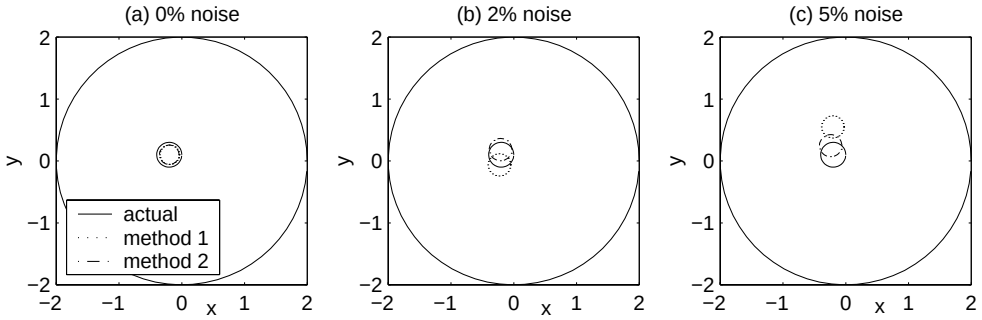


FIG. 3. *Third example.*

TABLE 3
Numerical table for Figure 3. *See explanation for Table* 1.

z	r	noise(%)	$\bar{r}$	$\bar{z_1}$	$\lvert z-\bar{z_1}\rvert$
				$\bar{z_2}$	$\lvert z-\bar{z_2}\rvert$
$(-0.2, 0.1)$	0.2	0	0.1559	$(-0.2000, 0.1000)$	1.2081e-005
				$(-0.2000, 0.1000)$	1.6354e-005
		2	0.1798	$(-0.2342, -0.0643)$	0.1678
				$(-0.2108, 0.1830)$	0.0837
		5	0.1785	$(-0.2120, 0.5493)$	0.4495
				$(-0.2405, 0.2464)$	0.1519

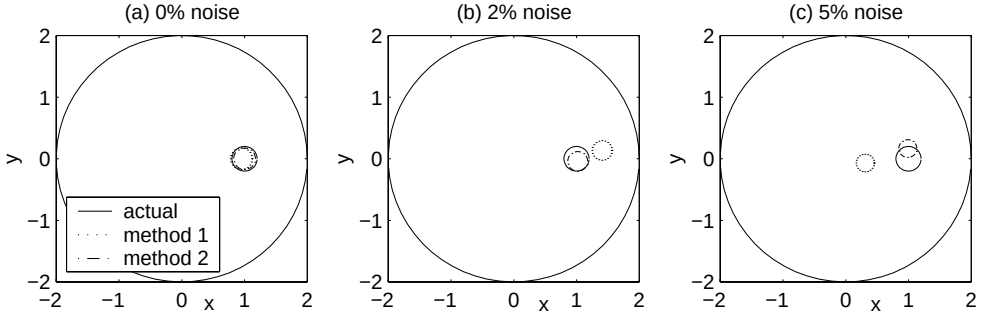


FIG. 4. *Fourth example.*

TABLE 4
Numerical table for Figure 4. *See explanation for Table* 1.

z	r	noise(%)	$\bar{r}$	$\bar{z_1}$	$\lvert z-\bar{z_1}\rvert$
				$\bar{z_2}$	$\lvert z-\bar{z_2}\rvert$
$(1, 0)$	0.2	0	0.1739	$(0.9447, -0.0000)$	0.0553
				$(1.0002, -0.0000)$	1.8443e-004
		2	0.1586	$(1.4031, 0.1347)$	0.4250
				$(1.0186, -0.0427)$	0.0465
		5	0.1436	$(0.3048, -0.0702)$	0.6987
				$(0.9891, 0.1656)$	0.1660

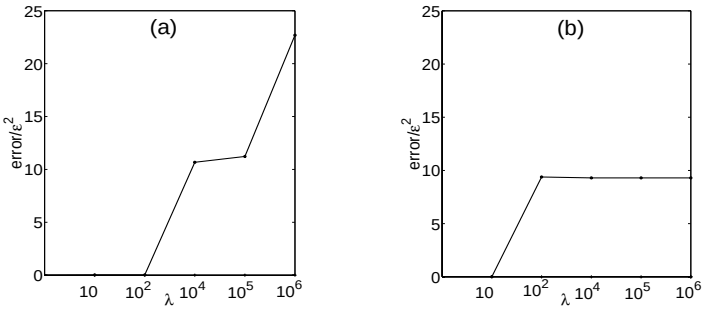


FIG. 5. *The graphs show how the quantities* $\frac{|z-\bar{z}|}{\epsilon^2}$ *change as* λ *goes to* ∞. (a) *The condition number of* $\tilde{A}-A$ *is* λ, *and the relative error increases.* (b) *The condition number of* $\tilde{A}-A$ *is* 1, *and the relative error does not increase.*

that long and thin inclusions, or cracklike inclusions (inclusions of high Lipschitz character) are hard to detect; see, for example, [4]. This experiment suggests that in addition to these geometric obstructions, in the anisotropic case there is another obstruction of high condition number of $\tilde{A}-A$.

Conclusion. Numerical results show that the second reconstruction algorithm performs better in the presence of noise. They also show that the reconstruction procedure works well even when $A-\tilde{A}$ is neither positive- nor negative-definite, and that the error of reconstruction increases as the condition number of $A-\tilde{A}$ increases. It would be interesting to investigate these points in a mathematically rigorous way.

REFERENCES

[1] H. Ammari and H. Kang, *High-order terms in the asymptotic expansions of the steady-state voltage potentials in the presence of conductivity inhomogeneities of small diameter*, SIAM J. Math. Anal., to appear.

[2] H. Ammari and H. Kang, *Properties of generalized polarization tensors*, Multiscale Model. Simul., to appear.

[3] H. Ammari, S. Moskow, and M. Vogelius, *Boundary integral formulas for the reconstruction of electromagnetic imperfections of small diameter*, ESAIM: Control Optim. Calc. Var., to appear.

[4] H. Ammari and J. K. Seo, *A New Formula for the Reconstruction of Conductivity Inhomogeneities*, Adv. Appl. Math., to appear.

[5] M. Brühl, M. Hanke, and M. Vogelius, *A Direct Impedance Tomography Algorithm for Locating Small Inhomogeneities*, Numer. Math., to appear.

[6] D. J. Cedio-Fengya, S. Moskow, and M. Vogelius, *Identification of conductivity imperfections of small diameter by boundary measurements: Continuous dependence and computational reconstruction*, Inverse Problems, 14 (1998), pp. 553–595.

[7] L. Escauriaza and J. K. Seo, *Regularity properties of solutions to transmission problems*, Trans. Amer. Math. Soc., 338 (1993), pp. 405–430.

[8] A. Friedman and M. Vogelius, *Identification of small inhomogeneities of extreme conductivity by boundary measurements: A theorem on continuous dependence*, Arch. Ration. Mech. Anal., 105 (1989), pp. 299–326.

[9] F. Hettlich and W. Rundell, *The determination of a discontinuity in a conductivity from a single boundary measurement*, Inverse Problems, 14 (1998), pp. 67–82.

[10] M. Ikehata and S. Siltanen, *Numerical method for finding the convex hull of an inclusion in conductivity from boundary measurements*, Inverse Problems, 16 (2000), pp. 1043–1052.

[11] H. Kang and J. K. Seo, *Identification of domains with near-extreme conductivity: Global stability and error estimates*, Inverse Problems, 15 (1999), pp. 851–867.

[12] H. Kang and J. K. Seo, *Recent progress in the inverse conductivity problem with single measurement*, in Inverse Problems and Related Fields, CRC Press, Boca Raton, FL, 2000, pp. 69–80.

[13] R. E. Kleinman and T. B. A. Senior, *Rayleigh scattering*, in Low and High Frequency Asymptotics, V. K. Varadan and V. V. Varadan, eds., North–Holland, Amsterdam, 1986, pp. 1–70.
[14] R. Kress, *Linear Integral Equations*, Appl. Math. Sci. 82, Springer, New York, 1989.
[15] O. Kwon, J. K. Seo, and J. R. Yoon, *A real-time algorithm for the location search of discontinuous conductivities with one measurement*, Comm. Pure Appl. Math., 55 (2002), pp. 1–29.
[16] A. B. Movchan and S. K. Serkov, *The Pólya–Szegö matrices in asymptotic models of dilute composite*, European J. Appl. Math., 8 (1997), pp. 595–621.
[17] G. Pólya and G. Szegö, *Isoperimetric Inequalities in Mathematical Physics*, Ann. of Math. Stud. 27, Princeton University Press, Princeton, NJ, 1951.
[18] M. Schiffer and G. Szegö, *Virtual mass and polarization*, Trans. Amer. Math. Soc., 67 (1949), pp. 130–205.
[19] C. Tomalsky and A. Wiegmann, *Recovery of small perturbations of an interface for an elliptic inverse problem via linearization*, Inverse Problems, 15 (1999), pp. 401–411.

SIAM J. APPL. MATH.
Vol. 63, No. 4, pp. 1292–1312

THE NO RESPONSE TEST—A SAMPLING METHOD FOR INVERSE SCATTERING PROBLEMS*

D. RUSSELL LUKE† AND ROLAND POTTHAST‡

Abstract. We describe a novel technique, which we call the no response test, to locate the support of a scatterer from knowledge of a far field pattern of a scattered acoustic wave. The method uses a set of sampling surfaces and a special test response to detect the support of a scatterer without a priori knowledge of the physical properties of the scatterer. Specifically, the method does not depend on information about whether the scatterer is penetrable or impenetrable nor does it depend on any knowledge of the nature of the scatterer (absorbing, reflecting, etc.). In contrast to previous sampling algorithms, the techniques described here enable one to locate obstacles or inhomogeneities from the far field pattern of only one incident field—the no response test is a one-wave method. We investigate the theoretical basis for the no response test and derive a one-wave uniqueness proof for a region containing the scatterer. We show how to find the object within this region. We demonstrate the applicability of the method by reconstructing sound-soft, sound-hard, impedance, and inhomogeneous medium scatterers in two dimensions from one wave with full and limited aperture far-field data.

Key words. inverse problems, scattering theory, image processing

AMS subject classifications. 35R30, 35P25, 68U10, 94A08

PII. S0036139902406887

1. Introduction.

Inverse scattering is concerned with recovering information about a medium and its embedded objects by exciting or illuminating the medium with acoustic or electromagnetic fields and measuring the resulting field. One of the fundamental problems is to determine the location and shape of scatterers that are either buried or located in some inaccessible region of a medium. Applications range from geoscience to medical imaging.

Over the past decade, several innovative and successful methods have been introduced into the area of inverse scattering. Here, we add to the list a methodology that we call the *no response test* and demonstrate its applicability. The idea of the method is to test the hypothesis that a scatterer lies within a given test domain given the far field data. We sample by construction the set of incident fields that are small on the test domain and large outside. The far field patterns corresponding to these incident fields are then calculated using the given far field data. We call the calculated far field patterns *responses*. If all the responses are small, then the unknown scatterer is shown to be a subset of the test domain, that is, the hypothesis is true. The unknown scatterer is located within the union of all test domains for which the hypothesis is true. Since small, rather than large, responses indicate the location of the scatterer, the methodology is called the "no response" method.

To place this methodology relative to other reconstruction techniques, we give a brief review of the different reconstruction approaches. The evolutionary tree of inverse scattering algorithms is diverse enough that some taxonomy is in order. We

*Received by the editors May 1, 2002; accepted for publication (in revised form) September 18, 2002; published electronically April 23, 2003.
http://www.siam.org/journals/siap/63-4/40688.html

†Simon Fraser University, Burnaby, British Columbia V5A 1S6, Canada (luke@cecm.sfu.ca). This author is a Post-Doctoral Fellow of the Pacific Institute for the Mathematical Sciences. His work was supported while at the University of Göttingen.

‡Institute for Numerical and Applied Mathematics, University of Göttingen, 37083 Göttingen, Germany (potthast@scienceatlas.de, http://www.scienceatlas.de/nfg).

separate reconstruction algorithms into three classes: *iterative*, *decomposition*, and *sampling/probe* methods.

Category	Methods
I	Iterative techniques
	Newton method Landweber scheme Least squares fits (depending on the setup) Conjugate gradient method
II	Decomposition techniques
	Colton–Monk method / dual space method Kirsch–Kress method Potthast / point source method
III	Probe and sampling techniques
	Colton–Kirsch method / linear sampling method Kirsch / factorization method Potthast / singular sources method Ikehata / probe method Ikehata / enclosure method Luke–Potthast / no response test

Iterative methods (category I in [16]) use the model of the full forward problem, or an appropriate approximation thereof, for the solution of the inverse problem. These techniques have the advantage that they use all information about the forward problem for the solution of the inverse problem, and they usually deliver quite good reconstructions. However, due to the need to solve the forward problem many times, they can be computationally intensive. Also, obtaining a localized reconstruction in a limited data setting is problematic since full data for solving the forward problem is presumed. Indeed, at the very least it is presumed that one knows *which* model the data should satisfy. Well-known examples of iterative techniques are the *Newton method*, the *Landweber method*, and various versions of *least squares fits*.

Decomposition algorithms (category II in [16]) consist of methods that split the inverse problem into an ill-posed part to reconstruct the scattered field and a well-posed part to find the unknown scatterer due to some boundary condition. Representatives of this type of method are the *dual space method* proposed by Colton and Monk [3, 4] and the technique of Kirsch and Kress [2], but also newer strategies like the *point source method* of Potthast [14, 15, 16], which turns out to be a type of adjoint method to the Kirsch–Kress technique.

Sampling and probe methods comprise the third and most recent class of algorithms (category III in [16]). These involve testing a given region with a model mapping the data to a point in the test region and locating the boundary of the unknown scatterer as the points where some unusual or characteristic behavior (usually some resolvable type of blow-up) occurs in the model functions. Where these techniques differ is in the construction of the model functions, which leads to fundamentally different algorithms. These methods share the advantage that they can be applied without knowing whether the scatterer is an impenetrable (sound-soft or sound-hard) or an inhomogeneous medium. The no response test belongs to this class. We discuss in more detail the different strategies within this class that have been proposed since 1995.

A scatterer is denoted by its support $\Omega \subset \mathbb{R}^m$ ($m = 2$ or 3) where Ω is bounded. For our purposes we need only assume that the boundary of the scatterer $\partial\Omega$ is Lipschitz; however, this introduces mathematical technicalities that cloud the central

ideas here. We therefore limit our discussion to twice continuously differentiable (C^2) boundaries. Readers interested in the details of boundaries with corners, or more generally Lipschitz boundaries, are referred to [11, 12]. We denote by $\nu(x_0)$ the *unit outward normal* to Ω at the point $x_0 \in \partial\Omega$, that is, $|\nu(x_0)| = 1$ and the vector product $(y - x_0, \nu(x_0)) \leq 0$ for every $y \in \overline{\Omega}$.

Let $u, u^s : \mathbb{R}^m \to \mathbb{C}$ and $u^\infty : \mathbb{S} \to \mathbb{C}$ denote the *total*, *scattered*, and *far* fields, respectively, due to excitation from an *incident plane wave* u^i at a fixed wavenumber $\kappa > 0$. Here $\mathbb{S} := \{x \in \mathbb{R}^m \mid |x| = 1\}$. We parameterize these fields by the *direction of incidence* $\hat{y}$ of the incident plane wave $u^i(x, \hat{y}) := e^{i\kappa x \cdot \hat{y}}$, $x \in \mathbb{R}^m$, $\hat{y} \in \mathbb{S}$, where $i = \sqrt{-1}$ in the exponential. Similarly we write the dependence on the direction of incidence explicitly in the argument of the other fields as $u(x, \hat{y})$, $u^s(x, \hat{y})$, and $u^\infty(\hat{x}, \hat{y})$, respectively. Here and elsewhere, the hat indicates a unit vector, $\hat{x} := \frac{x}{|x|}$. When the far field data is known only on an open subset Γ of $\mathbb{S}$, we call the data *limited aperture* data.

Method	Amount of far field data needed	Assumptions on the physical nature of the scatterer
Colton–Kirsch / linear sampling method	$u^\infty(\hat{x}, \hat{y})$ $\forall \hat{x}, \hat{y} \in \Gamma \subset \mathbb{S}$	none
Kirsch / factorization method	$u^\infty(\hat{x}, \hat{y})$ $\forall \hat{x}, \hat{y} \in \mathbb{S}$	none
Potthast / singular sources method	$u^\infty(\hat{x}, \hat{y})$ $\forall \hat{x}, \hat{y} \in \Gamma \subset \mathbb{S}$	none
Ikehata / probe method	$u^\infty(\hat{x}, \hat{y})$ $\forall \hat{x}, \hat{y} \in \mathbb{S}$	none
Ikehata / enclosure method	$u^\infty(\hat{x}, \hat{y})$ $\forall \hat{x} \in \mathbb{S}$ one $\hat{y} \in \mathbb{S}$	none
Luke–Potthast / no response test	$u^\infty(\hat{x}, \hat{y})$ $\forall \hat{x} \in \Gamma \subset \mathbb{S}$ one $\hat{y} \in \mathbb{S}$	none

The *linear sampling method* of Colton and Kirsch [1] characterizes the domain of an unknown scatterer by the behavior of the solution to the integral equation of the first kind:

$$\int_{\mathbb{S}} u^\infty(\hat{x}, \hat{y}) g(\hat{y}) ds(\hat{y}) = e^{i\kappa \hat{x} \cdot z}, \qquad \hat{x} \in \mathbb{S}. \tag{1.1}$$

Here a regularized solution g is calculated for all points z on a sampling grid $\mathcal{G}$. The unknown boundary is found where $||g(z)||$ becomes unbounded.

Kirsch [10] proposed a modified version of this method by constructing a spectral decomposition of the operator

$$(Fg)(\hat{x}) := \int_{\mathbb{S}} u^\infty(\hat{x}, \hat{y}) g(\hat{y}) ds(\hat{y}), \qquad \hat{x} \in \mathbb{S},$$

used in (1.1). He proposed to solve the equation

$$(F^* F)^{1/4} g(\hat{x}) = e^{i\kappa \hat{x} \cdot z}, \qquad \hat{x} \in \mathbb{S},$$

for all $z \in \mathcal{G}$ and showed that the equation is solvable if and only if z is in the interior of the unknown scatterer. This technique of Kirsch is known as the *modified linear sampling* or *factorization method*.

Ikehata and Potthast have independently proposed two related algorithms, the *probe method* [6] and the *method of singular sources* [16], respectively. These techniques are distinct from the (modified) linear sampling methods above in that they use different quantities that blow up when approaching the boundary of some scatterer.

The probe method of Ikehata uses Green's formula to define an indicator function that blows up when the virtual source touches the unknown obstacle. Let Λ be the Dirichlet-to-Neumann map for the boundary value problem in a domain B with the unknown domain $\Omega \subset B$ and Λ_0 be the Dirichlet-to-Neumann map for B without the existence of Ω. Ikehata proposed considering

$$I(z,f) := \int_{\partial B} \overline{(\Lambda - \Lambda_0)f} \cdot f ds$$

for specially constructed functions f. It can be shown that $I(z,f)$ tends to infinity if z tends to the boundary of the unknown domain. The Dirichlet-to-Neumann map can be calculated from the far field patterns $u^\infty(\hat{x},\hat{y})$ for all $\hat{x},\hat{y} \in \mathbb{S}$, i.e., from the far field pattern for scattering of all plane waves of one fixed frequency.

The singular sources method of Potthast uses a different functional which also blows up at the boundary of the obstacle. This functional is defined as the magnitude of the *scattered field* $\Psi^s(z,z)$ of singular sources $\Psi(\cdot,z)$ and is calculated by backprojection of the form

$$\Psi^s(y,z) \approx \int_{\mathbb{S}}\int_{\mathbb{S}} u^\infty(\hat{x},\hat{y}) g(\hat{x},y) g(-\hat{y},z) ds(\hat{y}) ds(\hat{x}), \qquad y,z \in \mathbb{R}^m \setminus \Omega,$$

for explicitly constructed kernels $g(\cdot,\cdot)$.

All of the linear sampling and probe methods share the advantage that no knowledge about the boundary condition of the unknown scatterer is needed. With the exception of the Kirsch factorization method, these methods are valid in the limited aperture case, where the far field data is not known on the full sphere but only on an open subset $\Gamma \subset \mathbb{S}$. The principle disadvantage of sampling and probe techniques, however, is that they all require the knowledge of far field patterns for a large number of incident plane waves. The current challenge facing these algorithms is to reduce the amount of data needed for reliable reconstructions.

Recent work by Ikehata has made significant progress toward the development of reconstruction algorithms using very limited data. His *enclosure method* [7, 8] enables one to find the support of convex polygons from the knowledge of one measured field. Ikehata uses a special harmonic incident field,

$$v = e^{\tau x \cdot (\omega + i\omega^\perp)},$$

to construct the following indicator function:

$$I_\omega(\tau,t) = e^{-\tau t}\left\{ \left\langle \left.\frac{\partial u}{\partial \nu}\right|_{\partial G}, v|_{\partial G} \right\rangle - \left\langle \left.\frac{\partial v}{\partial \nu}\right|_{\partial G}, u|_{\partial G} \right\rangle \right\}, \qquad \tau > 0, \quad t \in \mathbb{R}, \tag{1.2}$$

where $\omega \in \mathbb{S}$ is a direction vector, u is the unknown, weak solution to the scattering problem, and G is some domain containing the unknown scatterer, $\Omega \subset \operatorname{int} G$ the interior of G. Ikehata shows that at the corners of polygonal scatterers this indicator function becomes unbounded. He then exploits this property to uniquely reconstruct the scatterer. For details on implementation, see [9]. While for the purposes of analysis the presentation of the enclosure method is limited to specific settings, it appears that in practice the method is independent of the material properties of the scatterer.

In this work we propose another technique for locating a scatterer from a single incident wave that also exploits the behavior of a special indicator function in the

neighborhood of a scatterer. Since we look, rather, for where the indicator function does *not* become unbounded, we call the method the no response test. Like the enclosure method, the no response test can be used to locate scatterers from only one incident wave. Moreover, neither the enclosure method nor the no response test require a priori knowledge of the material properties of the scatterer. However, the indicator function in the no response test is a different functional on the measured data than that of Ikehata. Also, we do not make use of, nor place any particular constraints on, the geometric properties of the scatterer.

It is often the case that numerical algorithms precede by many years their mathematical justification. The absence of analytical results for a particular application does not preclude the successful implementation and numerical study of algorithms. At the expense of mathematical analysis limited to narrow settings, we have chosen to highlight the robustness of the no response test in a variety of settings by focusing on numerical results. We leave many questions unanswered; however, the demonstration of the applicability of the techniques discussed here helps to motivate and formulate the analysis that must follow. In section 3 we provide preliminary theoretical results to motivate the method. A convergence proof for one-wave reconstructions would include a one-wave uniqueness result. These results are not yet available. However, we can show that a set (depending on some test domain Ω_t^0) surrounding the unknown scatterer, which we call its *corona*, is uniquely determined by the one-wave far field pattern independent of the boundary condition.

The no response algorithm is given in section 4. In the same section we show reconstructions for scattering from scatterers with *Dirichlet*, *Neumann*, or *impedance* boundary conditions or for scattering from an inhomogeneous medium. We show results from each of these scatterers with full and limited aperture data. Preparatory to this, we briefly review in section 2 the fundamental scattering models for sound-soft, sound-hard, and mixed obstacles as well as inhomogeneous media.

2. Dirichlet, Neumann, impedance, and medium scattering problems. This section serves to review briefly the key elements of scattering by bounded objects or media and to provide some tools for the inversion method described in section 3. We also describe how we solved the forward problems to produce the data used for the demonstration of the no response test.

Scattering review. Let v^i be an incident field that satisfies the Helmholtz equation,

$$\triangle v + \kappa^2 v = 0,$$

with wave number $\kappa > 0$ on $\mathbb{R}^m$. The incident field produces a scattered field v^s that solves the Helmholtz equation on the exterior of the scatterer Ω and satisfies the *Sommerfeld radiation condition*

$$r^{\frac{m-1}{2}}\Big(\frac{\partial}{\partial r} - i\kappa\Big)v(x) \to 0, \qquad r = |x| \to \infty$$

uniformly in all directions. For impenetrable scatterers we consider cases where the scatterer is either sound-soft (a perfect conductor), sound-hard (a perfect reflector), or some mixture of these. Each of these types of scatterers is modeled by a total field,

$$v = v^i + v^s,$$

that satisfies either *Dirichlet*, *Neumann*, or *impedance* boundary conditions. These boundary conditions are given, respectively, as

$$v|_{\partial\Omega} = 0, \qquad \frac{\partial v}{\partial \nu}|_{\partial\Omega} = 0, \qquad \frac{\partial v}{\partial \nu}|_{\partial\Omega} + \lambda v|_{\partial\Omega} \ = \ 0,$$

with the impedance function $\lambda \in C(\partial\Omega)$. We also treat penetrable scatterers, where the inhomogeneity is modeled by a nonnegative refractive index $n : \mathbb{R}^m \to \mathbb{R}_+$ and where $n(x) := 1$ for $x \in \mathbb{R}^m \setminus \Omega$. Then the total field $v \in H^2_{loc}(\mathbb{R}^m)$ solves the inhomogeneous Helmholtz equation,

$$\triangle v + \kappa^2 n v = 0,$$

in $\mathbb{R}^m$, and $v^s = v - v^i$ satisfies the Sommerfeld radiation condition.

The following result enables us to calculate the scattered and far fields of any reasonable incident field as the weighted superposition of the corresponding fields generated by scattering from incident plane waves. This result is fundamental to the no response test.

THEOREM 2.1. *Let Γ be an open subset of $\mathbb{S}$, the unit sphere on $\mathbb{R}^m$ ($m = 2, 3$), and let $\Omega \subset \mathbb{R}^m$ denote the bounded support of a scattering body with C^2 boundary. Denote by $u^s : \mathbb{R}^m \to \mathbb{C}$ and $u^\infty : \mathbb{S} \to \mathbb{C}$ the scattered and far fields, respectively, due to excitation from an incident plane wave u^i at a fixed wavenumber $\kappa > 0$ with direction $-\hat{y}$, $u^i(x, -\hat{y}) := e^{i\kappa x\cdot(-\hat{y})}$, $x \in \mathbb{R}^m$, $\hat{y} \in \mathbb{S}$. Consider the superposition of plane waves*

$$v^i(x) \ = \ \int_\Gamma e^{i\kappa x\cdot(-\hat{y})} g(-\hat{y}) ds(\hat{y}), \qquad x \in \mathbb{R}^m, \tag{2.1}$$

where $g \in L^2(-\Gamma)$. The corresponding solution to the scattering problem with Dirichlet, Neumann, or impedance boundary conditions or scattering by an inhomogeneous medium is given by $v = v^i + v^s$, where

$$v^s(x) \ = \ \int_\Gamma u^s(x, -\hat{y}) g(-\hat{y}) ds(\hat{y}), \qquad x \in \mathbb{R}^m \setminus \overline{\Omega}.$$

The corresponding far field pattern is given by

$$v^\infty(\hat{x}) \ = \ \int_\Gamma u^\infty(\hat{x}, -\hat{y}) g(-\hat{y}) ds(\hat{y}), \qquad \hat{x} \in \mathbb{S}. \tag{2.2}$$

Proof. The proof relies only on the linearity and boundedness of the particular scattering problem. Linearity implies that the sum of two incident fields is scattered onto the sum of the single scattered fields. By boundedness of the scattering operator from $C(\Omega)$ into $C_{loc}(\mathbb{R}^m \setminus \overline{\Omega})$, the limit for the integration can be performed and we obtain the stated results. ∎

The signs in the expressions for v^i, v^s, and v^∞ above have been chosen so that the backprojection mapping between the far field measurements and the scattered field, which we derive below, has a natural interpretation in terms of a physical aperture in the far field. Note that the function g is defined on $-\Gamma$, where $-\Gamma$ is the mirror image of the interval Γ: $\hat{y} \in \Gamma \iff -\hat{y} \in -\Gamma$. Using the standard far field reciprocity relation $u^\infty(\hat{x}, -\hat{y}) = u^\infty(\hat{y}, -\hat{x})$ $(\hat{x}, \hat{y} \in \mathbb{S})$ we see that the far field is defined on Γ with any incident wave direction $-\hat{x}$. When $\Gamma = \mathbb{S}$ this virtual aperture is not as apparent. The incident field v^i given by (2.1) is called a *Herglotz wave function*. Since this function depends on the density g, we write this explicitly as $v^i[g](x)$. We denote the scattered field for scattering of a Herglotz wave function $v^i[g](x)$ by $v^s[g](x)$. Similarly, the corresponding far field pattern is given by $v^\infty[g](\hat{x})$.

Numerical considerations. As a basis both for the theoretical discussion and the implementations (that is, the generation of the simulated data), we briefly sketch the solution of the above scattering problems. For all proofs and a detailed discussion we refer to [2] and [16].

For the solution of the Dirichlet problem we represent the scattered field as a combined single- and double-layer potential

$$v^s(x) = \int_{\partial\Omega} \left\{ \frac{\partial\Phi(x,y)}{\partial\nu(y)} - i\Phi(x,y) \right\} \varphi(y) ds(y), \qquad x \in \mathbb{R}^m \setminus \partial\Omega.$$

For this representation of the scattered field and the boundary condition, the density φ must satisfy the integral equation

$$\varphi + K\varphi - iS\varphi = -2v^i, \tag{2.3}$$

where $\mathcal{S}$ is the single-layer operator,

$$(S\varphi)(x) := 2\int_{\partial\Omega} \Phi(x,y)\varphi(y) ds(y), \qquad x \in \partial\Omega,$$

and K is the double-layer operator,

$$(K\varphi)(x) := 2\int_{\partial\Omega} \frac{\partial\Phi(x,y)}{\partial\nu(y)}\varphi(y) ds(y), \qquad x \in \partial\Omega.$$

The equation has a unique solution that depends continuously on the right-hand side in $C(\partial\Omega)$.

For the Neumann problem we use the modified approach due to Panich [13]:

$$v^s(x) = \int_{\partial\Omega} \left\{ \Phi(x,y)\varphi(y) + i\frac{\partial\Phi(x,y)}{\partial\nu(y)}(S_0^2\varphi)(y) \right\} ds(y), \qquad x \in \mathbb{R}^m \setminus \partial\Omega, \tag{2.4}$$

where S_0 denotes the single-layer operator in the case $\kappa = 0$. For this representation of the scattered field, the density φ can be shown to satisfy the boundary integral equation

$$\varphi - K'\varphi - iTS_0^2\varphi = 2\frac{\partial v^i}{\partial\nu}, \tag{2.5}$$

where

$$(K'\varphi)(x) := 2\int_{\partial\Omega} \frac{\partial\Phi(x,y)}{\partial\nu(x)}\varphi(y) ds(y), \qquad x \in \partial\Omega,$$

and

$$(T\varphi)(x) := 2\frac{\partial}{\partial\nu(x)}\int_{\partial\Omega} \frac{\partial\Phi(x,y)}{\partial\nu(y)}\varphi(y) ds(y), \qquad x \in \partial\Omega.$$

Both (2.3) and (2.5) have unique solutions that depend continuously on the incident field in $C(\partial\Omega)$.

For the impedance boundary value problem we follow the same approach using the representation (2.4). An application of the jump relations leads to the equation

$$\left[I - K' - iTS_0^2 - \lambda S - i\lambda(I+K)S_0^2 \right]\varphi = 2\frac{\partial v^i}{\partial\nu} + 2\lambda v^i. \tag{2.6}$$

Under suitable assumptions on the impedance λ (basically ensuring uniqueness of the impedance scattering problem) the integral equation (2.6) has a unique solution which depends continuously on the incident field in $C(\partial\Omega)$.

For the penetrable inhomogeneous medium we use Green's formula applied to the total field to recast the solution to the scattering problem as the solution to the *Lippmann–Schwinger* equation

$$v(x) \;=\; v^i - \kappa^2 \int_{\mathbb{R}^m} \Phi(x,y) m(y) v(y), \qquad x \in \mathbb{R}^m,$$

where $m(y) := 1 - n(y)$ for the index of refraction $n : \mathbb{R}^m \to \mathbb{R}_+$. The Lippmann–Schwinger equation has a unique solution in $C(\Omega)$ that depends continuously on the incident field v^i.

3. The inverse problem and the no response test. The *inverse problem* we consider is to locate the scatterer Ω given an incident plane wave u^i and the far field data restricted to the aperture $u^\infty|_\Gamma$, where $\Gamma \subset \mathbb{S}$ is some open set. The solution to the inverse problem is often called the *reconstruction* of the scatterer. The no response method is a reconstruction algorithm that uses only one incident wave and does not use any a priori information about the physical characteristics of the scatterer.

We consider the hypothesis that a scatterer lies within a given domain. The no response test is a way to determine whether or not this hypothesis is true. We begin with a heuristic description of the reconstruction method based on this test. An explicit formulation of the full algorithm is given in the next section.

The scattering test response. Let $\Omega_t \subset \mathbb{R}^m$ be a bounded test domain with a C^2 boundary. Sample by construction (see (2.1)) the set of incident fields that are small on the test domain Ω_t and large outside. The far field patterns corresponding to these incident fields are then calculated via (2.2). We call the magnitude of the calculated far field patterns *responses*. If the maximum of the sampled responses is small, we show that this is an indication that the unknown scatterer is a subset of the test domain Ω_t. While general geometric properties of the test domain are not important (e.g., convexity, symmetry, and so forth), it is critical that the test domain be large enough that by translation the scatterer is contained in the interior. The no response algorithm makes use of a template test domain Ω_t^0 that is rotated and translated around the computational domain. The location and shape of the scatterer is then recovered by the behavior, with respect to these test domains, of the sampled *scattering test response*, that is, the supremum over all responses for a fixed test domain. This is defined below.

DEFINITION 3.1 (scattering test response). *Given the far field pattern u^∞ due to an incident plane wave u^i with direction $-\hat{x}$ and a scatterer Ω as in Theorem* 2.1*, let $v^i[g]$ denote a Herglotz wave function defined by* (2.1) *and $v^\infty[g]$ denote the corresponding far field pattern given by* (2.2)*. We define the* scattering test response *for the test domain Ω_t by*

$$\mu_\epsilon(\Omega_t, \Omega, \hat{x}) := \sup\left\{ |v^\infty[g](\hat{x})| : \;\; g \in L^2(-\Gamma) \text{ such that } ||v^i[g]||_{C(\Omega_t)} \le \epsilon \right\}. \tag{3.1}$$

We keep the direction $\hat{x}$ fixed in what follows, so to reduce notational clutter we drop the argument and use the notation $\mu_\epsilon(\Omega_t, \Omega)$ whenever there is no chance for confusion.

To calculate μ_ϵ from the far field pattern $u^\infty|_\Gamma$ for scattering of a plane wave u^i with direction $-\hat{x}$, we use the reciprocity relation $u^\infty(\hat{x},-\hat{y}) = u^\infty(\hat{y},-\hat{x})$ $(\hat{x},\hat{y}\in\mathbb{S})$ and Theorem 2.1 to obtain

$$\begin{aligned} v^\infty(\hat{x}) &= \int_\Gamma u^\infty(\hat{x},-\hat{y})g(-\hat{y})ds(\hat{y}) \\ &= \int_\Gamma u^\infty(\hat{y},-\hat{x})g(-\hat{y})ds(\hat{y}). \end{aligned} \tag{3.2}$$

Thus, from knowledge of the far field pattern $u^\infty(\hat{y},-\hat{x})$, $\hat{y}\in\Gamma$, for one wave with direction of incidence $-\hat{x}$, we can reconstruct $\mu_\epsilon(\Omega_t,\Omega)$ for any domain Ω_t by construction of appropriate kernels g of the (limited aperture) Herglotz wave functions. Before discussing in detail the construction of the densities g and the test domains Ω_t, we prove some basic results about the behavior of the scattering test response that motivate our numerical methods.

The no response test is built upon two observations. First, when the scatterer Ω is contained in the *interior* of the test domain Ω_t, the value $\mu_\epsilon(\Omega_t,\Omega)$ is *small or bounded*. Second, if the scatterer is in the *exterior* of the test domain, then $\mu_\epsilon(\Omega_t,\Omega)$ is *large or unbounded*. These facts are used to locate the support Ω of the scatterer as a region contained in the union of test domains where the scattering test response μ_ϵ is bounded. We summarize this critical behavior in the following theorem.

THEOREM 3.2 (behavior of the scattering test response). *If $\Omega\subset\Omega_t$, then there is a constant $c\in\mathbb{R}$ such that*

$$\mu_\epsilon(\Omega_t,\Omega)\le c\epsilon.$$

On the other hand, if $\overline{\Omega}\cap\overline{\Omega}_t=\emptyset$, and $\mathbb{R}^m\setminus(\overline{\Omega}\cup\overline{\Omega}_t)$ is connected, then we have

$$\mu_\epsilon(\Omega_t,\Omega)=\infty.$$

Proof. When $\Omega\subset\Omega_t$ the boundedness of the scattering map $v^i\mapsto v^\infty$ implies the existence of a constant c such that for all v^i satisfying

$$||v^i||_{C(\Omega_t)}\le\epsilon,$$

we have

$$||v^\infty||_{C(\mathbb{S})}\le c\epsilon.$$

This completes the proof of the first statement.

To prove the second statement, we consider two disjoint domains, Ω_t' and Ω', satisfying $\Omega_t\subset\Omega_t'$, $\Omega\subset\Omega'$, and $\overline{\Omega_t'}\cap\overline{\Omega'}=\emptyset$. We further require that the interior homogeneous Dirichlet problems for Ω_t' and Ω' have only the trivial solution. Then the Herglotz wave operator $H: L^2(-\Gamma)\to L^2(\partial(\Omega_t'\cup\Omega'))$, defined by

$$(Hg)(x) := v^i[g](x)\Big|_{\partial(\Omega_t'\cup\Omega')},$$

has dense range. This can be shown in a similar fashion to the proof of Lemma 3.1.2 of [16]. Choose $y\notin\overline{\Omega_t'}\cup\overline{\Omega'}$ such that the far field pattern $w^\infty(\hat{x},y)$ for scattering of $\Phi(\cdot,y)$ by Ω is not zero. This is always possible since, by the mixed reciprocity relation [16, Theorem 2.1.4], we have

$$w^\infty(\hat{x},y)=\gamma u^s(y,-\hat{x})$$

and $u^s(\cdot,-\hat{x})$ cannot vanish on an open subset of $\mathbb{R}^m$. Next, construct $v^i[g](x)$ satisfying

$$||v^i[g](x)||_{C(\Omega_t')} \leq \epsilon, \qquad ||v^i[g](x) - \beta\Phi(\cdot,y)||_{C(\Omega')} \leq \epsilon.$$

Then since $\Omega_t \subset \Omega_t'$, we have

$$\mu_\epsilon(\Omega_t,\Omega) \geq |v^\infty[g](\hat{x})|.$$

By definition $\Omega \subset \Omega'$; thus

$$\left|v^\infty[g](\hat{x}) - \beta w^\infty(\hat{x},y)\right| \leq c\epsilon,$$

with some constant c, which, by the triangle inequality, yields

$$|v^\infty[g](\hat{x})| \geq \beta|w^\infty(\hat{x},y)| - c\epsilon.$$

Thus we have

$$\mu_\epsilon(\Omega_t,\Omega) \geq \beta|w^\infty(\hat{x},y)| - c\epsilon$$

for all $\beta \in \mathbb{R}$. This completes the proof. ∎

REMARK 3.3. *In general we would like to know if the implication*

$$\Omega \not\subset \Omega_t \implies \mu_\epsilon(\Omega_t,\Omega) = \infty$$

is true. It would immediately yield a convergence proof of the no response test to find the support of unknown scatterers. This implication is strongly linked to the uniqueness question for the inverse scattering problem under consideration, for which to date there is no proof. Colton and Sleeman [5] *have proven uniqueness for the problem with Dirichlet boundary data given a finite number of incident fields and a priori information about the size of the scatterer. The number of incident fields required depends on the size of the scatterer and the wavelength of the incident field. Alternatively, we could try to prove*

$$\Omega \not\subset \Omega_t \implies \mu_\epsilon(\Omega_t,\Omega) > C$$

for the smallest constant $C = c\epsilon$ for which Theorem 3.2 *is true. This would also lead to a convergence proof for the no response test under the condition that the right constant C is chosen appropriately for the judgment about a test domain. Both problems will be part of future research.*

In the following corollary to Theorem 3.2, we use (3.2) to show that the far field pattern on a limited aperture Γ resulting from excitation by a single incident field uniquely determines the union of all translations of a fixed test domain Ω_t^0 for which μ_ϵ is finite. This is stated precisely below.

DEFINITION 3.4 (corona of Ω corresponding to Ω_t^0). *Let Ω_t^0 denote a fixed, bounded test domain with C^2 boundary. Denote translations of Ω_t^0 by $\Omega_t^0(z) := \Omega_t^0 + z$ for $z \in \mathbb{R}^m$. Define the* corona *of the scatterer Ω by*

$$M(\Omega_t^0,\Omega,\hat{x}) := \bigcup\left\{\Omega_t^0(z) : \ z \in \mathbb{R}^m, \mu_\epsilon(\Omega_t^0(z),\Omega,\hat{x}) < \infty\right\}. \tag{3.3}$$

COROLLARY 3.5 (uniqueness and bounds for the corona). *Let $\Omega_t^0 \subset \mathbb{R}^m$ with $\mathbb{R}^m \setminus \overline{\Omega_t^0}$ connected be a bounded domain large enough that there is some $z \in \mathbb{R}^m$ for which $\Omega \subset \Omega_t^0(z)$, where Ω denotes the support of the scatterer. Then we have*

$$M(\Omega_t^0,\Omega,\hat{x}) \subset \bigcup\left\{\Omega_t^0(z) : \ z \in \mathbb{R}^m, \overline{\Omega_t^0}(z) \cap \overline{\Omega} \neq \emptyset\right\} \tag{3.4}$$

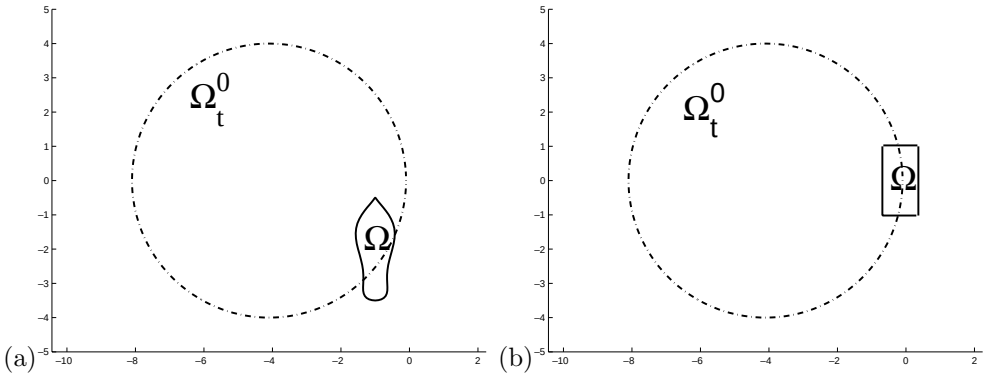


FIG. 1. *Scatterer Ω and test domain Ω_t^0 used in reconstruction simulations. The obstacle Ω in* (a) *is used for Dirichlet, Neumann, and impedance obstacle reconstructions. The scatterer shown in* (b) *is used for inhomogeneous media reconstructions.*

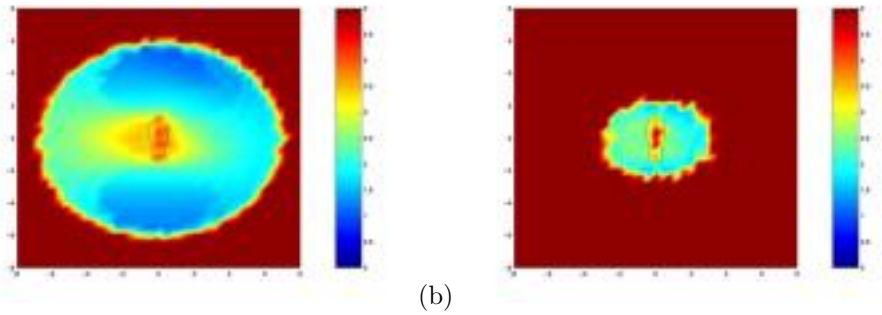
and the scatterer Ω is a subset of its corona, $M(\Omega_t^0, \Omega, \hat{x})$. Moreover, the corona is uniquely determined by the far field pattern for scattering of one plane wave with direction of incidence $-\hat{x}$.

Proof. For points z with $\mu_\epsilon(\Omega_t^0(z)) < \infty$, we apply Theorem 3.2 to conclude that $\overline{\Omega_t^0}(z) \cap \overline{\Omega} \neq \emptyset$, from which we immediately obtain the relation (3.4). For $\Omega \subset \Omega_t^0(z)$ we have $\mu_\epsilon(\Omega_t^0(z)) < \infty$ and thus the support of the scatterer is a subset of its corona: $\Omega \subset M$.

Using (3.2) the values of $\mu_\epsilon(\Omega_t^0(z), \Omega, \hat{x})$ can be calculated directly from the limited aperture far field pattern $u^\infty(\hat{y}, -\hat{x})$, $\hat{y} \in \Gamma$; that is, for fixed test domain Ω_t^0 and direction $\hat{x}$, the scattering test response $\mu_\epsilon(\Omega_t^0(z), \Omega, \hat{x})$ is a scalar-valued mapping of $\hat{x}$. Since the direction of incidence of a plane wave uniquely determines the far field pattern $u^\infty(\cdot, -\hat{x})$, then the corona is uniquely determined by the far field pattern u^∞. ∎

The corona corresponding to the circular test domain of a boat-shaped scatterer (see Figure 1) is shown in Figure 2. From the above uniqueness theorem we know that the unknown scatterer—whatever its physical nature might be—is located in the corona. Note that, at the very least, we can use the center of the corona for a single incident wave as an estimate for the center of the obstacle. In our experiments here, however, we are able to extract even more information about the scatterer from the corona. Recall from Remark 3.3 that we cannot say anything specific about the behavior of μ_ϵ for $\overline{\Omega} \cap \overline{\Omega_t^0}(z) \neq \emptyset$ when $\Omega \not\subset \Omega_t^0(z)$. We observe numerically that the value of the scattering test response increases as the intersection $\overline{\Omega} \cap \overline{\Omega_t^0}(z) \neq \emptyset$ becomes smaller. We therefore propose a technique that allows us to detect these increases, and thereby detect the location and shape of the scatterer within the corona. We begin by describing the choice of the test domain and the calculation of the scattering test response. Details for efficient implementation together with the algorithm are given in the following section.

The test domain Ω_t^0. For fixed scatterers Ω and incident wave directions $\hat{x}$, the scattering test response μ_ϵ takes as input the test domains $\Omega_t^0(z)$ and returns a scalar value as output. We would like to know which test domains $\Omega_t^0(z)$ yield small values for μ_ϵ without having to work with the unwieldy domains themselves. For this, we construct a mapping from the domain $\Omega_t^0(z)$ to the point $z' \in \Omega_t^0(z)$ and assign to



(a) (b)

FIG. 2. *The figure demonstrates the calculated corona and the corresponding bound for the location of the unknown obstacle when* Ω_t^0 *is a circle with radius* $r_t = 4$ *as shown in Figure* 1(a). *For this bound we do not need to know the physical nature of the scatterer and only* one *scattered wave is necessary. Here we used the wave number* $\kappa = 5$, *aperture opening* $\theta = 0.9\pi$, *regularization parameter* $\alpha = 10^{-11}$ *for an incident wave with direction* $(-1, 0)$. *The far field pattern contains* 1–2% *errors. The figure demonstrates the change in the corona for different choices of the approximation domain:* (a) *shows the corona for an approximation domain of radius* $r_t = 4$, (b) *shows the corona for an approximation domain of radius* $r_t = 1.3$.

that point the corresponding value of μ_ϵ. It is how we choose the point z' that allows us to get much more information about the obstacle than we would expect.

The key property that we exploit is the observation that μ_ϵ grows as the intersection $\Omega \cap \Omega_t^0(z)$ becomes smaller. We emphasize that this observation is *empirical*, since at this time we cannot prove anything about the behavior of μ_ϵ in this situation. In order to detect this growth, assign the domain $\Omega_t^0(z)$ to a point z' on the *boundary* $\partial\Omega_t^0(z)$. To avoid keeping track of more points than necessary, we construct the generating domain Ω_t^0 such that $0 \in \partial\Omega_t^0$ and map the translated domain $\partial\Omega_t^0(z)$ to the point $z' = z$. When $z \in \Omega$, the scatterer will not to fall entirely within the domain $\Omega_t^0(z)$. In this case we observe that the scattering test response μ_ϵ is significantly higher than when z is in the parts of the corona that do not intersect with the scatterer, representing the situation where $\Omega \subset \Omega_t^0(z)$.

We now describe a special realization of the no response test. We assign to the point z the value of the scattering test response

$$f^*(z; \Omega_t^0) := \mu_\epsilon(\Omega_t^0(z), \Omega). \tag{3.5}$$

However, by restriction to the point z from the full set $\overline{\Omega_t^0}(z)$, we lose information: we obtain small values for $f^*(z; \Omega_t^0)$ only on one side of the unknown object as shown in Figure 3, where $f^*(z; \Omega_t^0)$ is plotted. The full information is recovered by rotating the generating domain Ω_t^0 around the origin and repeating the above procedure. This is described in detail next.

Rotations or other variations of the test domain are necessary because of our choice of the mapping from $\Omega_t^0(z)$ to the point $z' \in \Omega_t^0(z)$. Had we chosen a radially symmetric generating domain Ω_t^0 and mapped this domain to its center, rotations would not be necessary. However, in this case the image does not directly reflect the behavior of the test response that we use to reconstruct the obstacle, not just its corona, that is, the behavior of the test response when the boundary of the test domain intersects the scatterer. Note that the idea of monitoring the behavior at the boundary of the test domain also appears in the enclosure method, where the test domain is a half-space and the behavior of the indicator function (1.2) indicates which

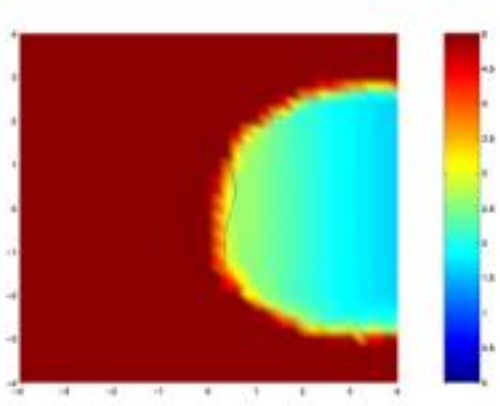


FIG. 3. *The figure shows a plot of the function* $f^*(z;\Omega_t^0)$ *given by* (3.5) *on a grid containing the unknown scatterer. The scatterer is indicated by the black curve. Here, we used a Dirichlet boundary condition.*

half-space the obstacle belongs to [6, 9]. Alternatively, we could use an arbitrary set of generating domains Ω_t^0 with $0 \in \partial\Omega_t^0$; however, it is much more convenient to work with a single generating domain rotated about the origin.

The value $f^*(z;\Omega_t^0)$ assigned to the point z via (3.5) at one rotation of the domain Ω_t^0 does not necessarily correspond to the value of the same point at a different rotation. To see this, suppose that the scatterer Ω is contained in a small circle of radius 1 centered at the point $(-1,0)$. Suppose further that Ω_t^0 is a circle of radius 2 with center $(-2,0)$. In this case $\Omega \subset \Omega_t^0$ and $f^*(0;\Omega_t^0)$ will therefore be small. If we rotate Ω_t^0 about the origin by 180° and denote the resulting domain by $\widetilde{\Omega_t}^0$, then $\Omega \cap \widetilde{\Omega_t}^0 = \{0\}$. In this case we observe that the corresponding value for $f^*(0;\widetilde{\Omega_t}^0)$ will be large. In each case we assign a value to the point $z = 0$, but clearly the values do not correspond to the same situation. In order to prevent the large values of one orientation from drowning out the information contained in the small values from other rotations, we take the minimum of the values assigned to the points z over all rotations.

Let R_θ denote the rotation operator mapping the domain Ω_t^0 onto the rotated domain $R_\theta\Omega_t^0$. If at a point z the value $F(z;\Omega_t^0)$,

$$F(z;\Omega_t^0) := \inf_{\theta\in[0,2\pi]} f^*(z;R_\theta\Omega_t^0), \tag{3.6}$$

is large, then we suppose that the unknown obstacle lies partly outside all rotations of the test domain about this point. In this way, by sampling all points z in and around the unknown scatterer Ω we are able to reconstruct aspects of the shape, location, and size of Ω. Details about how we implement this are given next.

4. Implementation and numerical demonstrations.

Calculating the densities g. As prescribed in (3.1), we construct incident fields that are small on the test domain $\Omega_t^0(z)$. For this we approximate the fundamental solution to the Helmholtz equation $\Phi(x,y)$ where the singularity is located at a point $y \in \mathbb{R}^m$ sufficiently far away from $\Omega_t^0(z)$. To construct the densities g corresponding

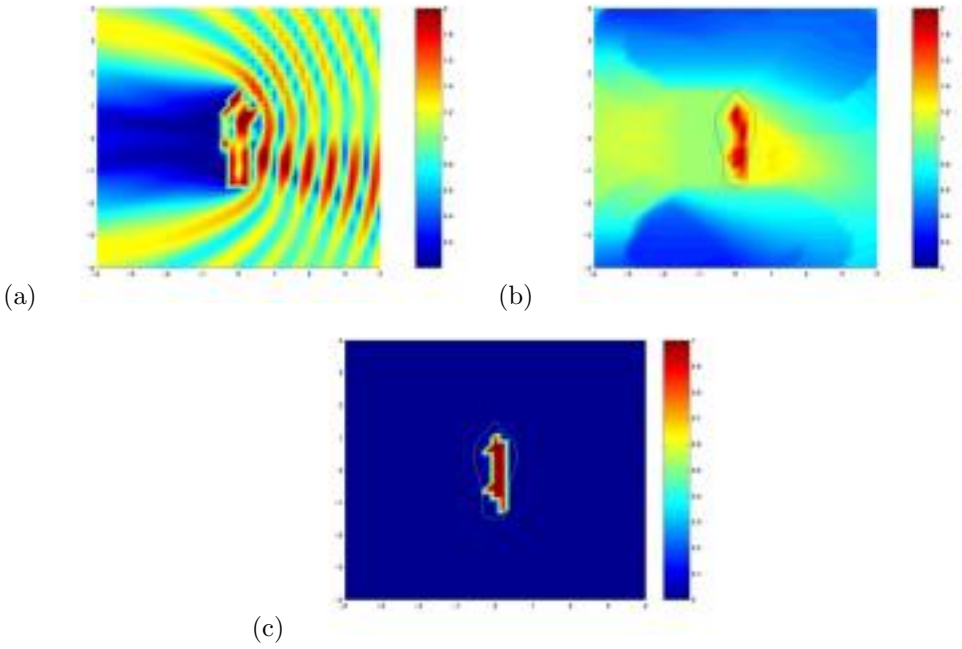


FIG. 4. (a) *Original total field for scattering by a Dirichlet obstacle.* (b) *A plot of the function* $F(z)$ *defined by* (3.6) *for* $z \in \mathcal{G}$*, the computational grid.* (c) *Thresholded version of the function* F *with* $C = 1.4$ *(see Algorithm* 4.1*). Here we used the wave number* $\kappa = 5$*, aperture opening* $\theta = 1.8\pi$*, regularization parameter* $\alpha = 10^{-11}$ *for an incident wave with direction* $(-1, 0)$*. The far field pattern contains* 1–2% *errors.*

to these incident fields, we use Tikhonov regularization to approximately solve the ill-posed equation

$$\Big(H_z g(\cdot, y)\Big)(x) = \Phi(x, y) \qquad \text{for } x \in \partial\Omega_t^0(z),$$

where $H_z : L^2(\Gamma) \to L^2(\partial\Omega_t^0(z))$ is a (limited angle) Herglotz wave operator defined by

$$(H_z g)(x) := \int_\Gamma e^{i\kappa x\cdot(-\hat{y})} g(-\hat{y}) ds(\hat{y}), \qquad x \in \partial\Omega_t^0(z). \tag{4.1}$$

Specifically, for the regularization parameter $\alpha > 0$, we define

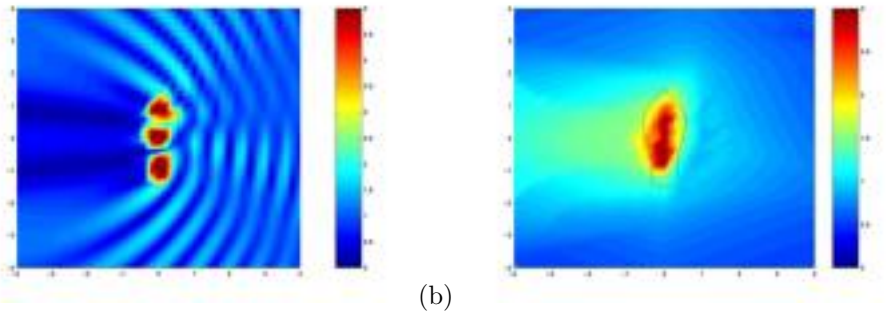
$$g_{z,\alpha}(\cdot, y) := (\alpha I + H_z^* H_z)^{-1} H_z^* \Phi(\cdot, y), \tag{4.2}$$

where the argument y of the density $g_{z,\alpha}$ denotes the dependence of the density on the location of the singularity in Φ. The subscripts z and α on g denote the dependence of the density on the regularization parameter α and the test domain $\Omega_t^0(z)$. This yields

$$v^i[g_{z,\alpha}(\cdot, y)](\cdot) \approx \Phi(\cdot, y) \quad \text{on } \partial\Omega_t^0(z).$$

On $\Omega_t^0(z)$, for $d(y, \Omega_t^0(z)) \geq \rho$ we have, for all $\alpha \in [0, \alpha_0]$ for fixed α_0 sufficiently small,

$$|v^i[g_{z,\alpha}(x, y)]| \leq \begin{cases} \frac{c}{\sqrt{\rho}}, & m = 2, \\ \frac{c}{\rho}, & m = 3, \end{cases} \qquad x \in \Omega_t^0(z), \tag{4.3}$$



(a) (b)

FIG. 5. (a) *Original total field for scattering by an impedance obstacle with* $\lambda = i$. (b) *A plot of the function* $F(z)$ *defined by* (3.6) *for* $z \in \mathcal{G}$, *the computational grid. Here we used the wave number* $\kappa = 5$, *aperture opening* $\theta = 1.8\pi$, *regularization parameter* $\alpha = 10^{-11}$ *for an incident wave with direction* $(-1, 0)$. *The far field pattern contains* 1–2% *errors.*

with some constant c that is typically of size smaller than 10^1. Thus by knowing the value of $v^i[g_{z,\alpha}]$ at the point closest to the source point we obtain upper bounds on the size of the incident field on all of $\Omega_t^0(z)$. This is used in the calculation of the scattering test response μ_ϵ. On the exterior region $\mathbb{R}^m \setminus \overline{\Omega}_t$ the magnitude of $v^i[g_{z,\alpha}]$ is in the range of $c/2\alpha$. For example, $|v^i[g_{z,\alpha}]|$ is of size 50 if $c = 1$ and $\alpha = 10^{-2}$. This corresponds to a data error of one percent.

Translations of the test domain. We describe a quick method to calculate lower estimates for the test response μ_ϵ for a large number of translated test domains $\Omega_t^0(z)$ with generating domain Ω_t^0. In the moving reference frame of the test domain, spatial translations look like translations of the incident field. We use this and the fact that phase shifts in the far field correspond to spatial translations in the near field in order to translate the generating domain Ω_t^0 around the computational domain.

Let Ω_t^0 be a generating test domain and define $\Omega_t^0(z) = \Omega_t^0 + z$ to be the corresponding translated test domain. Translations of the Herglotz wave function $v^i[g]$ can be easily performed by the multiplication of the density g by the complex factor $e^{-i\kappa z \cdot d}$. At points $z \in \mathbb{Q}$ covering the area where the unknown scatterer is supposed to be (that is, $\mathbb{Q}$ is the computational domain and satisfies $\Omega \subset \mathbb{Q}$) we calculate translations $\Omega_t^0(z)$ of the test domain Ω_t^0 by the corresponding translation of the Herglotz wave function $v^i[g]$:

$$v^i[g](x - z) = v^i[e^{-i\kappa z \cdot (\cdot)} g(\cdot)](x), \qquad x \in \mathbb{R}^m. \tag{4.4}$$

We define the function $|v^\infty[g](\hat{x}, z)|$ to be the far field pattern at the point $\hat{x} \in \mathbb{S}$ for scattering of the shifted incident field $v^i[g](x - z)$. Then from Theorem 2.1 and (3.2) we obtain

$$|v^\infty[g](\hat{x}, z)| = \left| \int_\Gamma u^\infty(\hat{y}, -\hat{x}) e^{i\kappa z \cdot \hat{y}} g(-\hat{y}) ds(\hat{y}) \right|. \tag{4.5}$$

In words, the magnitude of the far field v^∞ at the point $\hat{x} \in \mathbb{S}$ with test domain $\Omega_t^0(z)$ is given by the magnitude of the weighted superposition of the measured far field pattern due to a single incident plane wave excitation; the weight $g(-\hat{y})$ is determined by the generating domain Ω_t^0 and the phase shift is determined by the translation of Ω_t^0.

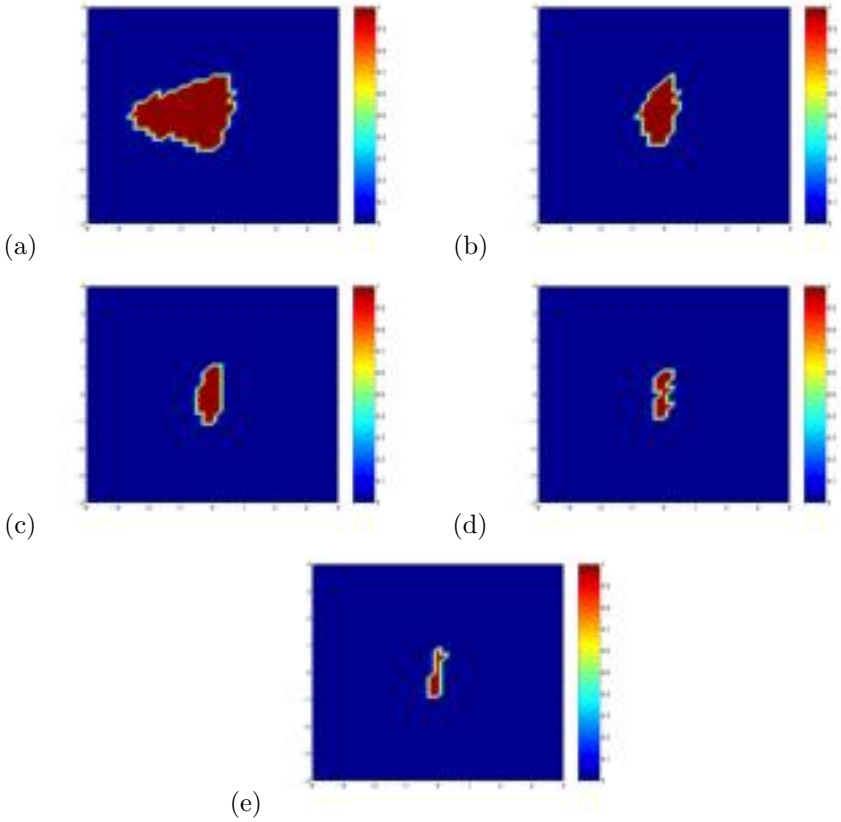


FIG. 6. *We show different thresholded versions of the reconstruction to demonstrate the influence of the cut-off parameter C. The wave number is $\kappa = 5$, aperture opening $\theta = 1.8\pi$, regularization parameter $\alpha = 10^{-11}$ for an incident wave with direction $(-1, 0)$; that is, the incident wave is coming from the right-hand side. The far field pattern contains 1–2% errors.*

Sampling the scattering test response. The scattering test response $\mu_\epsilon(\Omega_t^0(z), \Omega)$ is the supremum over $|v^\infty[g](\hat{x}, z)|$, where $g \in L^2(\Gamma)$ is chosen such that

$$||v^i[g]||_{C(\Omega_t^0(z))} \leq \epsilon. \tag{4.6}$$

We choose a finite subset $\{g_1, \ldots, g_{n_y}\}$ of densities g such that (4.6) is satisfied and calculate the maximum of the values $|v^\infty[g_j](\hat{x}, z)|$ for $j = 1, \ldots, n_y$ via (4.5). To obtain different densities g, we solve (4.2) at the points y_j, $j = 1, \ldots, n_y$, in the exterior of $\Omega_t^0(z)$. In our experiments we chose $n_y \approx 20$. Using the efficient translations in (4.4), we need only solve (4.2) for g with Ω_t^0 for each $j = 1, \ldots, n_y$, rather than solving for g for every translated domain $\Omega_t^0(z)$. Also, from the discussion following (4.3), if we choose the points y_j appropriately, there is no need to check explicitly if condition (4.6) is satisfied.

The no response algorithm. We finish this section with a detailed prescription for using the no response test to locate an unknown obstacle.

ALGORITHM 4.1 (no response test).

- *Choose an appropriate test domain Ω_t^0 with $0 \in \partial\Omega_t^0$ that is large enough such that translations of Ω_t^0 and its rotated versions may contain the unknown*

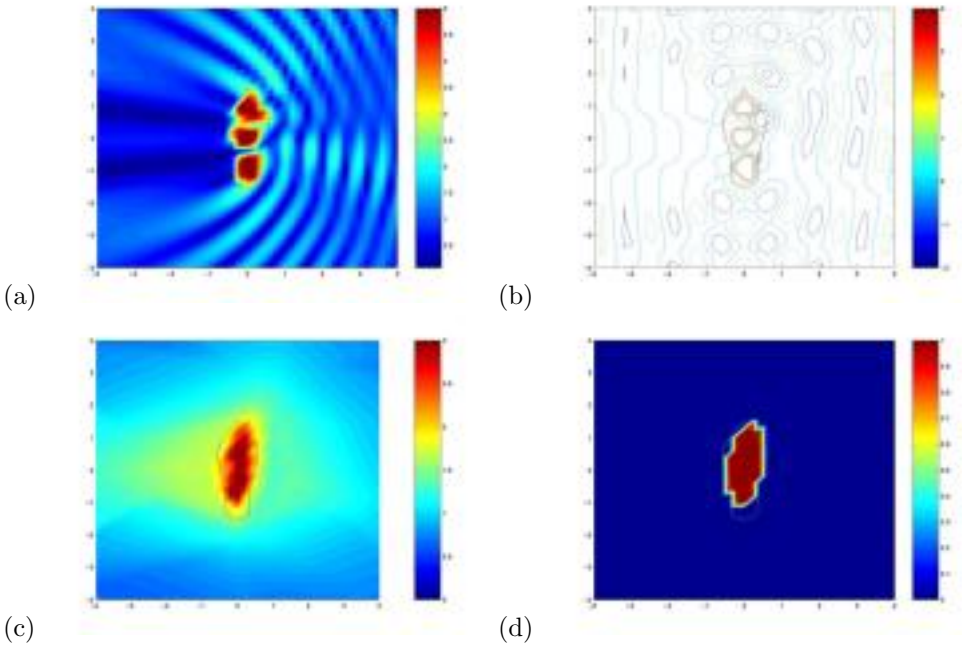


FIG. 7. (a)–(b) *Original total field for scattering by an obstacle with Neumann boundary condition; we show a surface and a contour plot of the field.* (c) *A plot of the function* $F(z)$ *defined by* (3.6) *for* $z \in \mathcal{G}$ *and* (d) *a thresholded version of the reconstruction with* $C = 2.0$. *Here we used the wave number* $\kappa = 5$, *aperture opening* $\theta = 1.8\pi$, *regularization parameter* $\alpha = 10^{-11}$ *for an incident wave with direction* $(-1, 0)$. *The far field pattern contains* 1–2% *errors.*

scatterer (see Figure 1*).*

- *For the angles* $\theta_l := 2\pi l/n_r$ *with* $l = 1, \ldots, n_r$, *let* $R_{\theta_l}\Omega_t^0$ *be the domain that is obtained from* Ω_t^0 *by rotation around the origin by angle* θ_l *as described in section* 3. *For each* $l = 1, \ldots, n_r$ *do the following:*
  - *Choose points* y_j, $j = 1, \ldots, n_y$, *in the exterior of* $R_{\theta_l}\Omega_t^0$ *and calculate the density* $g_{l,j}$ *by*

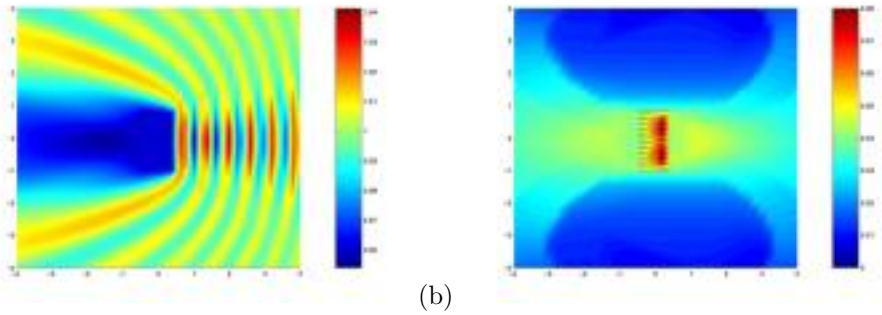
$$g_{l,j} := (\alpha I + H_{\theta_l}^* H_{\theta_l})^{-1} H_{\theta_l}^* \Phi(\cdot, y_j),$$

    where H_{θ_l} *is the Heglotz wave operator (see* (4.1)*) corresponding to the rotated domain* $R_{\theta_l}\Omega_t^0$.
  - *For each* $j = 1, \ldots, n_y$ *calculate*

$$f_j(z; R_{\theta_l}\Omega_t^0) := \left| \int_\Gamma u^\infty(\hat{y}, -\hat{x}) e^{i\kappa z \cdot \hat{y}} g_{l,j}(-\hat{y}) ds(\hat{y}) \right|$$

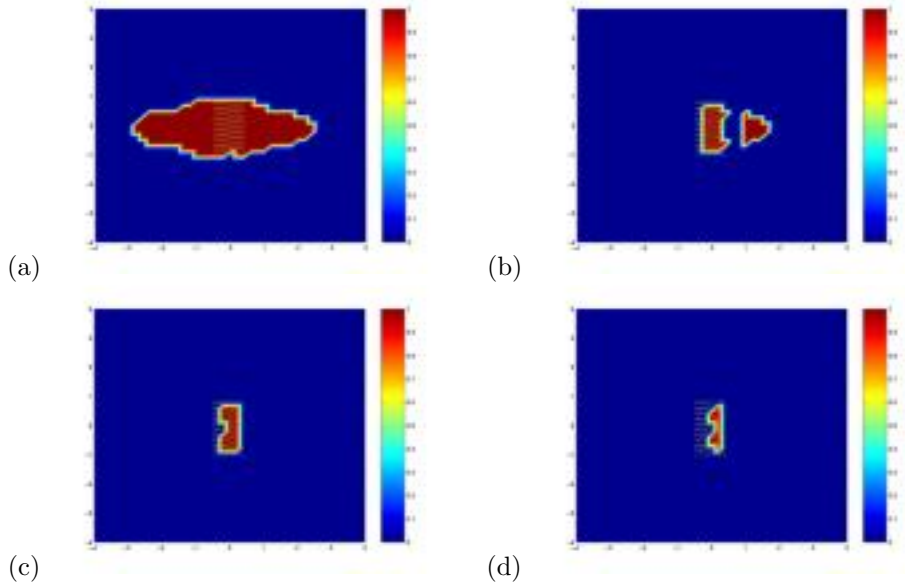
    for all $z \in \mathcal{G}$, *the computational grid, from the one-wave far field pattern* $u^\infty(\hat{y}, -\hat{x})$, $\hat{y} \in \Gamma$.
  - *Calculate the maximum with respect to the densities* $g_{l,j}$; *that is, calculate the sampled version of* (3.5)*:*

$$f^*(z; R_{\theta_l}\Omega_t^0) := \max_{j=1,\ldots,n_y} f_j(z; R_{\theta_l}\Omega_t^0), \qquad z \in \mathcal{G}.$$



(a) (b)

FIG. 8. (a) *Original total field for scattering by a homogeneous penetrable medium with* $n := 4$ *in* Ω *where the inhomogeneity is shown in Figure* 1. (b) *A plot of the function* $F(z)$ *defined by* (3.6) *for* $z \in \mathcal{G}$. *Here we used the wave number* $\kappa = 5$, *aperture opening* $\theta = 1.8\pi$, *regularization parameter* $\alpha = 10^{-11}$ *for an incident wave with direction* $(-1, 0)$. *The far field pattern contains* 1–2% *errors.*



(a) (b)

(c) (d)

FIG. 9. (a)–(d) *We show several thresholded versions of the function* F *for the reconstruction of an inhomogeneous medium, where we used the thresholds* $C = 0.1$, $C = 0.06$, $C = 0.05$, *and* $C = 0.045$.

- *Calculate the minimum with respect to the rotations* θ_l, *that is, the sampled version of* (3.6)*:*

$$F(z; \Omega_t^0) := \min_{l=1,\ldots,n_r} f^*(z; R_{\theta_l} \Omega_t^0), \qquad z \in \mathcal{G}.$$

- *Choose a threshold* C *and calculate*

$$\Omega_{rec} := \{ z \in \mathcal{G} : \ F(z; \Omega_t^0) \geq C \}.$$

Now, an approximation for the support Ω *of the unknown scatterer is given by the components of* Ω_{rec} *that are not connected with infinity.*

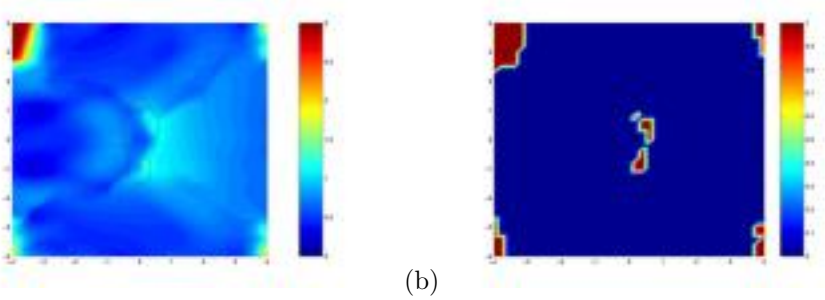


FIG. 10. *Limited aperture reconstruction of a Dirichlet obstacle.* (a) *A plot of the function* $F(z)$ *defined by* (3.6) *for* $z \in \mathcal{G}$*, the computational grid.* (b) *Thresholded version of the function* F*. Here we used the wave number* $\kappa = 5$*, aperture opening* $\theta = 0.6\pi$*, regularization parameter* $\alpha = 10^{-11}$ *for an incident wave with direction* $(-1, 0)$*. The far field pattern contains* 1–2% *errors.*

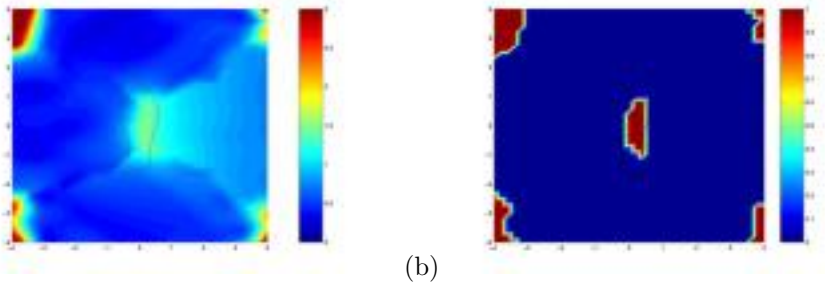


FIG. 11. *Limited aperture reconstruction of a Neumann obstacle.* (a) *A plot of the function* $F(z)$ *defined by* (3.6) *for* $z \in \mathcal{G}$*.* (b) *Thresholded version of the function* F*. Here we used the wave number* $\kappa = 5$*, aperture opening* $\theta = 0.6\pi$*, regularization parameter* $\alpha = 10^{-11}$ *for an incident wave with direction* $(-1, 0)$*. The far field pattern contains* 1–2% *errors.*

For the choice of the constant C we propose dynamical thresholding on the image $F(z)$ that is informed by a priori knowledge about the approximate size of the object.

Numerical results. All the following numerical reconstruction procedures are based on the same algorithm independent of the boundary condition or physical nature of the scatterer. All reconstructions use the far field data for *one* wave only. We show results for full and limited aperture data.

To compare different reconstructions for obstacles with different boundary condition for all the following pictures, we used the far field pattern for one wave with direction of incidence $(-1, 0)$. We first show results for full aperture and demonstrate the influence of the cut-off parameter (Figures 4–9).

In a second part, we restrict our measurements to a limited aperture. Here, we would like to show that even with limited aperture the method yields reasonable results. Figures 10 to 13 show limited aperture reconstructions for the Dirichlet, Neumann, and impedance boundary condition and for the inhomogeneous medium. We used $\kappa = 5$ and an aperture of 0.6π, or $108°$.

5. Concluding remarks. The no response test is a novel sampling technique for reconstructing the support of unknown scatterers. The method does not require a priori knowledge about the physical or geometric properties of the unknown scatterer. Reconstructions can be obtained from the far field pattern for scattering of a single

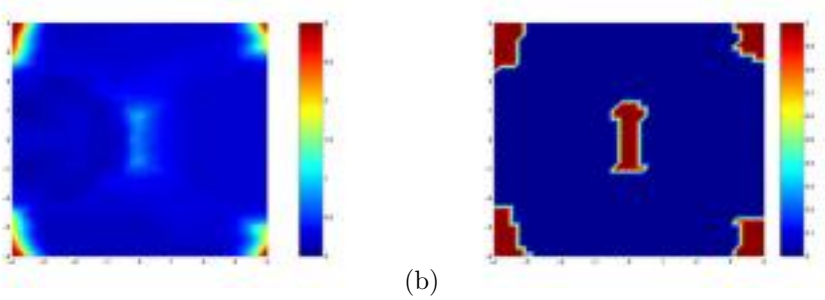


FIG. 12. *Limited aperture reconstruction of an impedance obstacle with* $\lambda = i$. (b) *A plot of the function* $F(z)$ *defined by* (3.6) *for* $z \in \mathcal{G}$, *the computational grid.* (c) *Thresholded version of the function* F. *Here we used the wave number* $\kappa = 5$, *aperture opening* $\theta = 0.6\pi$, *regularization parameter* $\alpha = 10^{-11}$ *for an incident wave with direction* $(-1, 0)$. *The far field pattern contains* 1–2% *errors.*

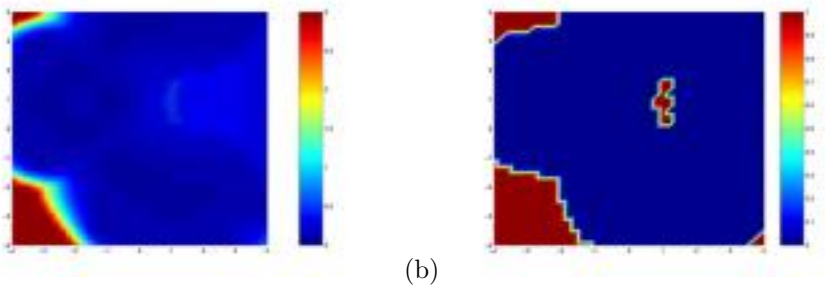


FIG. 13. *Limited aperture reconstruction of the support of an inhomogeneous medium.* (a) *A plot of the function* $F(z)$ *defined by* (3.6) *for* $z \in \mathcal{G}$. (b) *Thresholded version of the function* F. *Here we used the wave number* $\kappa = 5$, *aperture opening* $\theta = 0.6\pi$, *regularization parameter* $\alpha = 10^{-11}$ *for an incident wave with direction* $(-1, 0)$. *The far field pattern contains* 1–2% *errors.*

incident wave. The method appears to be robust and can be used in limited aperture settings. The results and open questions discussed in this work offer a new perspective on some old questions (for example, the uniqueness question for one wave) and provide new directions for future research, both in numerical techniques and analysis.

REFERENCES

[1] D. COLTON AND A. KIRSCH, *A simple method for solving inverse scattering problems in the resonance region*, Inverse Problems, 12 (1996), pp. 383–393.

[2] D. COLTON AND R. KRESS, *Inverse Acoustic and Electromagnetic Scattering Theory*, 2nd ed., Springer–Verlag, Berlin, 1998.

[3] D. COLTON AND P. MONK, *A novel method for solving the inverse scattering problem for time-harmonic acoustic waves in the resonance region*, SIAM J. Appl. Math., 45 (1985), pp. 1039–1053.

[4] D. COLTON AND P. MONK, *A novel method for solving the inverse scattering problem for time-harmonic acoustic waves in the resonance region* II, SIAM J. Appl. Math., 46 (1986), pp. 506–523.

[5] D. COLTON AND B. D. SLEEMAN, *Uniqueness theorems for the inverse problem of acoustic scattering*, IMA J. Appl. Math., 31 (1983), pp. 253–259.

[6] M. IKEHATA, *Reconstruction of an obstacle from the scattering amplitude at a fixed frequency*, Inverse Problems, 14 (1998), pp. 949–954.

[7] M. IKEHATA, *Enclosing a polygonal cavity in a two-dimensional bounded domain from Cauchy data*, Inverse Problems, 15 (1999), pp. 1231–1241.

[8] M. Ikehata, *On reconstruction in the inverse conductivity problem with one measurement*, Inverse Problems, 16 (2000), pp. 785–793.
[9] M. Ikehata and T. Ohe, *A numerical method for finding the convex hull of polygonal cavities using the enclosure method*, Inverse Problems, 18 (2002), pp. 111–124.
[10] A. Kirsch, *Characterization of the shape of a scattering obstacle using the spectral data of the far field operator*, Inverse Problems, 14 (1998), pp. 1489–1512.
[11] R. Lies, *Initial Boundary Value Problems in Mathematical Physics*, Teubner, Stuttgart, 1985.
[12] D. Mitrea, M. Mitrea, and M. Taylor, *Layer potentials, the Hodge Laplacian, and global boundary problems in nonsmooth Riemannian manifolds*, Mem. Amer. Math. Soc., 150 (2001).
[13] O.I. Panich, *On the question of the solvability of the exterior boundary-value problems for the wave equation and Maxwell's equations*, Uspehi Mat. Nauk., 20 (1965), pp. 221–226 (in Russian).
[14] R. Potthast, *A fast new method to solve inverse scattering problems*, Inverse Problems, 12 (1996), pp. 731–742.
[15] R. Potthast, *A point-source method method for inverse acoustic and electromagnetic obstacle scattering problems*, IMA J. Appl. Math., 61 (1998), pp. 119–140.
[16] R. Potthast, *Point Sources and Multipoles in Inverse Scattering Theory*, Chapman and Hall, London, 2001.

SIAM J. APPL. MATH.
Vol. 63, No. 4, pp. 1313–1327

VIRUS DYNAMICS: A GLOBAL ANALYSIS*

PATRICK DE LEENHEER† AND HAL L. SMITH†

Abstract. Exploiting the fact that standard models of within-host viral infections of target cell populations by HIV, developed by Perelson and Nelson [*SIAM Rev.*, 41 (1999), pp. 3–44] and Nowak and May [*Virus Dynamics*, Oxford University Press, New York, 2000], give rise to competitive three dimensional dynamical systems, we provide a global analysis of their dynamics. If the basic reproduction number $R_0 < 1$, the virus is cleared and the disease dies out; if $R_0 > 1$, then the virus persists in the host, solutions approaching either a chronic disease steady state or a periodic orbit. The latter can be ruled out in some cases but not in general.

Key words. virus dynamics, global stability, oscillations, HIV

AMS subject classifications. 34D23, 34A34

PII. S0036139902406905

1. Introduction.

Recently there has been a substantial effort in the mathematical modeling of virus dynamics, primarily motivated by the AIDS epidemic and HIV; see, e.g., [9, 11, 15]. Perelson and Nelson [14] and Nowak and May [12] provide excellent reviews and many more citations. The latter has a somewhat broader focus, also treating SIV (the simian version of HIV) and the hepatitis B viral infections. These models focus on the disease dynamics within an infected individual and contrast with an earlier parallel literature on the dynamics within the human population. Simple HIV models have played a significant role in the development of a better understanding of the disease and the various drug therapy strategies used against it. For example, they provided a quantitative understanding of the level of virus production during the long asymptomatic stage of HIV infection; see [13, 14, 12].

We focus primarily on HIV models here but note, following [12], that the basic model applies to many other viral infections. Moreover, similar models exist which describe infections of marine bacteria by bacteriophages; see [1].

A brief review of the salient features of the role of HIV in the disease will be useful. The course of an HIV infection is as follows. First, HIV enters its target, a T cell. Inside this cell it makes a DNA copy of its viral RNA; hence it falls into the class of so-called retroviruses. In this process it needs the enzyme reverse transcriptase (RT). The viral DNA is then inserted into the DNA of the T cell, which will henceforth produce viral particles that can bud off the cell to infect other uninfected T cells. Before leaving the host cell, the virus particle is equipped with protease, an enzyme used to cleave a long protein chain. If this feature is lost, the virus particle is not capable of successfully infecting other T cells.

The models considered in [14, 12] have three state variables: T, the concentration of uninfected T cells; T^*, the concentration of productively infected T cells; and V, the concentration of free virus particles in the blood. In chemical reaction notation, the model can be written

$$T + V \to T^* \to NV,$$

*Received by the editors May 1, 2002; accepted for publication (in revised form) September 30, 2002; published electronically April 23, 2003.
http://www.siam.org/journals/siap/63-4/40690.html
†Department of Mathematics and Statistics, Arizona State University, Tempe, AZ 85287 (leenheer@math.la.asu.edu, halsmith@asu.edu). The research of the second author was supported by NSF grant DMS 9700910.

because mass action reaction terms are used and each infected T cell is assumed to produce N viral particles over its lifespan. The interaction between these cells and virus particles is then given by the following equations:

$$\begin{aligned} \dot{T} &= f(T) - kVT, \\ \dot{T}^* &= -\beta T^* + kVT, \\ \dot{V} &= -\gamma V + N\beta T^*, \end{aligned} \tag{1}$$

where we have relabeled many of the parameters used in [14, 12]. The functional form of f is defined differently by different authors:

1. Perelson and Nelson [14]: $f(T) = f_1(T) \equiv \delta - \alpha T + pT(1 - \frac{T}{T_{max}})$.
2. Nowak and May [12]: $f(T) = f_2(T) \equiv \delta - \alpha T$.

The parameters α, β, γ, δ, k, N, p, and $T_{\max}$ are positive.

We briefly summarize the interpretation of the different parameters in the model. Parameters α, β, and γ are the death rates of the uninfected T cells, the infected T cells, and the virus particles, respectively. k is the contact rate between uninfected T cells and virus particles. δ represents a constant production of T cells in the thymus. In the literature this process is not assumed to be constant, but to depend on virus loads. Usually δ is then replaced by a decreasing function of the concentration of virus particles; see, e.g., [15]. N is the average number of virus particles produced by an infected T cell. In the case $f = f_1$, healthy T cells are assumed to proliferate logistically, although the control mechanisms for T cell proliferation are largely unknown. The p and $T_{\max}$ are the growth rate (respectively, carrying capacity) associated with a logistic growth of uninfected T cells in the absence of virus particles, infected T-cells, and natural body sources such as the thymus. Note that simplification of the logistic term $pT(1 - (T + T^*)/T_{\max})$ to $pT(1 - T/T_{\max})$ is not always performed; see, e.g., [15]. From a mathematical point of view, this simplification leads to a competitive system, which opens up a whole arsenal of tools in the subsequent analysis. We will elaborate on this below. Another simplification, found in all models in the literature, is that (logistic) proliferation of T^* cells has been neglected.

Both Perelson and Nelson and Nowak and May ignore the loss term $-kVT$, which should appear in the V equation, i.e.,

$$\dot{V} = -\gamma V + N\beta T^* - kVT, \tag{2}$$

representing the loss of a free virus particle once it enters the target cell, arguing that this small term can be absorbed into the loss term $-\gamma V$. We will consider (1) with and without this added term.

An important feature of this model is that it ignores the reaction of the immune system, and therefore the model describes a worst-case scenario in some sense; see [12, 11] for models which include an immune response to the virus. More realistic models also include a compartment for latently infected T cells [14, 12, 15], which are capable of but not actively producing virus. A related modeling approach consists of incorporating a delay term describing the delay between the time of infection of a T cell and the time of emission of virus particles from this cell [3]. Our model also neglects virus mutations, which occur very frequently and on a fast time-scale. Some of these mutations cause drug resistance, which makes effective treatment very difficult.

System (1), with or without the $-kVT$ term in the V equation, is competitive with respect to the cone $K := \{(X, Y, Z) \in R^3 | \, X, Z \geq 0, Y \leq 0\}$—see p. 49 in [16]—and thus solutions with initial states ordered according to the order of K (i.e., their

difference is a vector in K) remain ordered for backward time. Indeed, the Jacobian matrix of system (1) (respectively of system (1) with the V-equation replaced by (2)) at an arbitrary point of R^3_+ possesses the following structures:

$$\begin{pmatrix} * & 0 & - \\ + & * & + \\ 0 & + & * \end{pmatrix}, \quad \begin{pmatrix} * & 0 & - \\ + & * & + \\ - & + & * \end{pmatrix}, \tag{3}$$

where some of the + and − signs can actually be zero for points on the boundary of R^3_+. Note that these matrices are sign-symmetric; i.e., for every $i \neq j$, the product of the (i, j)th and the (j, i)th entry of these matrices is nonnegative. The incidence graph associated with this matrix, where edges between the nodes are furnished with a + or a − sign, depending on the sign of one of the corresponding entries in the above Jacobian matrix, satisfies the following property: Every closed loop in this graph possesses an odd number of edges with − signs. This property implies that the system is competitive. Alternatively, the change of variables $T^* \to -T^*$ results in a system the Jacobian for which has nonpositive off-diagonal terms on the relevant domain and hence is competitive in the usual sense. The theory of competitive (and cooperative) systems was initiated by Hirsch in a series of six well-known papers, of which we list [5, 6, 7, 8]. Contributions to this theory were also made by Smith, e.g., [17, 18, 20]; see [16] for a review. A particular consequence of the theory of competitive systems is a generalization of the Poincaré–Bendixson theorem to dimension 3; see, e.g., [5, 6] or Theorem 4.1 in [16]: A compact limit set of a competitive system in R^3 which contains no steady states is a periodic orbit. Furthermore, a periodic orbit of a competitive system in R^3 must contain a steady state inside a certain topological closed ball on the surface of which lies the periodic orbit; see Theorem 2.4 in [17]. These results will play a major role in our analysis.

We will also exploit the "isomorphism" between system (1) with $f = f_2$ and the standard SEIR model with constant population size, analyzed by Li and Muldowney in their well-known paper [10]. Although this isomorphism breaks down when $f \neq f_2$ or when the $-kVT$ term is included in the V equation, the method used by Li and Muldowney to prove orbital asymptotic stability of any periodic orbit, and thereby to derive a contradiction to their existence, extends under suitable restrictions.

We identify a basic reproduction number R_0 for the model, which gives the number of infected T cells produced by a single infected T cell in a healthy individual. Our main results are formulated in terms of this number and extend the existing ones in the following five directions:

1. If $R_0 < 1$, we show that the virus is cleared.
2. If $R_0 > 1$, then a chronic disease steady state exists which is globally asymptotically stable under certain conditions. In particular, these conditions are satisfied for the special case $f = f_2$ using parameter values appropriate for HIV.
3. For $f = f_1$, orbitally asymptotically stable periodic orbits are shown to exist and to attract almost all solutions under suitable conditions if $R_0 > 1$. These conditions are apparently not satisfied for HIV. We note that sustained oscillations were observed from numerical simulations by Perelson, Kirschner, and de Boer [15] in a four dimensional model including a compartment of latently infected T cells.
4. Since the function f, which models healthy T cell dynamics, is poorly understood, we start analyzing our model with only minimal assumptions on f. We

show that particular choices for f may lead to different qualitative behavior. For example, for $f = f_2$ the chronic disease steady state, if it exists, is always locally asymptotically stable, while for $f = f_1$ this steady state may be unstable and sustained oscillations may occur. This sensitivity of the behavior to f, in particular, calls for a better understanding of the mechanisms of T cell proliferation.

5. Applications are made to drug therapy following Perelson and Nelson's treatment in [14].

2. Main results. We consider a model of a virus infecting a target cell population. Denoting by T the target cell and using the same symbol for its concentration in the appropriate bodily fluid, we assume that the target cell population is regulated in a healthy individual according to some dynamics given by

$$\dot{T} = f(T),$$

where f is a smooth function. We expect homeostasis to be maintained in a healthy individual with T cell levels at some positive steady state $\bar{T} > 0$. Therefore, assume that f satisfies

$$f(T) > 0,\ 0 \le T < \bar{T}, \quad f(\bar{T}) = 0, \quad f'(\bar{T}) < 0, \quad \text{and} \quad f(T) < 0,\ T > \bar{T}. \tag{4}$$

Consider an individual infected with a virus V which attacks target cells, producing productively infected cells T^*, which, in turn, produce on average N virus particles during their life spans. Following [14, 12], we obtain the following system for the dynamics of T, T^*, V:

$$\begin{aligned} \dot{T} &= f(T) - kVT, \\ \dot{T}^* &= -\beta T^* + kVT, \\ \dot{V} &= -\gamma V + N\beta T^* - ikVT, \end{aligned} \tag{5}$$

where $i = 0$ if we choose, following [14, 12], to ignore the loss of a viral particle when it enters a target cell, or $i = 1$ when we do not.

The basic reproduction number for the model is intuitively determined by considering the fate of a single productively infected cell in an otherwise healthy individual with normal target cell level $T = \bar{T}$. This infected cell produces N virions, each with life span γ^{-1}, which will infect $k\bar{T}N\gamma^{-1}$ healthy target cells. Thus we expect the amplification factor to be $k\bar{T}N\gamma^{-1}$. In fact, a local stability calculation, carried out in the proof of Lemma 3.2 below, leads to

$$R_0 = \frac{k\bar{T}(N - i)}{\gamma}, \tag{6}$$

reflecting the loss of the original productively infected cell if $i = 1$. In any case, as N is typically large, this is a minor point.

Our main result, proved in a series of results in the next section, shows that the global dynamics is largely determined by R_0.

THEOREM 2.1.

1. *For $R_0 < 1$ the only steady state is the disease-free state $E_0 \equiv (\bar{T}, 0, 0)$, and it is globally attracting; the virus is cleared.*

2. *For $R_0 > 1$, in addition to the disease-free state, which is unstable, there is a "chronic disease" steady state $E_e \equiv (T_e, T_e^*, V_e)$ given by*

$$T_e = \frac{\gamma}{k(N-i)}(\equiv \bar{T}/R_0), \quad T_e^* = \frac{\gamma V_e}{(N-i)\beta}, \quad V_e = \frac{f(T_e)}{kT_e}, \tag{7}$$

which is locally attracting if $f'(T_e) \le 0$, e.g., when $f = f_2$.
In particular, with R_0 as a bifurcation parameter, E_0 exchanges its local stability properties with E_e when R_0 passes through 1, making E_e locally attracting if $R_0 > 1$ and $R_0 - 1$ small.
The disease persists in the sense that there exist $\epsilon > 0$ and $M > 0$, independent of initial data (T_0, T_0^*, V_0) satisfying $T_0^* + V_0 > 0$, such that

$$\epsilon < T(t), T^*(t), V(t) < M$$

for all large t.
The omega limit set of every solution with initial conditions as restricted above either contains E_e or is a nontrivial periodic orbit.
If $f'(T) < 0$ for $T \in [0, \bar{T}]$, and denoting $0 < \alpha^* = -\max_{T\in[0,\bar{T}]} f'(T)$, E_e is a globally asymptotically stable steady state for system (5) with respect to initial conditions not on the T axis in case $i = 0$ or in case $i = 1$ and $kf(0) - \min(\alpha^*, \beta)\beta < 0$.
In the special case $f = f_1$, for both $i = 0, 1$ there exist parameter values for which E_e is unstable with a two dimensional unstable manifold (see Lemma 3.4). In this case, there exists an orbitally asymptotically stable periodic orbit; every solution except those with initial data on the one dimensional stable manifold of E_e or on the T axis converges to a nontrivial periodic orbit.

Observe that, as $f(T) > 0$ only if $T < \bar{T}$, the positivity of V_e requires that $T_e < \bar{T}$, or equivalently, $R_0 > 1$.

Our main result says that if a typical productively infected target cell, introduced into an otherwise healthy individual where $T = \bar{T}$, cannot replace itself by producing virus that infects at least one healthy target cell, then the virus is eventually cleared and the individual returns to the disease-free state. However, if the infected cell can replace itself, then the disease persists indefinitely into the future in the sense that the viral load is ultimately bounded from below by an initial-condition-independent value. Moreover, the omega limit set either contains the chronic disease state E_e, coinciding with it in case it is locally attracting, or is a nontrivial periodic orbit. In the latter case, the viral load and the target cell populations cycle periodically.

If $f = f_2$ and $R_0 > 1$, then $f' = -\alpha < 0$ is automatically satisfied and therefore E_e is globally asymptotically stable if $i = 0$ or if $i = 1$ and $kf_2(0) - \min(\alpha^*, \beta)\beta = k\delta - \min(\alpha, \beta)\beta < 0$. In case of HIV, $\alpha \le \beta$ is expected to hold, reflecting the fact that removal rates for healthy target cells are lower than those for infected target cells, and thus the last condition reduces to $k\delta - \alpha\beta < 0$, which is easily verified for the (biologically plausible) numerical data for HIV in [15].

In the special case $f = f_1$, E_e is asymptotically stable when $R_0 > 1$ and $R_0 - 1$ is small, but this stability can be lost for certain parameter values. In Figure 1 below, we show that periodic oscillations in the viral load and T cell populations are possible. The parameter values are not chosen to match those for a particular viral infection; they are chosen simply to establish the possibility for oscillations. As in [15], time is measured in days and T, T^*, V have units mm^{-3}. See Lemma 3.4 for more information about parameter ranges for which periodic solutions are expected.

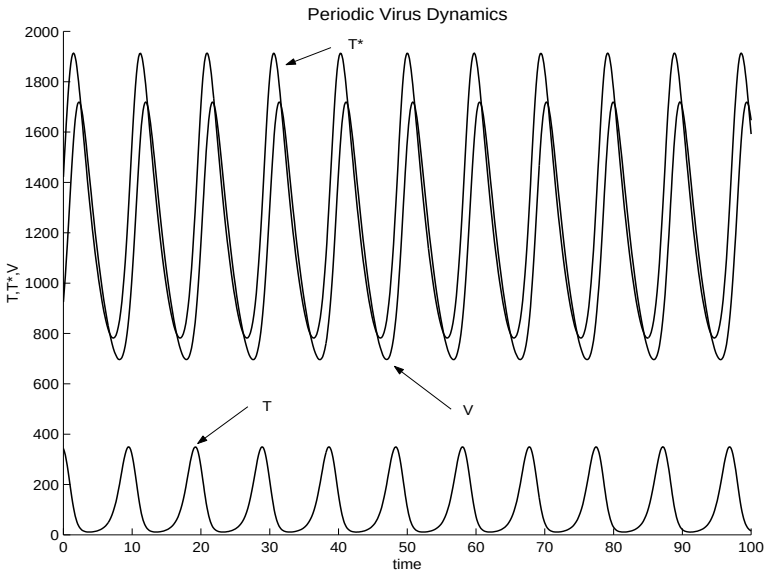


FIG. 1. *Periodic solution for* $f = f_1$. *Parameters:* $\delta = 10day^{-1}mm^{-3}, \alpha = 0.02day^{-1}$, $p = 3day^{-1}$, $T_{max} = 1500mm^{-3}$, $\beta = 0.24day^{-1}$, $\gamma = 2.4day^{-1}$, $k = 0.0027mm^3day^{-1}$, $N = 10$, *and* $i = 1$.

Our results can be used to give a mathematically rigorous justification for the plausible approximation arguments employed by Perelson and Nelson [14] to show that combination drug therapy can be effective in clearing the virus. Currently, the main drugs are RT inhibitors and protease inhibitors, and in practice, cocktails of several of these drugs have been most successful. The first type inhibits the copying of viral RNA to DNA and results in unsuccessful infection of the T cell by the virus. The second type results in virus particles that are noninfectious. Following [14], the short-term behavior after infection is given by the following system describing uninfected and infected T cells, infectious virus V_I, and noninfectious virus V_{NI}:

$$\begin{aligned} \dot{T} &= f(T) - k(1-\eta_{RT})V_I T, \\ \dot{T}^* &= -\beta T^* + k(1-\eta_{RT})V_I T, \\ \dot{V}_I &= -\gamma V_I + N\beta(1-\eta_{PI})T^* - ikV_I T, \\ \dot{V}_{NI} &= -\gamma V_{NI} + N\beta\eta_{PI}T^*, \end{aligned} \tag{8}$$

where, again, $i = 0$ corresponds to the system treated in [14], and $i = 1$ takes account of the loss of a virus particle when it enters a target cell (whether or not the virus is able to convert its RNA to DNA and insert itself in the host genome). The "effectiveness" coefficients η_{RT} for RT inhibitor and η_{PI} for protease inhibitor are assumed to lie somewhere between zero, meaning totally ineffective, and one, which represents 100% effectiveness.

Of course, the primary focus of drug therapy is on the possibility of clearing the virus. Observing that the first three equations are decoupled from the last one and that this subsystem is essentially similar to (5), we can calculate the basic reproduction number R_0^c under combination therapy by linearizing about the disease-free state E_0

to obtain

$$R_0^c = \frac{k\bar{T}[N(1-\eta_{RT})(1-\eta_{PI}) - i]}{\gamma}. \tag{9}$$

Comparing this with (6), we see that, in essence, N has been reduced to $N(1-\eta_{RT})(1-\eta_{PI})$. As i is typically much smaller than N and can be neglected, we see that the two inhibitors act in concert to reduce R_0 in (6) by the factor $(1-\eta_{RT})(1-\eta_{PI})$. If $R_0^c < 1$, the virus is cleared.

COROLLARY 2.2. *If $R_0^c < 1$, then the disease-free steady state E_0 is globally attracting. If $R_0^c > 1$, then E_0 is unstable.*

Assuming that current treatment does not allow for HIV eradication in an individual, this result implies one of the following: The efficiency of drugs is never high enough to make $R_o^c < 1$, or model (8) is not appropriate to describe HIV dynamics in a treated individual. It is argued in the recent paper by Callaway and Perelson that the first explanation is not viable. The second is adopted instead, and modified models are proposed to bring reality and theory closer to each other; see [2] for details.

3. Proofs.

3.1. Boundedness and stability of the disease-free steady state. First we show that solutions of model (5) are bounded.

LEMMA 3.1. *The closed positive orthant is positively invariant for* (5) *and there exists $M > 0$ such that all solutions satisfy $T(t), T^*(t), V(t) < M$ for all large t.*

Proof. The positive invariance of the positive orthant is trivial; we sketch the ultimate boundedness argument. Since $\dot{T} < f(T)$, we see that $T(t) < \bar{T} + 1$ for all large t, say $t > t_0$. Let $S = \max_{T \geq 0} f(T)$. Adding the first two equations gives $\dot{T} + \dot{T}^* = f(T) - \beta T^* \leq S - \beta T^*$. Let $A > 0$ be such that $\beta A > S + 1$. Then, so long as $T(t) + T^*(t) \geq A + \bar{T} + 1$ and $t > t_0$, we have $\dot{T} + \dot{T}^* < -1$. Clearly, there must exist $t_1 > t_0$ such that $T(t) + T^*(t) < A + \bar{T} + 1$ for all $t > t_1$.

The asymptotic bound for $T^*(t)$, namely, $T(t)^* \leq A + \bar{T} + 1$, together with the differential inequality $\dot{V} \leq -\gamma V + N\beta[A + \bar{T} + 1]$, which holds for large t, leads immediately to the asymptotic bound $V(t) \leq \gamma^{-1} N\beta[A + \bar{T} + 1]$. ∎

Next we consider the local stability behavior of (5) at the disease-free steady state E_0.

LEMMA 3.2. *If $R_0 < 1$, then the disease-free state E_0 is a locally asymptotically stable steady state of system* (5)*; if $R_0 > 1$, then it is unstable.*

Proof. The Jacobian matrix of the vector field corresponding to system (5), evaluated at E_0, is

$$J_0 := \begin{pmatrix} f'(\bar{T}) & 0 & -k\bar{T} \\ 0 & -\beta & k\bar{T} \\ 0 & N\beta & -\gamma - ik\bar{T} \end{pmatrix}. \tag{10}$$

Here $f'(\bar{T}) < 0$ is an eigenvalue, and the remaining eigenvalues derive from the two-by-two lower right submatrix, whose trace is negative and determinant is $\beta\gamma[1 - R_0]$. The result follows immediately. ∎

We remark that the same result holds for (8) with drug therapy, where R_0^c replaces R_0.

The following result deals with the global stability behavior of the disease-free steady state E_0.

LEMMA 3.3. *If $R_0 < 1$, then all solutions approach the disease-free state E_0.*

Proof. On consideration of the competitive vector field given by (5) on the three faces of the positive orthant, we see that any nontrivial periodic orbit must lie entirely in the interior of the positive orthant. If P denotes such a nontrivial periodic orbit, then it follows that the smallest box B containing P whose sides are parallel to the coordinate planes must also lie interior to the positive orthant. We can express B as $B = [p, q]_K$, where K denotes the cone $K \equiv \{(T, T^*, V) : T, V \geq 0, T^* \leq 0\}$. Indeed, if X^P (respectively, X_P) denotes the maximum (respectively, minimum) of coordinate $X = T, T^*, V$ on the periodic orbit P, then $p = (T_P, T^{*P}, V_P)$ and $q = (T^P, T^*_P, V^P)$. By Proposition 4.3 of [16], B must contain a steady state of (5). However, E_0 is the only steady state and $E_0 \notin B$. We conclude that no nontrivial periodic orbit exists. By the Poincaré–Bendixson theory for three dimensional competitive systems and the local stability of E_0, all solutions must approach E_0 in the limit. ∎

The same result holds for (8), with R_0^c in place of R_0. The entirely similar argument uses the fact that an endemic steady state exists only when the disease-free state is unstable ($R_0^c > 1$).

3.2. Local stability of the disease steady state. The local stability of the disease steady state is discussed next.

LEMMA 3.4. *Let $R_0 > 1$ and $f'(T_e) \leq 0$; then the nontrivial steady state $E_e \in \mathrm{int}(R^3_+)$ is locally asymptotically stable for system (5), for $i = 0, 1$. If $R_0 > 1$ and $f = f_1$, then E_e is unstable with a two dimensional unstable manifold under each of the following conditions:*

(a) *$i = 0$ with T_{max} large enough and (19) holds.*

(b) *$i = 1$ with kT_{max} large (see (20)) and p large enough.*

Proof. A calculation shows that the Jacobian matrix of the vector field corresponding to system (5), evaluated at E_e, takes the following form:

$$J_1 := \begin{pmatrix} -a & 0 & -kT_e \\ kV_e & -\beta & kT_e \\ -ikV_e & N\beta & -c \end{pmatrix}, \tag{11}$$

where

$$a := -f'(T_e) + kV_e \quad \text{and} \quad c := \gamma + ikT_e. \tag{12}$$

The characteristic equation associated with J_1 is given by

$$\lambda^3 + (a + \beta + c)\lambda^2 + [a(\beta + \gamma) - ikT_e f'(T_e)]\lambda + k\beta\gamma V_e = 0, \tag{13}$$

where we have used the expressions (6), (7) to simplify the coefficient of first and zeroth order. If $f'(T_e) \leq 0$, then it is easy to see that all coefficients are positive.

To finish the proof by means of the Routh–Hurwitz criterion, we need to show that

$$\Delta \equiv (a + \beta + c)(a(\beta + \gamma) - ikT_e f'(T_e)) - k\beta\gamma V_e \tag{14}$$

is positive. Using (12), it follows that

$$\begin{aligned} \Delta &= (-f'(T_e) + kV_e + \beta + \gamma + ikT_e)[(-f'(T_e) + kV_e)(\beta + \gamma) - ikT_e f'(T_e)] - k\beta\gamma V_e \\ &= (\beta + \gamma)^2(kV_e - f'(T_e)) - (\beta + \gamma)ikT_e f'(T_e) + (\beta + \gamma)ikT_e(kV_e - f'(T_e)) \\ &\quad - (ikT_e)^2 f'(T_e) + (\beta + \gamma)(kV_e - f'(T_e))^2 - ikT_e f'(T_e)(kV_e - f'(T_e)) \\ &\quad - k\beta\gamma V_e. \end{aligned} \tag{15}$$

If $f'(T_e) \leq 0$, then all terms in (15) are nonnegative except the last. However, the very first term $(\beta+\gamma)^2(kV_e - f'(T_e))$ can be expanded, yielding a term $2\beta\gamma kV_e$, which exceeds the last term $-k\beta\gamma V_e$. This implies that Δ is positive.

Hereafter, we consider the case in which $f = f_1$. A calculation yields

$$a = \frac{\delta}{T_e} + \frac{pT_e}{T_{max}} > 0, \tag{16}$$

and thus the coefficients of the zero and second powers of λ in the characteristic polynomial are positive. Together with the claim (which is proved below) that the Jacobian matrix has a real eigenvalue which is strictly less than the real parts of any other eigenvalue, it follows that if the Jacobian is hyperbolic and unstable, then there can be only one eigenvalue with negative real part (in fact it is negative) and two with positive real part. Further, hyperbolicity can only fail by a pair of pure imaginary eigenvalues and one negative eigenvalue.

Proof of claim. We prove that the Jacobian matrix possesses a real eigenvalue which is strictly less than the real part of the other eigenvalues. This follows from an application of the Perron–Frobenius theorem. Recall that the Perron–Frobenius theorem holds for nonnegative matrices and states that these matrices possess a real eigenvalue which is nonnegative. In addition, the modulus of every eigenvalue is not larger than this real eigenvalue. Now notice that the linear transformation $(x, y, z) \to (x, -y, z)$ puts J_1 in the following form:

$$\tilde{J}_1 := \begin{pmatrix} -a & 0 & -kT_e \\ -kV_e & -\beta & -kT_e \\ -ikV_e & -N\beta & -c \end{pmatrix}. \tag{17}$$

Of course, the eigenvalues of J_1 and $\tilde{J}_1$ are the same. Finally, observe that $-\tilde{J}_1$ is a nonnegative matrix for which the Perron–Frobenius theorem holds. The claim then follows immediately since the eigenvalues of $-\tilde{J}_1$ are the opposites of the eigenvalues of J_1.

If $i = 0$ and $f = f_1$, then all coefficients of (13) are positive as noted above. Inserting (16) and the values of V_e, T_e into (15) leads to

$$\begin{aligned} \Delta &= (\beta+\gamma)^2 a + (\beta+\gamma)a^2 - k\beta\gamma V_e \\ &= m\left(\frac{p}{T_{max}}\right)^2 + n\frac{p}{T_{max}} + q, \end{aligned} \tag{18}$$

where

$$\begin{aligned} m &= \frac{(\beta+\gamma)\gamma^2}{(Nk)^2}, \\ n &= \frac{(\beta+\gamma)^2\gamma}{Nk} + 2\delta(\beta+\gamma) - \beta\gamma T_{max} + \frac{\beta\gamma^2}{Nk}, \\ q &= (\beta+\gamma)^2\frac{Nk\delta}{\gamma} + (\beta+\gamma)\frac{(Nk\delta)^2}{\gamma^2} - \beta\delta Nk + \beta\gamma\alpha. \end{aligned}$$

Clearly, $m > 0$ and, less obviously, $q > 0$ since the first term exceeds the third in absolute value. By choosing T_{max} large, we may make $n < 0$ and as large in absolute value as we desire. In particular, if $n < 0$ and $n^2 > 4om$, then the quadratic (18) in

p/T_{max} is negative for an interval of values of p/T_{max} centered on

$$\frac{p}{T_{max}} = \frac{-n}{2m}, \tag{19}$$

ensuring that $\Delta < 0$. It follows that E_e is hyperbolic and unstable with a two dimensional unstable manifold.

If $i = 1$ and $f = f_1$, then a straightforward calculation shows that the coefficient of λ in (13) is given by

$$a_2 \equiv \frac{\gamma p}{N-1}\left[\frac{\beta+\gamma}{kT_{max}} + \frac{2\gamma}{k(N-1)T_{max}} - 1\right] + \frac{\gamma\alpha}{N-1} + \frac{(\beta+\gamma)\delta(N-1)k}{\gamma},$$

which can be negative when the term in brackets is negative, provided that p is large enough. Fixing kT_{max} so large that

$$kT_{max} > \beta + \gamma + \frac{2\gamma}{N-1} \tag{20}$$

ensures that the term in brackets is negative. Then, provided that p is large enough, it follows that E_e is hyperbolic and unstable with a two dimensional unstable manifold. □

3.3. Disease persistence. We discuss persistence of the disease next.

LEMMA 3.5. *If $R_0 > 1$, then there exists $\epsilon > 0$, independent of initial conditions satisfying $T^*(0) + V(0) > 0$, such that $\liminf_{t\to\infty} X(t) > \epsilon$ for $X = T, T^*, V$.*

Proof. The result follows from an application of Theorem 4.6 in [19], with $X_1 = \text{int}(R^3_+)$ and $X_2 = \text{bd}(R^3_+)$. This choice is in accordance with the conditions stated in this theorem. Furthermore, note that by virtue of Lemma 3.1 there exists a compact set B in which all solutions of system (5) initiated in R^3_+ ultimately enter and remain forever after. The compactness condition $(C_{4.2})$ is easily verified for this set B. Denoting the omega limit set of the solution $x(t, x_0)$ of system (5) starting in $x_0 \in R^3_+$ by $\omega(x_0)$ (which exists by Lemma 3.1), we need to determine the following set:

$$\Omega_2 = \cup_{y\in Y_2}\omega(y), \quad \text{where } Y_2 = \{x_0 \in X_2 |\, x(t, x_0) \in X_2, \forall t > 0\}. \tag{21}$$

From the system equations (5) it follows that all solutions starting in $\text{bd}(R^3_+)$ but not on the T axis leave $\text{bd}(R^3_+)$ and that the T axis is an invariant set, implying that $Y_2 = \{(T, T^*, V)^T \in \text{bd}(R^3_+) |\ T^* = V = 0\}$. Furthermore, it is easy to see that $\Omega_2 = \{E_0\}$ as all solutions initiated on the T axis converge to E_0. Then E_0 is a covering of Ω_2, which is isolated (since E_0 is a hyperbolic steady state under the assumption of the theorem) and acyclic (because there is no nontrivial solution in $\text{bd}(R^3_+)$ which links E_0 to itself). Finally, if it is shown that E_0 is a weak repeller for X_1, the proof will be done.

By definition, E_0 is a weak repeller for X_1 if for every solution starting in $x_0 \in X_1$

$$\limsup_{t\to+\infty} d(x(t, x_0), E_0) > 0. \tag{22}$$

We claim that (22) is satisfied if the following holds:

$$W^s(E_0) \cap \text{int}(R^3_+) = \emptyset, \tag{23}$$

where $W^s(E_0)$ denotes the stable manifold of E_0. To see this, suppose that (22) does not hold for some solution $x(t, x_0)$ starting in $x_0 \in X_1$. In view of the fact that the closed positive orthant is positively invariant for system (5) (recall Lemma 3.1), it follows that $\liminf_{t\to+\infty} d(x(t, x_0), E_0) = \limsup_{t\to+\infty} d(x(t, x_0), E_0) = 0$ and thus that $\lim_{t\to+\infty} x(t, x_0) = E_0$, which is clearly impossible if (23) holds.

What remains to be shown is that (23) holds. To that end, recall that the Jacobian matrix of system (5) at E_0, given in (10), is unstable if $R_0 > 1$. In particular, J_0 possesses one eigenvalue with positive real part, which we denote as λ_+, and two eigenvalues with negative real part, $f'(\bar{T})$, and an eigenvalue which we denote as λ_-. (Note that λ_- may be equal to $f'(\bar{T})$.) We proceed by determining the location of $E^s(E_0)$, the stable eigenspace of E_0. Clearly $(1, 0, 0)^T$ is an eigenvector of J_0 associated to $f'(\bar{T})$. If $\lambda_- \neq f'(\bar{T})$, then the eigenvector associated to λ_- has the following structure: $(0, p_2, p_3)^T$, where p_2 and p_3 satisfy the eigenvector equation

$$\begin{pmatrix} -\beta & k\bar{T} \\ N\beta & -\gamma - ik\bar{T} \end{pmatrix} \begin{pmatrix} p_2 \\ p_3 \end{pmatrix} = \lambda_- \begin{pmatrix} p_2 \\ p_3 \end{pmatrix}. \tag{24}$$

If $\lambda_- = f'(\bar{T})$, then λ_- is a repeated eigenvalue, and an associated generalized eigenvector will possess the following structure: $(*, p_2, p_3)^T$, where the value of $*$ is irrelevant for what follows and p_2 and p_3 also satisfy (24).

We claim that in both cases (i.e., $\lambda_- \neq f'(\bar{T})$ and $\lambda_- = f'(\bar{T})$) the vector $(p_2, p_3)^T \notin R^2_+$. The matrix in (24) is an irreducible Metzler matrix. A Metzler matrix is a matrix with nonnegative off-diagonal entries. For the definition of an irreducible matrix, see [4]. Observe that adding a sufficiently large positive multiple of the identity matrix to the matrix in (24) results in a nonnegative irreducible matrix for which the Perron–Frobenius theorem [4] holds. Consequently, the matrix in (24) possesses a simple real eigenvalue which is larger than the real part of any other eigenvalue, also called the *dominant eigenvalue*. Clearly, the dominant eigenvalue here is λ_+. But the Perron–Frobenius theorem also implies that every eigenvector that is not associated with the dominant eigenvalue does not belong to the closed positive orthant. Applied here, this means that $(p_2, p_3) \notin R^2_+$. Consequently, $E_s(E_0) \cap \mathrm{int}(R^3_+) = \emptyset$, and therefore also $W^s(E_0) \cap \mathrm{int}(R^3_+) = \emptyset$, which concludes the proof. □

3.4. Oscillations. Lemma 3.4 provides sufficient conditions for the Jacobian at E_e to have two eigenvalues with positive real part and one negative eigenvalue. The dynamical consequences of this are described in the following result.

LEMMA 3.6. *If $R_0 > 1$, the omega limit set of a solution which is not initiated on the T axis either contains E_e or is a nontrivial periodic orbit. If $R_0 > 1$ and if the Jacobian matrix at E_e has two eigenvalues with positive real part and one negative eigenvalue, then there exists an orbitally asymptotically stable periodic orbit. Every solution except those with initial data on the one dimensional stable manifold of E_e or on the T axis approaches a nontrivial periodic orbit.*

Proof. For $R_0 > 1$ it follows from the persistence result in Lemma 3.5 that the omega limit set of a solution which is not initiated on the T axis cannot contain a point on the T axis. Since there is only one steady state E^e which does not belong to the T axis, the first statement of the theorem follows from the generalized Poincaré–Bendixson theorem for competitive systems in dimension 3.

The assertions regarding the existence of an orbitally asymptotically stable periodic orbit follow from Theorem 1.2 in [20] and the fact that nonlinearities in (5) are analytic. In order to apply that result, we take the domain for (5) to be the interior of the positive orthant, in which the only steady state is E_e. Lemmas 3.1 and 3.5

imply that the dissipativity hypothesis of Theorem 1.2 is satisfied. The negativity of the Jacobian determinant, also required for Theorem 1.2, follows from our hypotheses concerning the eigenvalues. The assertion that suitably restricted forward orbits approach a periodic orbit follows from Theorem 4.2 in [16]. That result is stated for systems which are competitive in the traditional sense and so it applies to (5) since it can be transformed to a system which is competitive in the traditional sense. See also the remarks following Theorem 4.2, where it is noted that the second hypothesis of Theorem 4.2 holds if the Jacobian matrix is irreducible. □

3.5. Global asymptotic stability of the disease steady state. Finally we provide sufficient conditions preventing oscillations and leading to a globally asymptotically stable disease steady state.

LEMMA 3.7. *Suppose that* $R_0 > 1$, $f'(T) < 0$ *for* $T \in [0, \bar{T}]$, *and denote* $0 < \alpha^* = -\max_{T\in[0,\bar{T}]} f'(T)$. *If* $i = 0$ *or if* $i = 1$ *and* $kf(0) - \min(\alpha^*, \beta)\beta < 0$, *then* E_e *is a globally asymptotically stable steady state for system* (5) *with respect to initial conditions not on the* T *axis.*

Proof. The proof is based on an extension of the Poincaré–Bendixson theorem for the class of three dimensional competitive systems [16] and a powerful theory of second compound equations to prove asymptotic orbital stability of periodic solutions; see [10] and references cited therein. We do not wish to repeat the details of a precise proof here, because many of the arguments are the same as in [10], where a global stability result for a related epidemiological model is proved. Instead we provide only a sketch of the proof and go into details only where our proof is different. Under the assumptions of this lemma, system (5) possesses a steady state $E_e \in \text{int}(R^3_+)$, which is unique in $\text{int}(R^3_+)$. Moreover, from the proof of Lemma 3.5 it follows that the omega limit sets of solutions not initiated on the T axis are in $\text{int}(R^3_+)$. We claim that the only possible omega limit sets of solutions of system (5) are E_e or nontrivial periodic orbits. Indeed, if an omega limit set of a solution does not possess E_e, then it cannot contain another steady state (E_e is the unique steady state in $\text{int}(R^3_+)$), and thus it must be a nontrivial periodic orbit according to the extension of the Poincaré–Bendixson theorem for competitive systems. On the other hand, if an omega limit set does contain E_e, it *is* $\{E_e\}$, because E_e is a locally asymptotically stable steady state of system (5) according to Lemma 3.4 (notice that the condition needed to apply this Lemma, $f'(T_e) \le 0$, is satisfied here because $T_e = \bar{T}/R_0 < \bar{T}$ and $f' < 0$ in $[0, \bar{T}]$ by assumption), which establishes the claim. Finally we will show below that if system (5) possesses a nontrivial periodic solution, then this solution must be asymptotically orbitally stable. This fact will imply that E_e is a globally asymptotically stable steady state of system (5) with respect to initial conditions not on the T axis, which concludes the proof of this theorem. A proof of this implication can be found in [10]. The argument is that if E_e would not be globally asymptotically stable, then there would have to be a nontrivial periodic solution in $\text{int}(R^3_+)$. But it can then be shown that the region of attraction of E_e would have nonempty intersection with the region of attraction of the periodic solution, a contradiction. We prove the following: If system (5) possesses a nontrivial period solution, then this solution is asymptotically orbitally stable. Denote the periodic solution by $p(t) \equiv (p_1(t), p_2(t), p_3(t))^T$ and suppose that its minimal period is $\omega > 0$. Recall that from the proof of Lemma 3.1

$$0 \le p_1(t) \le \bar{T} \qquad \forall t \in [0, \omega]. \tag{25}$$

To establish asymptotic orbital stability of a periodic solution, we resort to the so-called method of the second compound equation; see [10] and references cited therein.

The second compound equation is the following periodic linear system:

$$\dot{z} = \frac{\partial f^{[2]}}{\partial x}(p(t))z, \tag{26}$$

where $z = (z_1, z_2, z_3)^T$ and $\frac{\partial f^{[2]}}{\partial x}$ is derived from the Jacobian matrix of system (5) and defined as follows:

$$\begin{aligned}\frac{\partial f^{[2]}}{\partial x} &:= \begin{pmatrix} j_{11}+j_{22} & j_{23} & -j_{13} \\ j_{32} & j_{11}+j_{33} & j_{12} \\ -j_{31} & j_{21} & j_{22}+j_{33} \end{pmatrix} \\ &= \begin{pmatrix} f'(T)-\beta-kV & kT & kT \\ N\beta & f'(T)-\gamma-k(iT+V) & 0 \\ ikV & kV & -\beta-\gamma-ikT \end{pmatrix}, \end{aligned} \tag{27}$$

where j_{kl} is the (k, l)th entry of the Jacobian matrix associated with system (5). The importance of the second compound equation is that if system (26) is asymptotically stable, then the periodic solution $p(t)$ is asymptotically orbitally stable for system (5); see [10]. We will show that the function

$$V(z_1, z_2, z_3; p(t)) := \sup\left\{|z_1|, \frac{p_2(t)}{p_3(t)}(|z_2|+|z_3|)\right\} \tag{28}$$

is a Lyapunov function for system (26). This function is positive, but not differentiable everywhere. Fortunately, this lack of differentiability can be remedied by using the right derivative of V, denoted as $D_+V(t)$. We have

$$D_+(|z_1(t)|) \le -(-f'(p_1(t)) + \beta + kp_3(t)).|z_1(t)| + k\frac{p_1(t)p_3(t)}{p_2(t)}.\frac{p_2(t)}{p_3(t)}(|z_2(t)|+|z_3(t)|) \tag{29}$$

and

$$\begin{aligned} D_+\left(\frac{p_2(t)}{p_3(t)}(|z_2(t)|+|z_3(t)|)\right) &= \left(\frac{\dot{p}_2(t)}{p_2(t)} - \frac{\dot{p}_3(t)}{p_3(t)}\right).\frac{p_2(t)}{p_3(t)}(|z_2(t)|+|z_3(t)|) \\ &\quad + \frac{p_2(t)}{p_3(t)}D_+(|z_2(t)|+|z_3(t)|) \\ &\le \left(\frac{p_2(t)}{p_3(t)}(N\beta + ikp_3(t))\right).|z_1(t)| \\ &\quad + \left(\frac{\dot{p}_2(t)}{p_2(t)} - \frac{\dot{p}_3(t)}{p_3(t)} - \gamma - ikp_1(t)\right).\frac{p_2(t)}{p_3(t)}(|z_2(t)|+|z_3(t)|) \\ &\quad - \frac{p_2(t)}{p_3(t)}(-f'(p_1(t))|z_2(t)| + \beta|z_3(t)|) \\ &\le \left(\frac{p_2(t)}{p_3(t)}(N\beta + ikp_3(t))\right).|z_1(t)| \\ &\quad + \left(\frac{\dot{p}_2(t)}{p_2(t)} - \frac{\dot{p}_3(t)}{p_3(t)} - \gamma - ikp_1(t) \right. \\ &\qquad \left. - \min(\alpha^*, \beta)\right).\frac{p_2(t)}{p_3(t)}(|z_2(t)|+|z_3(t)|), \end{aligned}$$

where the last inequality was obtained using the definition of α^* and (25).

Defining the following functions,

$$g_1(t) = -(-f'(p_1(t)) + \beta + kp_3(t)) + k\frac{p_1(t)p_3(t)}{p_2(t)}$$

$$= -(-f'(p_1(t)) + kp_3(t)) + \frac{\dot{p}_2(t)}{p_2(t)}, \tag{30}$$

$$g_2(t) = \frac{p_2(t)}{p_3(t)}(N\beta + ikp_3(t)) + \frac{\dot{p}_2(t)}{p_2(t)} - \frac{\dot{p}_3(t)}{p_3(t)} - \gamma - ikp_1(t) - \min(\alpha^*, \beta)$$

$$= ikp_2(t) + \frac{\dot{p}_2(t)}{p_2(t)} - \min(\alpha^*, \beta), \tag{31}$$

where the second equalities in (30) and (31) stem from the fact that $p(t)$ satisfies the system equations (5), we obtain that

$$D_+V(t) \le \sup(g_1(t), g_2(t))V(t). \tag{32}$$

Using the definition of α^* and (25), it follows from (30) that $g_1(t) \le -\alpha^* + \dot{p}_2(t)/p_2(t)$, and thus that $g_1(t) \le g_2(t)$. Then (32) can be rewritten as

$$D_+V(t) \le g_2(t)V(t). \tag{33}$$

We claim that the following holds:

$$\int_0^\omega g_2(t)dt < 0. \tag{34}$$

If this is established, it will follow from (33) that V is a Lyapunov function for system (26), and this will conclude the proof of the theorem.

(a) When $i = 0$, (34) is immediate from (31).

(b) When $i = 1$, using the fact that $p(t)$ is a periodic solution of (5), we see that

$$\int_0^\omega \beta p_2(t)dt = \int_0^\omega kp_3(t)p_1(t)dt = \int_0^\omega f(p_1(t))dt \le f(0)\omega$$

because, by assumption, $f'(T) < 0$ for all $T \in [0, \bar{T}]$ and since (25) holds. Consequently,

$$\int_0^\omega g_2(t)dt = \int_0^\omega [kp_2(t) - \min(\alpha^*, \beta)]dt \le \left[k\frac{f(0)}{\beta} - \min(\alpha^*, \beta)\right]\omega, \tag{35}$$

and it follows, under the assumption that $kf(0) - \min(\alpha^*, \beta)\beta < 0$, that (34) holds as claimed. □

REFERENCES

[1] E. Beretta and Y. Kuang, *Modeling and analysis of a marine bacteriophage infection*, Math. Biosci., 149 (1998), pp. 57–76.

[2] D.S. Callaway and A.S. Perelson, *HIV-1 infection and low steady state viral loads*, Bull. Math. Biol., 64 (2002), pp. 29–64.

[3] R.V. Culshaw and S. Ruan, *A delay-differential equation model of HIV infection of CD4$^+$ T-cells*, Math. Biosci., 165 (2000), pp. 27–39.

[4] F.R. Gantmacher, *The Theory of Matrices*, Chelsea, New York, 1959.

[5] M.W. Hirsch, *Systems of differential equations which are competitive or cooperative.* I: *Limit sets*, SIAM J. Math. Anal., 13 (1982), pp. 167–179.
[6] M.W. Hirsch, *Systems of differential equations that are competitive or cooperative* II: *Convergence almost everywhere*, SIAM J. Math. Anal., 16 (1985), pp. 423–439.
[7] M.W. Hirsch, *Systems of differential equations which are competitive or cooperative* III: *Competing species*, Nonlinearity, 1 (1988), pp. 51–71.
[8] M.W. Hirsch, *Systems of differential equations that are competitive or cooperative.* IV: *Structural stability in three-dimensional systems*, SIAM J. Math. Anal., 21 (1990), pp. 1225–1234.
[9] S. Merrill, *Modeling the interaction of HIV with the cells of the immune system*, in Mathematical and Statistical Approaches to AIDS Epidemiology, Lecture Notes in Biomath. 83, Springer-Verlag, New York, 1989.
[10] M.Y. Li and J.S. Muldowney, *Global stability for the SEIR model in epidemiology*, Math. Biosci., 125 (1995), pp. 155–164.
[11] M.A. Nowak and C.R.M. Bangham, *Population dynamics of immune responses to persistent viruses*, Science, 272 (1996), pp. 74–79.
[12] M.A. Nowak and R.M. May, *Virus Dynamics*, Oxford University Press, New York, 2000.
[13] A.S. Perelson, A.U. Neumann, M. Markowitz, J.M. Leonard, and D.D. Ho, *HIV-1 dynamics in vivo: Virion clearance rate, infected cell life span, and viral generation time*, Science, 271 (1996), pp. 1582–1585.
[14] A.S. Perelson and P.W. Nelson, *Mathematical analysis of HIV-1 dynamics in vivo*, SIAM Rev., 41 (1999), pp. 3–44.
[15] A.S. Perelson, D.E. Kirschner, and R. de Boer, *Dynamics of HIV infection of $CD4^+$ T cells*, Math. Biosci., 114 (1993), pp. 81–125.
[16] H.L. Smith, *Monotone Dynamical Systems*, AMS, Providence, RI, 1995.
[17] H.L. Smith, *Periodic orbits of competitive and cooperative systems*, J. Differential Equations, 65 (1986), pp. 361–373.
[18] H.L. Smith, *Systems of ordinary differential equations which generate an order preserving flow. A survey of results*, SIAM Rev., 30 (1988), pp. 87–113.
[19] H.R. Thieme, *Persistence under relaxed point-dissipativity (with application to an endemic model)*, SIAM J. Math. Anal., 24 (1993), pp. 407–435.
[20] H.R. Zhu and H.L. Smith, *Stable periodic orbits for a class of three dimensional competitive system*, J. Differential Equations, 110 (1994), pp. 143–156.

SIAM J. APPL. MATH.
Vol. 63, No. 4, pp. 1328–1358

HIGH FREQUENCY BEHAVIOR OF THE FOCUSING NONLINEAR SCHRÖDINGER EQUATION WITH RANDOM INHOMOGENEITIES*

ALBERT FANNJIANG†, SHI JIN‡, AND GEORGE PAPANICOLAOU§

Abstract. We consider the effect of random inhomogeneities on the focusing singularity of the nonlinear Schrödinger equation in three dimensions, in the high frequency limit. After giving a phase space formulation of the high frequency limit using the Wigner distribution, we derive a nonlinear diffusion equation for the evolution of the wave energy density when random inhomogeneities are present. We show that this equation is linearly stable even in the case of a focusing nonlinearity, provided that it is not too strong. The linear stability condition is related to the variance identity for the nonlinear Schrödinger equation in an unexpected way. We carry out extensive numerical computations to get a better understanding of the interaction between the focusing nonlinearity and the randomness.

Key words. focusing NLS, semiclassical limit, Wigner transformation, random media, diffusion approximation

AMS subject classifications. 35Q55, 60H15

PII. S003613999935559X

1. Introduction.

The nonlinear Schrödinger equation (NLS)

$$i\frac{\partial\phi}{\partial t}+\frac{1}{2}\Delta\phi-\beta|\phi|^2\phi=0, \qquad \phi(0,\mathbf{x})=\phi_0(\mathbf{x}), \tag{1}$$

with $\mathbf{x}$ in three dimensions, arises as the subsonic limit of the Zakharov model of Langmuir equations in plasma physics [20] and in many other contexts. The NLS (1) is in dimensionless form, with β a parameter that measures the strength of the nonlinearity relative to wave dispersion. When $\beta<0$ the nonlinearity is focusing, and when $\beta>0$ it is defocusing. An important property of the NLS is that, in three dimensions, the solution in the focusing case may develop a singularity at some finite time. This result is based on the existence of two invariants with respect to time: the *mass*

$$M=\int_{R^3}|\phi(t,\mathbf{x})|^2\,d\mathbf{x} \tag{2}$$

and the *energy*

$$H=\int_{R^3}\left(\frac{1}{2}|\nabla\phi(t,\mathbf{x})|^2+\frac{1}{2}\beta|\phi(t,\mathbf{x})|^4\right)d\mathbf{x}, \tag{3}$$

*Received by the editors May 5, 1999; accepted for publication (in revised form) September 27, 2002; published electronically April 23, 2003.
http://www.siam.org/journals/siap/63-4/35559.html

†Department of Mathematics, University of California, Davis, CA 95616 (cafannjiang@ucdavis.edu). The research of this author was supported by NSF grants DMS-9600119, DMS-9707756, and DMS-9971322.

‡Department of Mathematics, University of Wisconsin-Madison, Madison, WI 53706 (jin@math.wisc.edu). The research of this author was supported by AFOSR grant F49620-92-J0098 and NSF grants DMS-9404157 and DMS-9704957.

§Department of Mathematics, Stanford University, Stanford, CA 94305 (papanico@math.stanford.edu). The research of this author was supported by AFOSR grant F49620-98-1-0211 and by NSF grant DMS-9622854.

together with the *variance identity*

$$\frac{d^2}{dt^2}\int_{R^3}|\mathbf{x}|^2|\phi(t,\mathbf{x})|^2\,d\mathbf{x} = 4H + \beta\int_{R^3}|\phi(t,\mathbf{x})|^4\,d\mathbf{x}. \tag{4}$$

In the focusing case $\beta < 0$ and with a negative energy $H < 0$, the solution cannot remain bounded for all time. More precisely, it follows from the variance identity and the uncertainty inequality that the L^2 norm of the gradient of the solution blows up in finite time [6]. Many other properties of the NLS can be found in [18].

The goal of this paper is to investigate the effect of random inhomogeneities on the focusing NLS in the *high frequency* regime. In the high frequency limit the dimensionless time and propagation distance are long compared to the scale of variation of the potential $V(\mathbf{x})$. To make this precise we introduce slow time and space variables $t \to t/\epsilon$, $\mathbf{x} \to \mathbf{x}/\epsilon$, with ϵ a small parameter, and the scaled wave function $\phi^\epsilon(t,\mathbf{x}) = \phi(t/\epsilon, \mathbf{x}/\epsilon)$, which satisfies the scaled Schrödinger equation

$$\begin{aligned} &i\epsilon\phi^\epsilon_t + \frac{\epsilon^2}{2}\Delta\phi^\epsilon - V(t,\mathbf{x})\phi^\epsilon = 0,\\ &V(t,\mathbf{x}) = \beta|\phi(t,\mathbf{x})|^2 + V_0(\mathbf{x}). \end{aligned} \tag{5}$$

In the absence of the regularizing effect of the random inhomogeneities, the initial value problem for the focusing NLS is, in this regime, catastrophically ill-posed, even if the original NLS does not blow up. The random inhomogeneities are modeled by a potential that is a zero mean, stationary random function with correlation length comparable to the wavelength and with small variance. It takes the form of $v_0(\mathbf{x}) = \sqrt{\epsilon}\,V_1(\frac{\mathbf{x}}{\epsilon})$ since the wavelength is of order ϵ. Here $V(t,\mathbf{x})$ is the slowly varying background, without the nonlinear part, and $V_1(\mathbf{y})$ is a mean zero, stationary random function with correlation length of order one. This scaling allows the random potential to interact fully with the waves. We shall also assume that the fluctuations are statistically homogeneous and isotropic so that

$$\langle V_1(\mathbf{x})V_1(\mathbf{y})\rangle = R(|\mathbf{x}-\mathbf{y}|), \tag{6}$$

where $\langle,\rangle$ denotes statistical averaging and $R(|\mathbf{x}|)$ is the covariance of random fluctuations. The power spectrum of the fluctuations is defined by

$$\hat{R}(\mathbf{k}) = \left(\frac{1}{2\pi}\right)^3\int e^{i\mathbf{k}\cdot\mathbf{x}}R(\mathbf{x})\,d\mathbf{x}. \tag{7}$$

When (6) holds, the fluctuations are isotropic and $\hat{R}$ is a function only of $|\mathbf{k}|$. Because of the statistical homogeneity, the Fourier transform of the random potential V_1 is a generalized random process with orthogonal increments

$$\langle\hat{V}_1(\mathbf{p})\hat{V}_1(\mathbf{q})\rangle = \hat{R}(\mathbf{p})\delta(\mathbf{p}+\mathbf{q}). \tag{8}$$

Using the Wigner phase space form of the Schrödinger equation, we derive a nonlinear mean field transport approximation, in the high frequency and weak fluctuation limit. When, moreover, the transport mean free path is small, this nonlinear phase space transport equation can be further approximated by a nonlinear degenerate diffusion equation (e.g., (76)). This is the main result of this paper, and it captures in a precise way the interaction between the focusing nonlinearity and the random medium,

in the high frequency limit. A linear stability analysis of this diffusion equation reveals in a simplified but physically clear way the form of the nonlinearity-randomness interaction. We find that the condition that the linearized diffusion equation be stable reduces to the *positivity of the right-hand side of the variance identity* (4) *of the NLS*, in the high frequency limit (e.g., (94)). This is a surprising result because it is precisely the opposite of this condition, a negative right-hand side for (4), which produces a focusing singularity in NLS (1). We see that this condition, or rather its opposite

$$4H + \beta \int_{R^3} |\phi(t, \mathbf{x})|^4 \, d\mathbf{x} > 0, \tag{9}$$

becomes a *stability* condition, in a well defined high frequency regime, provided that the focusing mechanism is regularized by random inhomogeneities.

The paper is organized as follows. In section 2 we briefly review the nonlinear high frequency limit in its usual form, and in section 3 we reconsider that limit in its phase space form, using the Wigner distribution. In section 4 the random initial data is discussed. In section 5 we introduce random inhomogeneities and describe the mean field transport approximation for the Wigner distribution. In section 6 we rewrite the nonlinear transport equation in parity form, introducing the odd and even parts of the Wigner distribution, and in section 7 we derive the diffusion approximation in the small mean free path limit. The linearized stability condition for this degenerate nonlinear diffusion equation is obtained in section 8.

In section 9 we introduce a numerical scheme for the mean field nonlinear transport equation and present the results of several numerical calculations. In section 10 we do the same for the degenerate nonlinear diffusion equation for the wave energy density. Our numerical results indicate that in the high frequency regime the random inhomogeneities slow down the propagation of wave energy, in both the linear and defocusing cases. In the focusing case, the randomness is able to interact fully with the focusing nonlinearity as long as the nonlinearity is not too strong. In the diffusive regime, the randomness interacts fully with the focusing or defocusing nonlinearity, in a diffusive way, provided that the stability condition (94) holds. We end with section 11, which contains a brief summary and conclusions.

2. Nonlinear high frequency limit. We briefly review the high frequency asymptotic analysis for solutions of (1) with oscillatory initial data. The potential has the nonlinear part from the NLS and a linear part that we may add since it does not affect the analysis as long as it does not depend on ϵ. In the usual high frequency approximation [11] we consider initial data of the form

$$\phi^\epsilon(0, \mathbf{x}) = e^{iS_0(\mathbf{x})/\epsilon} A_0(\mathbf{x}) \tag{10}$$

with a smooth real valued initial phase function $S_0(\mathbf{x})$ and a smooth compactly supported complex valued initial amplitude $A_0(\mathbf{x})$. We then look for an asymptotic solution of (5) in the same form as the initial data (10), with evolved phase and amplitude

$$\phi^\epsilon(t, \mathbf{x}) \sim e^{iS(t,\mathbf{x})/\epsilon} A(t, \mathbf{x}). \tag{11}$$

Inserting this form into (5) and equating powers of ϵ, we get approximate evolution equations for the phase and amplitude

$$S_t + \frac{1}{2}|\nabla S|^2 + V(t, \mathbf{x}) = 0, \qquad S(0, \mathbf{x}) = S_0(\mathbf{x}) \tag{12}$$

and

$$(|A|^2)_t + \nabla \cdot (|A|^2 \nabla S) = 0, \qquad |A(0,\mathbf{x})|^2 = |A_0(\mathbf{x})|^2. \tag{13}$$

The phase equation (12) is the *eiconal* and the amplitude equation (13) the *transport* equation. The terminology for the latter is standard in the high frequency approximation but should not be confused with the radiative transport equation that will be derived later. These equations can be rewritten using the high frequency dispersion relation ω of the Schrödinger equation

$$\omega(t,\mathbf{x},\mathbf{k}) = \frac{1}{2}|\mathbf{k}|^2 + V(t,\mathbf{x}). \tag{14}$$

The energy in the high frequency regime is obtained by using the ansatz (11) in the energy (3) so that for small ϵ

$$H \approx \int_{R^3} \left(\frac{1}{2}|\nabla S|^2 + \frac{\beta}{2}|A|^2 + V_0 \right) |A|^2 \, d\mathbf{x}. \tag{15}$$

The potential is $V(t,\mathbf{x}) = \beta|A(t,\mathbf{x})|^2 + V_0(\mathbf{x})$. Even when it does not depend on the amplitude $|A|$, in the linear case, the eiconal equation (12) is nonlinear, and its solution exists in general only up to some time t^* that depends on the initial phase $S_0(\mathbf{x})$ and $V_0(\mathbf{x})$. This solution can be constructed by the method of characteristics, and singularities form when these characteristics (rays) cross. The eiconal and the transport equations are decoupled in the linear case.

To see more clearly the form of the eiconal and transport equations in the NLS case, we let $\rho = |A|^2$, $\mathbf{u} = \nabla S$, take the gradient of (12), with only the nonlinear potential $V(t,\mathbf{x}) = \beta|A(t,\mathbf{x})|^2$, and rewrite this differentiated eiconal and (13) in conservation law form:

$$\rho_t + \nabla \cdot (\rho \mathbf{u}) = 0, \tag{16}$$

$$(\rho \mathbf{u})_t + \nabla \cdot (\rho \mathbf{u}\mathbf{u}) + \nabla p(\rho) = 0, \tag{17}$$

where

$$p(\rho) = \frac{\beta}{2}\rho^2. \tag{18}$$

Now the eiconal and transport equations are fully coupled. When $\beta > 0$, this system of conservation laws are the isentropic gas dynamics equations, with equation of state given by (18) (γ-law gas with $\gamma = 2$). It is hyperbolic, and the solution may become discontinuous at a finite time. The velocity $\mathbf{u}$ is irrotational since it is a gradient, so $\nabla \times \mathbf{u} = 0$. The eiconal equation (12) is the Bernoulli form of the momentum conservation law (17) for time dependent and irrotational flows. Another form of the momentum conservation law is

$$\rho(\mathbf{u}_t + \mathbf{u} \cdot \nabla \mathbf{u}) + \nabla p = 0, \tag{19}$$

and the conservation of energy is exactly as in (15), with $V_0 = 0$, which we rewrite in the fluid variables

$$\frac{\partial}{\partial t} H = \frac{\partial}{\partial t} \int_{R^3} \left(\frac{1}{2}\rho|\mathbf{u}|^2 + p \right) d\mathbf{x} = 0. \tag{20}$$

In the one-dimensional defocusing case, the nonlinear high frequency limit was analyzed in detail in [9, 10]. In the higher-dimensional defocusing case, mathematical results are available only for the more regular Schrödinger–Poisson high frequency equations [5, 14, 7, 21]. When $\beta < 0$, the system of conservation laws (16)–(17) has complex characteristics, and the initial value problem is catastrophically ill-posed. This is the case even if the original NLS does not have solutions that blow up, when the Hamiltonian $H > 0$, for example. The high frequency limit for the focusing NLS has been studied only in the one-dimensional case with analytical initial data [12].

3. The Wigner distribution. An essential step in deriving phase space transport equations from wave equations is the introduction of the Wigner distribution [19, 15]. We begin with a brief review of some basic facts and then give the phase space form of the high frequency limit.

For any smooth function ϕ, rapidly decaying at infinity, the Wigner distribution is defined by

$$W(\mathbf{x},\mathbf{k}) = \left(\frac{1}{2\pi}\right)^3 \int_{R^3} e^{i\mathbf{k}\cdot\mathbf{y}} \phi\left(\mathbf{x}-\frac{\mathbf{y}}{2}\right) \overline{\phi}\left(\mathbf{x}+\frac{\mathbf{y}}{2}\right) d\mathbf{y}, \tag{21}$$

where $\bar{\phi}$ is the complex conjugate of ϕ. The Wigner distribution is defined on phase space and has many important properties. It is real, and its $\mathbf{k}$-integral is the modulus square of the function ϕ,

$$\int_{R^3} W(\mathbf{x},\mathbf{k})\, d\mathbf{k} = |\phi(\mathbf{x})|^2, \tag{22}$$

so we may think of $W(\mathbf{x},\mathbf{k})$ as wave number–resolved mass density. This is not quite right though, because $W(\mathbf{x},\mathbf{k})$ is not always positive, but it does become positive in the high frequency limit. The energy flux is expressed through $W(\mathbf{x},\mathbf{k})$ by

$$\mathcal{F} = \frac{1}{2i}(\phi\nabla\overline{\phi} - \overline{\phi}\nabla\phi) = \int_{R^3} \mathbf{k} W(\mathbf{x},\mathbf{k})\, d\mathbf{k}, \tag{23}$$

and its second moment in $\mathbf{k}$ is

$$\int |\mathbf{k}|^2 W(\mathbf{x},\mathbf{k})\, d\mathbf{k} = |\nabla\phi(\mathbf{x})|^2. \tag{24}$$

The Wigner distribution possesses an important $\mathbf{x}$-to-$\mathbf{k}$ duality given by the alternative definition

$$W(\mathbf{x},\mathbf{k}) = \int e^{i\mathbf{p}\cdot\mathbf{x}} \hat{\phi}\left(-\mathbf{k}-\frac{\mathbf{p}}{2}\right) \overline{\hat{\phi}\left(-\mathbf{k}+\frac{\mathbf{p}}{2}\right)}\, d\mathbf{p}, \tag{25}$$

where $\hat{\phi}$ is the Fourier transform of ϕ,

$$\hat{\phi}(\mathbf{k}) = \frac{1}{(2\pi)^3} \int e^{i\mathbf{k}\cdot\mathbf{x}} \phi(\mathbf{x})\, d\mathbf{x}. \tag{26}$$

These properties make the Wigner distribution a good quantity for analyzing the evolution of wave energy in phase space.

Given a wave function of the form (11), that is, an inhomogeneous wave with phase $S(t,\mathbf{x})/\epsilon$, its scaled Wigner distribution has the weak limit

$$W^\epsilon(\mathbf{x},\mathbf{k}) = \frac{1}{\epsilon^3} W\left(\mathbf{x},\frac{\mathbf{k}}{\epsilon}\right) \to |A(\mathbf{x})|^2 \delta(\mathbf{k}-\nabla S(\mathbf{x})), \tag{27}$$

as a generalized function, as $\epsilon \to 0$. This suggests that the correct scaling for the high frequency limit is

$$W^\epsilon(t,\mathbf{x},\mathbf{k}) = \left(\frac{1}{2\pi}\right)^3 \int e^{i\mathbf{k}\cdot\mathbf{y}} \phi^\epsilon\left(t,\mathbf{x}-\frac{\epsilon\mathbf{y}}{2}\right)\overline{\phi^\epsilon\left(t,\mathbf{x}+\frac{\epsilon\mathbf{y}}{2}\right)}\,d\mathbf{y}, \tag{28}$$

where ϕ^ϵ satisfies (5). From (27) we conclude that, as $\epsilon \to 0$, the scaled Wigner distribution of the solution $\phi^\epsilon(t,\mathbf{x})$ of (5) with initial data (10) is given by

$$W(t,\mathbf{x},\mathbf{k}) = |A(t,\mathbf{x})|^2 \delta(\mathbf{k}-\nabla S(t,\mathbf{x})), \tag{29}$$

where $S(t,\mathbf{x})$ and $A(t,\mathbf{x})$ are solutions of the eiconal and transport equations (12) and (13), respectively.

We will now sketch the derivation of the high frequency approximation of the scaled Wigner distribution directly from the Schrödinger equation. Let us assume that the initial Wigner distribution $W_0^\epsilon(\mathbf{x},\mathbf{k})$ tends to a smooth function $W_0(\mathbf{x},\mathbf{k})$ that has compact support. Note that this is not the case with the Wigner function corresponding to $\phi^\epsilon(0,\mathbf{x})$ given by (10), but it may be the case for random initial wave functions. We explain this briefly in the next section. The evolution equation for $W^\epsilon(t,\mathbf{x},\mathbf{k})$ corresponding to the Schrödinger equation (5) is the *Wigner equation*

$$W_t^\epsilon + \mathbf{k}\cdot\nabla_{\mathbf{x}} W^\epsilon + \mathcal{L}^\epsilon W^\epsilon = 0. \tag{30}$$

Here the operator $\mathcal{L}^\epsilon$ is defined by

$$\mathcal{L}^\epsilon Z(\mathbf{x},\mathbf{k}) = i\int_{R^3} e^{-i\mathbf{p}\cdot\mathbf{x}}\hat{V}(\mathbf{p})\frac{1}{\epsilon}\left[Z\left(\mathbf{x},\mathbf{k}+\frac{\epsilon\mathbf{p}}{2}\right) - Z\left(\mathbf{x},\mathbf{k}-\frac{\epsilon\mathbf{p}}{2}\right)\right]d\mathbf{p} \tag{31}$$

on any smooth function Z in phase space. The Fourier transform of V is $\hat{V}$.

From (31) we can easily find the limit of the operator $\mathcal{L}^\epsilon$ as $\epsilon \to 0$, in the linear case where the potential V does not depend on the solution. For any smooth and decaying function $Z(\mathbf{x},\mathbf{k})$ we have

$$\mathcal{L}^\epsilon Z(\mathbf{x},\mathbf{k}) \to -\nabla_{\mathbf{x}} V\cdot\nabla_{\mathbf{k}} Z. \tag{32}$$

Thus, the limit Wigner equation is the *Liouville equation* in phase space,

$$W_t + \mathbf{k}\cdot\nabla_{\mathbf{x}} W - \nabla V\cdot\nabla_{\mathbf{k}} W = 0, \tag{33}$$

with the initial condition $W(0,\mathbf{x},\mathbf{k}) = W_0(\mathbf{x},\mathbf{k})$. When the initial Wigner distribution has the form

$$W_0(\mathbf{x},\mathbf{k}) = |A_0(\mathbf{x})|^2\delta(\mathbf{k}-\nabla S_0(\mathbf{x})), \tag{34}$$

then it is easy to see that, up to the time of singularity formation, the solution of (33) is given by

$$W(t,\mathbf{x},\mathbf{k}) = |A(t,\mathbf{x})|^2\delta(\mathbf{k}-\nabla S(t,\mathbf{x})), \tag{35}$$

where $S(t,\mathbf{x})$ and $A(t,\mathbf{x})$ are solutions of the eiconal and transport equations (12) and (13), respectively. If the $W_0(\mathbf{x},\mathbf{k})$ and $V(\mathbf{x})$ are smooth, so will be the solution of (32), in the linear case.

In the nonlinear case the potential depends on the solution. The Liouville, or *Liouville–Vlasov*, equation is a nonlinear partial differential equation since

$$V = \beta\rho(t,\mathbf{x}) + V_0(\mathbf{x}) \quad \text{and} \quad \rho = \int_{R^3} W\, d\mathbf{k}. \tag{36}$$

For the initial conditions (34) it is better to use the fluid variables

$$\rho(t,\mathbf{x}) = |A(t,\mathbf{x})|^2 \quad \text{and} \quad \rho(t,\mathbf{x})\mathbf{u}(t,\mathbf{x}) = \rho(t,\mathbf{x})\nabla S(t,\mathbf{x}) = \int_{R^3} \mathbf{k} W\, d\mathbf{k}, \tag{37}$$

which solve the conservation laws (16)–(17). In the defocusing case, up to the time of shock formation the solution to the Liouville–Vlasov equation is given by

$$W(t,\mathbf{x},\mathbf{k}) = \rho(t,\mathbf{x})\delta(\mathbf{k} - \mathbf{u}(t,\mathbf{x})). \tag{38}$$

We see, therefore, that from the Wigner distribution we can recover all the information about the high frequency approximation, when it makes sense. In addition, it provides flexibility to deal with initial data that are not of the form (34).

4. Random initial data. Let us consider initial wave functions of the form $\phi_0(\frac{\mathbf{x}}{\epsilon},\mathbf{x})$, where $\phi_0(\mathbf{y},\mathbf{x})$ is a stationary random field in $\mathbf{y}$ for each $\mathbf{x}$ with mean zero and covariance

$$\langle \phi_0(\mathbf{y}_1,\mathbf{x}_1)\bar{\phi}_0(\mathbf{y}_2,\mathbf{x}_2)\rangle = R_0(\mathbf{y}_1 - \mathbf{y}_2, \mathbf{x}_1, \mathbf{x}_2). \tag{39}$$

Then with

$$W_0^\epsilon(\mathbf{x},\mathbf{k}) = \frac{1}{(2\pi)^3}\int e^{i\mathbf{k}\cdot\mathbf{y}}\phi_0\left(\frac{\mathbf{x}}{\epsilon} - \frac{\mathbf{y}}{2}, \mathbf{x} - \frac{\epsilon\mathbf{y}}{2}\right)\bar{\phi}_0\left(\frac{\mathbf{x}}{\epsilon} + \frac{\mathbf{y}}{2}, \mathbf{x} + \frac{\epsilon\mathbf{y}}{2}\right) d\mathbf{y} \tag{40}$$

we have that

$$\langle W_0^\epsilon(\mathbf{x},\mathbf{k})\rangle \to \hat{R}_0(\mathbf{k},\mathbf{x},\mathbf{x}) \tag{41}$$

pointwise in $\mathbf{k}$ and $\mathbf{x}$. Here $\hat{R}_0(\mathbf{k},\mathbf{x},\mathbf{x})$ is the diagonal part of the power spectral density of R_0, that is, its Fourier transform in $\mathbf{y}$ with $\mathbf{x}_1 = \mathbf{x}_2 = \mathbf{x}$. We also have that for any test function $\psi(\mathbf{x},\mathbf{k})$

$$\int W_0^\epsilon(\mathbf{x},\mathbf{k})\psi(\mathbf{x},\mathbf{k})\, d\mathbf{x}\, d\mathbf{k} \to \int \hat{R}_0(\mathbf{k},\mathbf{x},\mathbf{x})\psi(\mathbf{x},\mathbf{k})\, d\mathbf{x}\, d\mathbf{k} \tag{42}$$

in probability as $\epsilon \to 0$. This means that W_0^ϵ converges to $\hat{R}_0$ weakly in probability. However, it does not converge in mean square; that is, the mean fluctuation $\langle ||W_0^\epsilon - \hat{R}_0||_{L^2}^2\rangle$ does not go to zero. This can be seen from the fact that $\langle ||W_0^\epsilon||_{L^2}^2\rangle$ does not tend to $||\hat{R}_0||_{L^2}^2$.

From the above example we see how smooth and compactly supported initial Wigner functions can arise. For linear waves in random media there are no additional complications when dealing with random initial data that are statistically independent from the medium. The situation is much more complicated in the case of nonlinear waves and essentially unexplored mathematically.

5. High frequency limit with random inhomogeneities. We now consider small random perturbations of the potential $V(t,\mathbf{x})$. It is well known that, in one space dimension, linear waves in a random medium get localized even when the random perturbations are small [16], so our analysis is restricted to three dimensions. The two-dimensional case is difficult because the mean field approximation that we use in three dimensions is most likely incorrect.

We consider the linear case first. We assume that the correlation length of the random perturbation is of the same order as the wavelength, so the potential has the form $V(t,\mathbf{x}) + \sqrt{\epsilon}\, V_1(\frac{\mathbf{x}}{\epsilon})$ since the wavelength is of order ϵ. Here $V(t,\mathbf{x})$ is the slowly varying background, without the nonlinear part, and $V_1(\mathbf{y})$ is a mean zero, stationary random function with correlation length of order one. This scaling allows the random potential to interact fully with the waves. We shall also assume that the fluctuations are statistically homogeneous and isotropic so that

$$\langle V_1(\mathbf{x}) V_1(\mathbf{y})\rangle = R(|\mathbf{x}-\mathbf{y}|), \tag{43}$$

where $\langle , \rangle$ denotes statistical averaging and $R(|\mathbf{x}|)$ is the covariance of random fluctuations. The power spectrum of the fluctuations is defined by

$$\hat{R}(\mathbf{k}) = \left(\frac{1}{2\pi}\right)^3 \int e^{i\mathbf{k}\cdot\mathbf{x}} R(\mathbf{x})\, d\mathbf{x}. \tag{44}$$

When (43) holds, the fluctuations are isotropic and $\hat{R}$ is a function only of $|\mathbf{k}|$. Because of the statistical homogeneity, the Fourier transform of the random potential V_1 is a generalized random process with orthogonal increments

$$\langle \hat{V}_1(\mathbf{p}) \hat{V}_1(\mathbf{q})\rangle = \hat{R}(\mathbf{p})\delta(\mathbf{p}+\mathbf{q}). \tag{45}$$

If the amplitude of these fluctuations is strong, then scattering will dominate and waves will be localized [4], at least in the linear case. This means that we cannot assume that the fluctuations in the random potential $V_1(\mathbf{y})$ are large. If the random fluctuations are too weak, they will not affect energy transport at all. In order that the scattering produced by the random potential and the influence of the slowly varying background affect energy transport in comparable ways, the fluctuations in the random potential must be of order $\sqrt{\epsilon}$. This makes the transport mean free time, the reciprocal of Σ below, of order one and independent of ϵ. The scaled equation (5) becomes

$$\begin{aligned} &i\epsilon \frac{\partial \phi^\epsilon}{\partial t} + \frac{\epsilon^2}{2}\Delta\phi^\epsilon - \left(V(t,\mathbf{x}) + \sqrt{\epsilon}\, V_1\left(\frac{\mathbf{x}}{\epsilon}\right)\right)\phi^\epsilon = 0, \\ &\phi^\epsilon(0,\mathbf{x}) = \phi_0\left(\frac{\mathbf{x}}{\epsilon},\mathbf{x}\right). \end{aligned} \tag{46}$$

To describe the passage from (46) to the transport equation in its simplest form, we will set $V(t,\mathbf{x}) = 0$. A smooth and ϵ independent potential $V(t,\mathbf{x})$ that is not zero will not change the scattering terms in the phase space transport equation. It will affect only the Liouville part in the linear case. Now (30) for W^ϵ has the form

$$\frac{\partial W^\epsilon}{\partial t} + \mathbf{k}\cdot\nabla_{\mathbf{x}} W^\epsilon + \frac{1}{\sqrt{\epsilon}}\mathcal{L}_{\frac{\mathbf{x}}{\epsilon}} W^\epsilon = 0, \tag{47}$$

where the operator $\mathcal{L}_{\frac{\mathbf{x}}{\epsilon}}$, a rescaled form of (31), is given by

$$\mathcal{L}_{\frac{\mathbf{x}}{\epsilon}} Z(\mathbf{x},\mathbf{k}) = i\int e^{-i\mathbf{p}\cdot\mathbf{x}/\epsilon}\hat{V}_1(\mathbf{p})\left(Z\left(\mathbf{x},\mathbf{k}+\frac{\mathbf{p}}{2}\right) - Z\left(\mathbf{x},\mathbf{k}-\frac{\mathbf{p}}{2}\right)\right) d\mathbf{p}. \tag{48}$$

The behavior of this operator as $\epsilon \to 0$ is very different from (32) when V_1 is slowly varying. We can find the correct results by a formal multiscale analysis (see [15]).

Let $\mathbf{y} = \mathbf{x}/\epsilon$ be a fast space variable (on the scale of the wavelength), and introduce an expansion of W^ϵ of the form

$$W^\epsilon(t,\mathbf{x},\mathbf{k}) = W(t,\mathbf{x},\mathbf{k}) + \epsilon^{1/2}W^{(1)}(t,\mathbf{x},\mathbf{y},\mathbf{k}) + \epsilon W^{(2)}(t,\mathbf{x},\mathbf{y},\mathbf{k}) + \cdots, \tag{49}$$

with $\mathbf{y} = \mathbf{x}/\epsilon$ on the right. We assume that the leading term does not depend on the fast scale and that the initial Wigner distribution $W^\epsilon(0,\mathbf{x},\mathbf{k})$ tends to a smooth function $W_0(\mathbf{x},\mathbf{k})$, which is decaying fast enough at infinity. Then the average of the Wigner distribution, $\langle W^\epsilon \rangle$ is close to W, which satisfies the transport equation

$$\begin{aligned} &\frac{\partial W}{\partial t} + \mathbf{k}\cdot\nabla_{\mathbf{x}}W - \nabla_{\mathbf{x}}V\cdot\nabla_{\mathbf{k}}W, = \overline{\mathcal{L}}W, \\ &W(0,\mathbf{x},\mathbf{k}) = W_0(\mathbf{x},\mathbf{k}), \end{aligned} \tag{50}$$

where we have inserted on the left the term due to the potential V in (36). The linear operator $\overline{\mathcal{L}}$ is given by

$$\overline{\mathcal{L}}W(\mathbf{x},\mathbf{k}) = 4\pi \int_{R^3} \hat{R}(\mathbf{p}-\mathbf{k})\delta(\mathbf{k}^2-\mathbf{p}^2)[W(\mathbf{x},\mathbf{p}) - W(\mathbf{x},\mathbf{k})]\, d\mathbf{p}. \tag{51}$$

The left-hand side of (50) has precisely the form (33) of the Liouville equation. The right-hand side is the linear transport operator with differential scattering cross section $\sigma(\mathbf{k},\mathbf{k}')$ given by

$$\sigma(\mathbf{k},\mathbf{p}) = 4\pi\hat{R}(\mathbf{p}-\mathbf{k})\delta(\mathbf{k}^2-\mathbf{p}^2), \tag{52}$$

and total scattering cross section $\Sigma(\mathbf{k})$ given by

$$\Sigma(\mathbf{k}) = 4\pi \int_{R^3} \hat{R}(\mathbf{k}-\mathbf{p})\delta(\mathbf{k}^2-\mathbf{p}^2)\, d\mathbf{p}. \tag{53}$$

Note also that the transport equation (50) has two important properties. First, the total energy

$$E(t) = \iint_{R^3\times R^3} W(t,\mathbf{x},\mathbf{k})\, d\mathbf{k}\, d\mathbf{x} \tag{54}$$

is conserved, and second, the positivity of the solution $W(t,\mathbf{x},\mathbf{k})$ is preserved; that is, if the initial Wigner distribution $W_0(\mathbf{x},\mathbf{k})$ is nonnegative, then $W(t,\mathbf{x},\mathbf{k}) \geq 0$ for $t > 0$.

The physical meaning of the transport approximation for the linear Schrödinger equation with random potential is as follows. The characteristic wavelength introduced by the initial data is comparable with the scale of the inhomogeneities of the random potential. When we observe the wave energy far from the source and after a long time, it appears to evolve in phase space according to a radiative transport equation with a mean free path that is comparable to the distance from the source of the waves. This kind of behavior is captured with the ϵ scaling that we have introduced. The scaling of the size of the fluctuations by $\sqrt{\epsilon}$ is introduced so that the mean free path between macroscopic scatterings is comparable to the propagation distance.

The mathematical analysis of the passage from waves to transport in the linear case is considered in [2, 3, 8, 17]. The paper of Ho, Landau, and Wilkins [8] has

extensive references, and the paper of Erdös and Yau [3] gives a result that is global in time.

In the nonlinear case the potential $V(t,\mathbf{x}) = V^\epsilon(t,\mathbf{x})$ in (46) depends on the solution. In terms of the Wigner function, the potential is $\beta\rho^\epsilon(t,\mathbf{x}) + V_0(\mathbf{x})$ with $\rho^\epsilon = \int W^\epsilon\, d\mathbf{k}$. We will make a *mean field* hypothesis here, which says that in the transport limit $\epsilon \to 0$ the nonlinear potential keeps its form in the transport equation (50). This amounts to assuming that $\rho^\epsilon \to \rho$ in a strong sense. Some evidence for this is provided in the appendix. However, the mean field hypothesis is very difficult to prove. It is also difficult to test numerically, since the fact that we are in three dimensions is expected to play an important role. There are no mathematical results that deal with the mean field approximation.

5.1. A linear stability analysis. Let $\boldsymbol{\xi}$ be the unit vector in the direction of $\mathbf{k}$, i.e., $\mathbf{k} = k\boldsymbol{\xi}$, where $k = |\mathbf{k}|$. For simplicity, in the remainder of the paper we assume that the power spectral density of the inhomogeneities is

$$\hat{R} = \frac{\alpha}{2\pi}$$

with α a constant. Then (50)–(51) can be written as

$$\partial_t W + \mathbf{k}\cdot\nabla_{\mathbf{x}} W - \nabla_{\mathbf{x}} V\cdot\nabla_{\mathbf{k}} W = \alpha \int_{|\boldsymbol{\xi}'|=1} W(t,\mathbf{x},k,\boldsymbol{\xi}')\, d\boldsymbol{\xi}' - 4\pi\alpha W, \tag{55}$$

with

$$V = \beta\rho \quad \text{and} \quad \rho = \int_{R^3} W\, d\mathbf{k}. \tag{56}$$

The initial condition (34) is now rewritten as

$$W(0,\mathbf{x},k,\boldsymbol{\xi}) = \frac{1}{4\pi k^2}\delta(k - |\nabla S_0(\mathbf{x})|)\delta\left(\boldsymbol{\xi} - \frac{\nabla S_0(\mathbf{x})}{|\nabla S_0(\mathbf{x})|}\right)|A_0(\mathbf{x})|^2. \tag{57}$$

In order to carry out a linear stability analysis we first take the first two moments of (55). Multiplying (55) by 1 and $\mathbf{k}$, respectively, and integrating over $\mathbf{k}$, we have

$$\partial_t\rho + \nabla_{\mathbf{x}}\cdot\rho\mathbf{u} = 0, \tag{58}$$

$$\partial_t\rho\mathbf{u} + \nabla_{\mathbf{x}}\cdot\int \mathbf{k}\mathbf{k}W\, d\mathbf{k} - \int \mathbf{k}\nabla_{\mathbf{x}} V\cdot\nabla_{\mathbf{k}} W\, d\mathbf{k} = -4\pi\alpha\rho\mathbf{u}. \tag{59}$$

Thus the random inhomogeneity contributes a damping effect. Of course, these equations are not closed since high moments are undefined. However, they can be used in the linear stability analysis.

First we note that

$$\rho_0 \equiv \bar{\varrho}, \quad u_0 \equiv 0, \qquad W_0 \equiv \delta(\mathbf{k})\bar{\varrho},$$

with $\bar{\varrho}$ a constant, is a solution of the moment equations (58)–(59). We look for a solution near these constant states, in the form

$$\rho = \overline{\rho} + \rho^{(1)}, \qquad \mathbf{u} = \mathbf{u}^{(1)}, \tag{60}$$

where

$$\rho^{(1)} \ll \overline{\rho}, \qquad |\mathbf{u}^{(1)}| \ll 1. \tag{61}$$

We also set

$$W = \delta(\mathbf{k} - \mathbf{u})\rho. \tag{62}$$

With this ansatz, the moment equations can be closed to give

$$\partial_t \rho^{(1)} + \nabla_{\mathbf{x}} \cdot \rho \mathbf{u}^{(1)} = 0, \tag{63}$$

$$\partial_t \rho \mathbf{u}^{(1)} + \nabla_{\mathbf{x}} \cdot \mathbf{u}^{(1)} \mathbf{u}^{(1)} \rho + \beta \rho \nabla_{\mathbf{x}} \rho = -4\pi\alpha\rho\mathbf{u}^{(1)}. \tag{64}$$

Using (61) and ignoring higher order terms, we obtain the leading order equations

$$\partial_t \rho^{(1)} + \overline{\rho} \nabla_{\mathbf{x}} \cdot \mathbf{u}^{(1)} = 0, \tag{65}$$

$$\partial_t \mathbf{u}^{(1)} + \beta \nabla_{\mathbf{x}} \rho^{(1)} = -4\pi\alpha\mathbf{u}^{(1)}. \tag{66}$$

This system is hyperbolic if $\beta \geq 0$ and so is stable in the linear and defocusing cases. However, in the focusing case $\beta < 0$, the system is elliptic. A dispersion relation analysis for (65)–(66) shows that there are three negative eigenvalues

$$-1, \quad -1, \quad -2\pi\alpha - \sqrt{4\pi^2\alpha^2 - \beta\overline{\rho}|\boldsymbol{\eta}|^2},$$

where $\boldsymbol{\eta}$ is the wave number, and a fourth one

$$-2\pi\alpha + \sqrt{4\pi^2\alpha^2 - \beta\overline{\rho}|\boldsymbol{\eta}|^2}.$$

This last one is always negative when $\beta > 0$, and zero when $\beta = 0$, suggesting stability in the defocusing and linear cases. It is always negative when $\beta < 0$. The focusing case near uniform solutions with $\mathbf{u} = 0$ is, therefore, linearly unstable.

This means that the only hope for linear stability in the focusing case is to have a nonzero $\mathbf{u}$. This is consistent with the linear stability of the diffusion approximation in the focusing case, which will be derived next.

6. The parity formulation. It is convenient to use the parity formulation of the transport equation (50). This allows us to obtain the diffusion approximation in a transparent way.

To get the parity form of (55), we split it into two equations, one for $\mathbf{k}$ and one for $-\mathbf{k}$:

$$\begin{aligned}\partial_t W(\mathbf{k}) + \mathbf{k} \cdot \nabla_{\mathbf{x}} W(\mathbf{k}) - \nabla_{\mathbf{x}} V \cdot \nabla_{\mathbf{k}} W(\mathbf{k}) \\ = \alpha \int_{|\boldsymbol{\xi}'|=1} W(t, \mathbf{x}, k, \boldsymbol{\xi}')\, d\boldsymbol{\xi}' - 4\pi\alpha W(\mathbf{k}),\end{aligned} \tag{67}$$

$$\begin{aligned}\partial_t W(-\mathbf{k}) - \mathbf{k} \cdot \nabla_{\mathbf{x}} W(-\mathbf{k}) + \nabla_{\mathbf{x}} V \cdot \nabla_{\mathbf{k}} W(-\mathbf{k}) \\ = \alpha \int_{|\boldsymbol{\xi}'|=1} W(t, \mathbf{x}, k, \boldsymbol{\xi}')\, d\boldsymbol{\xi}' - 4\pi\alpha W(-\mathbf{k}).\end{aligned} \tag{68}$$

Define the even and odd parities as

$$W^+ = \frac{1}{2}[W(t, \mathbf{x}, \mathbf{k}) + W(t, \mathbf{x}, -\mathbf{k})],$$

$$W^- = \frac{1}{2}[W(t, \mathbf{x}, \mathbf{k}) - W(t, \mathbf{x}, -\mathbf{k})].$$

Adding and subtracting (67) and (68) gives the parity form of the transport equation

$$\partial_t W^+ + \mathbf{k}\cdot\nabla_{\mathbf{x}} W^- - \nabla_{\mathbf{x}} V\cdot\nabla_{\mathbf{k}} W^- = \alpha \int_{|\boldsymbol{\xi}'|=1} W^+(t,\mathbf{x},k,\boldsymbol{\xi}')\,d\boldsymbol{\xi}' - 4\pi\alpha W^+, \tag{69}$$

$$\partial_t W^- + \mathbf{k}\cdot\nabla_{\mathbf{x}} W^+ - \nabla_{\mathbf{x}} V\cdot\nabla_{\mathbf{k}} W^+ = -4\pi\alpha W^-. \tag{70}$$

The parity formulation has the advantage that the diffusion approximation can be derived easily, as will be shown in the next section.

7. Nonlinear diffusion limit. The diffusion approximation is obtained from the parity equations (69) and (70) in the small mean free time limit $1/\alpha \to 0$, and with the time stretched $t \to \alpha t$. Then (69) implies that for α large

$$W^+(t,\mathbf{x},\mathbf{k}) = \frac{1}{4\pi}\int_{|\boldsymbol{\xi}'|=1} W^+(t,\mathbf{x},k,\boldsymbol{\xi}')\,d\boldsymbol{\xi}' \equiv W_0(t,\mathbf{x},k), \tag{71}$$

and so the leading term of W^+ is independent of $\boldsymbol{\xi}$. From (70) we have that

$$W^- = -\frac{1}{4\pi\alpha}(\mathbf{k}\cdot\nabla_{\mathbf{x}} W_0 - \beta\nabla_{\mathbf{x}}\rho_0\cdot\nabla_{\mathbf{k}} W_0), \tag{72}$$

where

$$\rho_0(t,\mathbf{x}) = \int k^2 W_0\,dk$$

and where the time derivative can be neglected on the long time scale. Using (72) in (69), we get

$$\begin{aligned}\partial_t W_0 - \mathbf{k}\cdot\nabla_{\mathbf{x}}\frac{1}{4\pi\alpha}&(\mathbf{k}\cdot\nabla_{\mathbf{x}} W_0 - \beta\nabla_{\mathbf{x}}\rho_0\cdot\nabla_{\mathbf{k}} W_0)\\ &+ \beta\nabla_{\mathbf{x}}\rho_0\cdot\nabla_{\mathbf{k}}\frac{1}{4\pi\alpha}(\mathbf{k}\cdot\nabla_{\mathbf{x}} W_0 - \beta\nabla_{\mathbf{x}}\rho_0\cdot\nabla_{\mathbf{k}} W_0) = 0.\end{aligned} \tag{73}$$

Taking the $\boldsymbol{\xi}$ average $\frac{1}{4\pi}\int_{|\boldsymbol{\xi}|=1}\cdot\,d\boldsymbol{\xi}$ in (73) yields

$$\begin{aligned}\frac{\partial W_0}{\partial t} &- \frac{k^2}{4\pi}\frac{1}{4\pi\alpha}\int_{|\boldsymbol{\xi}|=1}\boldsymbol{\xi}\cdot\nabla_{\mathbf{x}}(\boldsymbol{\xi}\cdot\nabla_{\mathbf{x}} W_0)\,d\boldsymbol{\xi} + \beta\frac{1}{4\pi}\frac{k}{4\pi\alpha}\int_{|\boldsymbol{\xi}|=1}\boldsymbol{\xi}\cdot\nabla_{\mathbf{x}}(\nabla_{\mathbf{x}}\rho_0\cdot\nabla_{\mathbf{k}} W_0)\,d\boldsymbol{\xi}\\ &+ \beta\frac{1}{4\pi\alpha}\nabla_{\mathbf{x}}\rho_0\cdot\frac{1}{4\pi}\int_{|\boldsymbol{\xi}|=1}\nabla_{\mathbf{k}}(\mathbf{k}\cdot\nabla_{\mathbf{x}} W_0)\,d\boldsymbol{\xi}\\ &- \frac{\beta^2}{4\pi}\nabla_{\mathbf{x}}\rho_0\cdot\frac{1}{4\pi\alpha}\int_{|\boldsymbol{\xi}|=1}\nabla_{\mathbf{k}}(\nabla_{\mathbf{x}}\rho_0\cdot\nabla_{\mathbf{k}} W_0)\,d\boldsymbol{\xi} = 0.\end{aligned} \tag{74}$$

By straightforward manipulations, (74) becomes

$$\begin{aligned}\alpha\frac{\partial W_0}{\partial t} &- \frac{1}{12\pi}k^2\Delta W_0 + \beta\frac{k}{12\pi}\nabla_{\mathbf{x}}\cdot\left(\frac{\partial W_0}{\partial k}\nabla_{\mathbf{x}}\rho_0\right)\\ &+ \beta\frac{1}{12\pi}\nabla_{\mathbf{x}}\rho_0\cdot\left[\nabla_{\mathbf{x}}\frac{\partial}{\partial k}(kW_0) + 2\nabla_{\mathbf{x}} W_0\right]\\ &- \frac{\beta^2}{12\pi}|\nabla_{\mathbf{x}}\rho_0|^2\left[\frac{\partial^2 W_0}{\partial k^2} + \frac{2}{k}\frac{\partial W_0}{\partial k}\right] = 0.\end{aligned} \tag{75}$$

This is the diffusion approximation of the transport equation (55) and can also be written in the form

$$\alpha\frac{\partial W_0}{\partial t} - \frac{1}{12\pi}\begin{pmatrix} \nabla_{\mathbf{x}} \\ \frac{\partial}{\partial k} \end{pmatrix}\cdot\begin{pmatrix} k^2 I_3 & -\beta k\nabla_{\mathbf{x}}\rho_0 \\ -\beta k\,(\nabla_{\mathbf{x}}\rho_0)^T & \beta^2|\nabla_{\mathbf{x}}\rho_0|^2 \end{pmatrix}\begin{pmatrix} \nabla_{\mathbf{x}} W_0 \\ \frac{\partial W_0}{\partial k} \end{pmatrix}$$
$$+\beta\frac{1}{6\pi}\nabla_{\mathbf{x}}\rho_0\cdot\nabla_{\mathbf{x}}W_0 - \frac{\beta^2}{6\pi k}|\nabla_{\mathbf{x}}\rho_0|^2\frac{\partial W_0}{\partial k} = 0. \tag{76}$$

Here I_3 is the 3×3 identity matrix. The diffusion coefficient matrix in (76) is

$$D = \frac{1}{12\pi}\begin{pmatrix} k^2 I_3 & -\beta k\nabla_{\mathbf{x}}\rho_0 \\ -\beta k(\nabla_{\mathbf{x}}\rho_0)^T & \beta^2|\nabla_{\mathbf{x}}\rho_0|^2 \end{pmatrix}, \tag{77}$$

and it is symmetric and nonnegative semidefinite. However, since $\det D = 0$, the diffusion matrix is degenerate. This is because the scattering operator in (51) is concentrated on the unit sphere. The derivation of the nonlinear diffusion equation (76) is the main result of this paper.

The diffusion equation (75) or (76) can be rewritten into a very simple form (see [1]). Let

$$e = \frac{k^2}{2}, \qquad \tilde{\nabla} = \nabla_{\mathbf{x}} - \beta\nabla_{\mathbf{x}}\rho\frac{\partial}{\partial e}; \tag{78}$$

then (75) becomes

$$12\pi\alpha\frac{\partial W_0}{\partial t} - \tilde{\nabla}\cdot(2e\tilde{\nabla}W_0) + \beta\nabla_{\mathbf{x}}\rho_0\cdot\tilde{\nabla}W_0 = 0. \tag{79}$$

We can get equations for moments of W_0 which, however, are not closed. First we multiply (75) by k^2, integrate over k, and then integrate by parts to get

$$12\pi\alpha\frac{\partial\rho_0}{\partial t} - 3\beta\nabla_{\mathbf{x}}\cdot(\rho_0\nabla_{\mathbf{x}}\rho_0) = -\Delta\int k^4 W_0\,dk. \tag{80}$$

This gives mass conservation

$$\frac{\partial}{\partial t}\int\rho_0(t,\mathbf{x})\,d\mathbf{x} = 0.$$

Let u_0 be given by

$$u_0 = \frac{1}{\rho_0}\int k^3 W_0(t,\mathbf{x}\,k)\,dk. \tag{81}$$

To get an equation for the second moment, we multiply (75) by k^3 and integrate over k. After integrating by parts, we obtain

$$\frac{\partial\rho_0 u_0}{\partial t} - \Delta\int k^5 W_0\,dk - 4\beta\nabla_{\mathbf{x}}\cdot(\rho_0 u_0\nabla_{\mathbf{x}}\rho_0) - \beta\nabla_{\mathbf{x}}\rho_0\cdot\nabla_{\mathbf{x}}\rho_0 u_0$$
$$-2\beta^2|\nabla_{\mathbf{x}}\rho_0|^2\int kW_0\,dk = 0. \tag{82}$$

Note that (80) and (82) are not closed since they involve higher k moments of W_0.

As in the usual diffusion theory of the transport equation, an initial layer analysis gives the initial condition for W_0 as

$$W_0(0,\mathbf{x},k) = \frac{1}{k^2}\delta(k - |\nabla S_0(\mathbf{x})|)\,|A_0(\mathbf{x})|^2. \tag{83}$$

8. Linear stability condition for the nonlinear diffusion equation. In this section we carry out a linear stability analysis on the diffusion equation (75). This stability analysis gives a simplified but clear picture of how the nonlinear and random effects interact.

We use the moment equations (80) and (82) for the stability analysis. First we note that

$$\rho_0 \equiv \bar{\varrho}, \quad u_0 \equiv \overline{u}, \qquad W_0 \equiv \frac{1}{k^2}\delta(k-\overline{u})\bar{\varrho},$$

where $\bar{\varrho}, \overline{u}$ are constants, is a solution of the moment equations (80) and (82). We look for a solution near these constant states, in the form

$$\rho_0 = \overline{\rho} + \rho^{(1)}, \qquad u_0 = \overline{u} + u^{(1)}, \tag{84}$$

where

$$\rho^{(1)} \ll \overline{\rho}, \qquad u^{(1)} \ll \overline{u}.$$

We also set

$$W_0 = \frac{1}{k^2}\delta(k-u_0)\rho_0. \tag{85}$$

With this ansatz, the moment equations can be closed to give

$$12\pi\alpha\frac{\partial\rho_0}{\partial t} - \Delta\rho_0 u_0^2 - 3\beta\nabla_{\mathbf{x}}\cdot(\rho_0\nabla_{\mathbf{x}}\rho_0) = 0, \tag{86}$$

$$\begin{aligned} 12\pi\alpha\frac{\partial\rho_0 u_0}{\partial t} &- \Delta\rho_0 u_0^3 - 4\beta\nabla_{\mathbf{x}}\cdot(\rho_0 u_0\nabla_{\mathbf{x}}\rho_0) \\ &- \beta\nabla_{\mathbf{x}}\rho_0\cdot\nabla_{\mathbf{x}}\rho_0 u_0 - 2\beta^2|\nabla_{\mathbf{x}}\rho_0|^2\frac{\rho_0}{u_0} = 0. \end{aligned} \tag{87}$$

With the linearization (84), the last two terms in (87) are insignificant and so will be neglected. The moment equations thus become

$$12\pi\alpha\frac{\partial\rho_0}{\partial t} - \Delta\rho_0 u_0^2 - 3\beta\nabla_{\mathbf{x}}\cdot(\rho_0\nabla_{\mathbf{x}}\rho_0) = 0, \tag{88}$$

$$12\pi\alpha\frac{\partial\rho_0 u_0}{\partial t} - \Delta\rho_0 u_0^3 - 4\beta\nabla_{\mathbf{x}}\cdot(\rho_0 u_0\nabla_{\mathbf{x}}\rho_0) = 0. \tag{89}$$

We now do the linearization (84) and, keeping only the leading terms, obtain the coupled system of linear diffusion equations

$$12\pi\alpha\frac{\partial\rho_1}{\partial t} = (\overline{u}^2 + 3\beta\bar{\varrho})\Delta\rho_1 + 2\overline{\varrho u}\Delta u_1, \tag{90}$$

$$12\pi\alpha\frac{\partial u_1}{\partial t} = \beta\overline{u}\Delta\rho_1 + \overline{u}^2\Delta u_1. \tag{91}$$

8.1. Linear stability from the diffusion matrix. The diffusion coefficient matrix of (90) and (91),

$$A = \begin{pmatrix} \overline{u}^2 + 3\beta\bar{\varrho} & 2\overline{\varrho u} \\ \beta\overline{u} & \overline{u}^2 \end{pmatrix}, \tag{92}$$

has two eigenvalues:

$$\lambda_\pm = \overline{u}^2 + \frac{3}{2}\beta\bar{\varrho} \pm \frac{1}{2}\sqrt{8\beta\overline{\varrho u}^2 + 9\beta^2\overline{u}^2}. \tag{93}$$

Clearly, $Re(\lambda_\pm) > 0$ if and only if

$$\overline{u}^2 + \frac{3}{2}\beta\bar{\varrho} > 0. \tag{94}$$

This means that the right-hand side of the variance identity (4) in the high frequency limit, where H has the form (15) or (20) in the $\rho, \mathbf{u}$ variables, must be positive for stability. Therefore, even in the focusing case $\beta < 0$, the initial value problems for the linear diffusion equations (90) and (91) are well posed as long as (94) is satisfied.

It is surprising that the stability condition (94) does not depend on the strength α of the random inhomogeneities, although the diffusion rate does. It is also surprising that the right-hand side of the variance identity comes up as a *stability* condition, while in the analysis of the NLS equation it is used to get focusing solutions (instability) when (9) is negative.

8.2. Linear stability for the energy. We now study the stability of (90)–(91) in the energy norm. In the linear case when $\beta = 0$, it is obvious that (90)–(91) is stable. We analyze the defocusing and focusing cases separately.

In the defocusing case, $\beta > 0$, we multiply (90) by $\beta\rho$, and (91) by $\bar{\varrho}u$, and then add the resulting equations and integrate over $\mathbf{x}$. Upon integration by parts, we obtain

$$\begin{aligned}
12\pi\alpha\frac{\partial}{\partial t}\int\left(\frac{\beta}{2}\rho^2 + \frac{1}{2}\bar{\varrho}u^2\right)d\mathbf{x} &= -\int \beta(\overline{u}^2 + 3\beta\bar{\varrho})|\nabla\rho|^2 + 3\beta\overline{\varrho u}\nabla\rho\cdot\nabla u + \overline{\varrho u}^2|\nabla u|^2\,d\mathbf{x} \\
&= -\int \beta\left(\overline{u}^2 + \frac{3}{4}\beta\bar{\varrho}\right)|\nabla\rho|^2\,d\mathbf{x} \\
&\quad -\int \frac{9}{4}\beta^2\bar{\varrho}|\nabla\rho|^2 + 3\beta\overline{\varrho u}\nabla\rho\cdot\nabla u + \overline{\varrho u}^2|\nabla u|^2\,d\mathbf{x} \\
&= -\int \beta\left(\overline{u}^2 + \frac{3}{4}\beta\bar{\varrho}\right)|\nabla\rho|^2\,d\mathbf{x} - \int \bar{\varrho}\left|\frac{3}{2}\beta\nabla\rho + \overline{u}\nabla u\right|^2 d\mathbf{x} \\
&\le 0. \qquad (95)
\end{aligned}$$

Thus, in the defocusing case the system (90)–(91) is always stable.

In the focusing case, $\beta < 0$, we again multiply (90) by $\beta\rho$, and (91) by $2\bar{\varrho}u$. We then *subtract* the first equation from the second one. By integrating over $\mathbf{x}$ and integrating by parts, we obtain

$$12\pi\alpha\frac{\partial}{\partial t}\int\left(\bar{\varrho}u^2 - \frac{\beta}{2}\rho^2\right)d\mathbf{x} = -2\overline{\varrho u}^2\int|\nabla u|^2\,d\mathbf{x} + \beta\int(\overline{u}^2 + 3\beta\bar{\varrho})|\nabla\rho|^2\,d\mathbf{x}. \tag{96}$$

Clearly, a sufficient condition for the above term to be nonpositive is

$$\overline{u}^2 + 3\beta\bar{\varrho} \ge 0. \tag{97}$$

This means that the coefficient of $\Delta\rho$ in (90) should be nonnegative, which means that the diagonal entries of the matrix A should be nonnegative.

9. Numerical solution of the nonlinear transport equation. In this section we will present some numerical results for the nonlinear transport equation (55)–(56). In order to have good resolution, we assume spherical symmetry and use a second order nonoscillatory upwind scheme [13]. The solution depends only on $r = |\mathbf{x}|$, k, and $\theta = \cos^{-1} \frac{\mathbf{x}\cdot\mathbf{k}}{rk}$, the angle between $\mathbf{x}$ and $\mathbf{k}$. In these variables, (55) and (56) become

$$\begin{aligned} &\frac{\partial W}{\partial t} + \mathrm{k}\cos\theta \frac{\partial W}{\partial r} - \frac{k}{r}\sin\theta \frac{\partial W}{\partial \theta} - \beta \frac{\partial V}{\partial r}\left(\cos\theta \frac{\partial W}{\partial k} - \frac{\sin\theta}{k}\frac{\partial W}{\partial \theta}\right) \\ &\qquad = 2\pi\alpha \int_0^\pi W \sin\theta \, d\theta - 4\pi\alpha W, \\ &V = V(\rho), \qquad \rho = \int_0^\infty \int_0^{2\pi} W k^2 \sin\theta \, dk \, d\theta. \end{aligned} \tag{98}$$

In terms of the direction cosine $-1 \le \mu = \cos\theta \le 1$, (98) can be rewritten in conservation form as

$$\begin{aligned} &\frac{\partial W}{\partial t} + \frac{\partial}{\partial r}(\mu k W) + \frac{\partial}{\partial k}\left(-\beta\mu \frac{\partial V}{\partial r} W\right) + \frac{\partial}{\partial \mu}\left[(1-\mu^2)\left(\frac{k}{r} - \frac{\beta}{k}\frac{\partial V}{\partial r}\right) W\right] \\ &\qquad = 2\pi\alpha \int_{-1}^1 W(t,r,k,\mu') \, d\mu' - 4\pi\alpha W, \\ &V = V(\rho), \qquad \rho = \int_0^\infty \int_{-1}^1 W k^2 \, dk \, d\mu, \end{aligned} \tag{99}$$

where $W = W(t, r, k, \mu)$. The initial condition for (99) is

$$W(0, r, k, \mu) = \frac{1}{k^2}\delta(k - u_0(r))\delta(\mu - a_0(r))\rho_0(r). \tag{100}$$

9.1. The numerical method. We will use a second order upwind scheme for spatial discretizations. This is a natural choice since (99) is a hyperbolic equation. We use the composite midpoint rule for angular integration, and the second order explicit Runge–Kutta method for time discretization. The overall accuracy is of second order.

Let $r_{i+1/2}$ $(0 \le i \le I)$ be the grid points in the r-direction, and likewise define $k_{j+1/2}$ $(0 \le j \le J)$ and $\mu_{l+1/2}$ $(0 \le l \le L)$. Let r_i, k_j, μ_l be the midpoints (for example, $r_i = \frac{1}{2}(r_{i+1/2} + r_{i-1/2})$). Let $\Delta r = r_{i+1/2} - r_{i-1/2}$, $\Delta k = k_{j+1/2} - k_{j-1/2}$, and $\Delta\mu = \mu_{l+1/2} - \mu_{l-1/2}$ be the uniform grid sizes in each direction. Let W_{ijl} be the numerical approximation of W at (r_i, k_j, μ_l), and $W_{i+1/2,j,l}$ be the approximation of W at $(r_{i+1/2}, s_j, \mu_l)$. We define $W_{i,j+1/2,l}$ and $W_{ij,l+1/2}$ in a similar way. A second order conservative approximation for (99) is

$$\begin{aligned} &\frac{\partial}{\partial t} W_{ijl} + \frac{\mu_l k_j}{\Delta r}(W_{i+1/2,j,l} - W_{i-1/2,j,l}) - \frac{\mu_l}{\Delta k}\frac{\partial V_i}{\partial r}(W_{i,j+1/2,l} - W_{i,j-1/2,l}) \\ &\qquad + \frac{1-\mu_l^2}{\Delta\mu}\left(\frac{k_j}{r_i} - \frac{1}{k_j}\frac{\partial V_i}{\partial r}\right)(W_{i,j,l+1/2} - W_{i,j,l-1/2}) \\ &\qquad = 2\pi\alpha \int_{-1}^1 W_{ij}(t,r,k,\mu') - 4\pi\alpha W_{ijl}. \end{aligned} \tag{101}$$

To get the flux $W_{i+1/2,jl}$ from the known quantity W_{ijl}, we use the second order

upwind scheme due to van Leer (see [13]):

$$W_{i+1/2,jl} = W_{ijl} + \frac{\Delta r}{2}\sigma_i^r \qquad \text{if} \quad \mu_l > 0, \tag{102}$$

$$W_{i+1/2,jl} = W_{i+1,jl} - \frac{\Delta r}{2}\sigma_{i+1}^r \qquad \text{if} \quad \mu_l < 0. \tag{103}$$

Here σ_i^r is the limited slope (see [13])

$$\sigma_i^r = \frac{1}{\Delta r}(W_{i+1,jl} - W_{ijl})\phi(\theta^r), \tag{104}$$

$$\theta^r = \frac{W_{ijl} - W_{i-1,jl}}{W_{i+1,jl} - W_{ijl}}, \tag{105}$$

$$\phi(\theta) = \frac{|\theta| + \theta}{1 + |\theta|}. \tag{106}$$

A limited slope scheme is necessary, especially for the defocusing case, since moments of the nonlinear transport equation are close to the gas dynamics equations. Without a limited slope the numerical solutions become oscillatory when shocks develop. We define the flux in the k-direction in a similar way:

$$W_{i,j+1/2,l} = W_{ijl} + \frac{\Delta k}{2}\sigma_j^k \qquad \text{if} \quad -\mu_l \frac{\partial V_i}{\partial r} > 0, \tag{107}$$

$$W_{i,j+1/2,l} = W_{i,j+1,l} - \frac{\Delta k}{2}\sigma_{j+1}^k \qquad \text{if} \quad -\mu_l \frac{\partial V_i}{\partial r} < 0, \tag{108}$$

where σ_i^k is defined as in (104). The flux in the μ-direction is given by

$$W_{i,j,l+1/2} = W_{ijl} + \frac{\Delta \mu}{2}\sigma_l^\mu \qquad \text{if} \quad \frac{k_j}{r_i} - \frac{1}{k_j}\frac{\partial V}{\partial r} > 0, \tag{109}$$

$$W_{i,j,l+1/2} = W_{i,j,l+1} - \frac{\Delta \mu}{2}\sigma_{l+1}^\mu \qquad \text{if} \quad \frac{k_j}{r_i} - \frac{1}{k_j}\frac{\partial V}{\partial r} < 0, \tag{110}$$

where σ_l^μ is defined as in (104).

To find $\frac{\partial V_i}{\partial r}$, given that $V'(r) = V'(\rho)\frac{\partial \rho}{\partial r}$, we need to evaluate $\frac{\partial \rho_i}{\partial r}$. We use the centered difference

$$\begin{aligned} \frac{\partial \rho_i}{\partial r} &\approx \frac{\rho_{i+1/2} - \rho_{i-1/2}}{\Delta r} \\ &= \frac{1}{\Delta r}\left(\int_0^\infty \int_{-1}^1 W_{i+1/2}(t,k,\mu)\, dk\, d\mu - \int_0^\infty \int_{-1}^1 W_{i-1/2}(t,k,\mu)\, dk\, d\mu\right) \\ &\approx \frac{\Delta k \Delta \mu}{\Delta r} \sum_{j,l} (W_{i+1/2,j,l} - W_{i-1/2,j,l}). \end{aligned} \tag{111}$$

Here we have used the composite midpoint rule to approximate the integral. It has second order accuracy in k and μ. We can use the flux $W_{i+1/2,j,l}$ already obtained from (103) in (111). To recover $V_{i+1/2}$, we use the integral

$$\rho_{i+1/2} = \rho_{1/2} + \int_0^{r_{i+1/2}} \frac{\partial \rho}{\partial r}\, dr \approx \rho_{1/2} + \sum_{j=0}^{i} \frac{\partial \rho_j}{\partial r}\Delta r, \tag{112}$$

where the composite midpoint rule has again been used.

Time discretization. We use the second order Runge–Kutta method.

Boundary conditions. Since the numerical flux has a five-point stencil, it is necessary to have two fictitious points outside the physical boundaries. To define W on the left of $r = 0$ we use the condition

$$W(t, -r, k, \mu) = W(t, r, k, -\mu). \tag{113}$$

This makes sense physically because (99) remains unchanged if we simultaneously replace r by $-r$ and μ by $-\mu$. To define W to the left of $k = 0$, we use a similar condition

$$W(t, r, -k, \mu) = W(t, r, k, -\mu). \tag{114}$$

The scattering term in (99) is small near $k = 0$. At outer boundaries of r and k we use outgoing boundary conditions. For the computations the domain in k is large enough so that at the outer boundary W is nearly zero. At $\mu = \pm 1$ we simply use the reflecting boundary condition $\frac{\partial W}{\partial \mu} = 0$.

More precisely, we fix the numerical boundary conditions at $r = 0$ as follows. If $\frac{\partial V}{\partial r} < 0$, we first extrapolate W to second order for $\mu < 0$ from the interior value of W to get $W_{-1/2,jl}$ and $W_{-2/3,jl}$, where $r_{-1/2} = -r_{1/2}$ and $r_{-3/2} = -r_{3/2}$ are the fictitious points. To obtain W at the fictitious points for $\mu > 0$, we use the condition (113) numerically, i.e., $W_{-1/2.jl} = W_{1/2,j,L-l+1}$ and $W_{-3/2,jl} = W_{3/2,j,L-l+1}$. If $\frac{\partial V}{\partial r} > 0$ we reverse this process. This corresponds to extrapolation in an upwind direction, which is necessary for a hyperbolic equation. At $k = 0$ we impose a similar boundary condition. At the outer boundaries we always assume zero incoming flux, and the outgoing flux is simply the outflow boundary condition.

9.2. Numerical results. We will now present the results of some numerical experiments for the nonlinear transport equation (99). We want to see the effect of randomness, which in (99) is the scattering term, on the focusing nonlinearity. We use the potential

$$V(\rho) = \beta\rho.$$

The initial energy density is

$$W(0, r, k, \mu) = \frac{1}{k^2}\delta(k-1)\delta(\mu)\rho_0(r), \tag{115}$$

but in the numerical computations we use the Gaussian

$$W(0, r, k, \mu) = \frac{\lambda^2}{\pi}\frac{1}{k^2}e^{-\lambda^2((k-1)^2+\mu^2)}\rho_0(r), \tag{116}$$

which gives

$$\rho_0(r) = e^{-\lambda^2 r^2} + 0.5. \tag{117}$$

This initial density regularizes the *delta*-function. The smoothing effect of the parameter λ is explored numerically below. The nonlinear transport equation with *delta*-function initial data is quite singular, and there is no theoretical justification for expecting it to correspond to the limiting behavior $\lambda \to \infty$ of the regularized solutions.

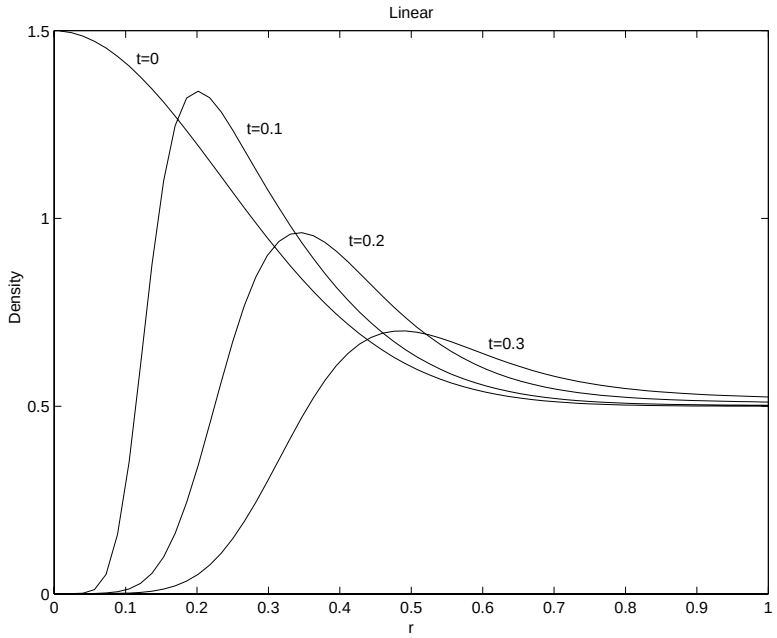


FIG. 1. *The density $\rho(t,r)$ versus r for the linear case without randomness;* ($\alpha = \beta = 0$) *and* $\lambda = 3$. *Note that the energy is propagating away from the origin.*

The computational domain in r, k, μ space is $[0,2] \times [0,2] \times [-1,1]$. We use 120 cells in r, 120 cells in k, and 40 cells in μ. We use $\Delta t = 10^{-4}$ for the linear case, $\Delta t = 5 \times 10^{-6}$ for the defocusing case, and $\Delta t = 10^{-5}$ for the focusing cases. In the focusing case the solution is less diffusive, and the CFL condition allows a slightly larger Δt than in the defocusing case. We first use $\lambda = 3$.

1. The linear case ($\beta = 0$). If there is no randomness ($\alpha = 0$), wave energy moves away from the origin, as in Figure 1. By turning on the random terms ($\alpha = 0.7$), we see that it tends to slow down the spreading of the energy, as shown in Figure 2.
2. The defocusing case ($\beta = 1$). The energy density without randomness ($\alpha = 0$) is shown in Figure 3. It propagates away from the origin much faster than in the linear case, Figure 1. In Figure 4 we show the energy density in the defocusing case with randomness ($\alpha = 0.7$). The solution is more spread out than in the linear case.
3. The focusing case. If there is no randomness ($\alpha = 0$), the numerical solution becomes highly oscillatory (one wave length per grid point; see Figure 5), reflecting the unstable nature of the problem. By turning on the randomness, for example, at $\alpha = 0.7$, numerical results are stable, at least when $|\beta|$ is not too large. We compare the results of $\beta = -0.2$ and $\beta = -0.5$ in Figures 6 and 7, respectively. In both cases the solutions spread out more slowly than in the linear case, and larger $|\beta|$ slows down the spreading of the solution. We also consider the smoothing effect of the initial data. In Figures 8 and 9 we compare numerical solutions with $\lambda = 6$ and $\lambda = 9$, respectively. The solutions are quite close to each other. The larger λ slows down spreading a bit, but the qualitative behavior of the numerical solutions remains the same.

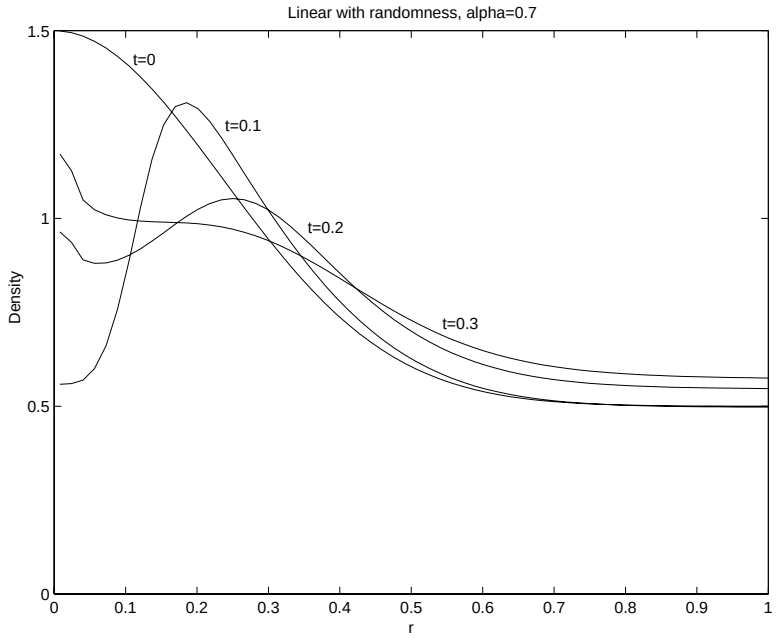


FIG. 2. *The density $\rho(t,r)$ versus r for the linear case with randomness $\alpha = 0.7$ and $\lambda = 3$. Wave energy spreading is slower than that of Figure 1 without randomness.*

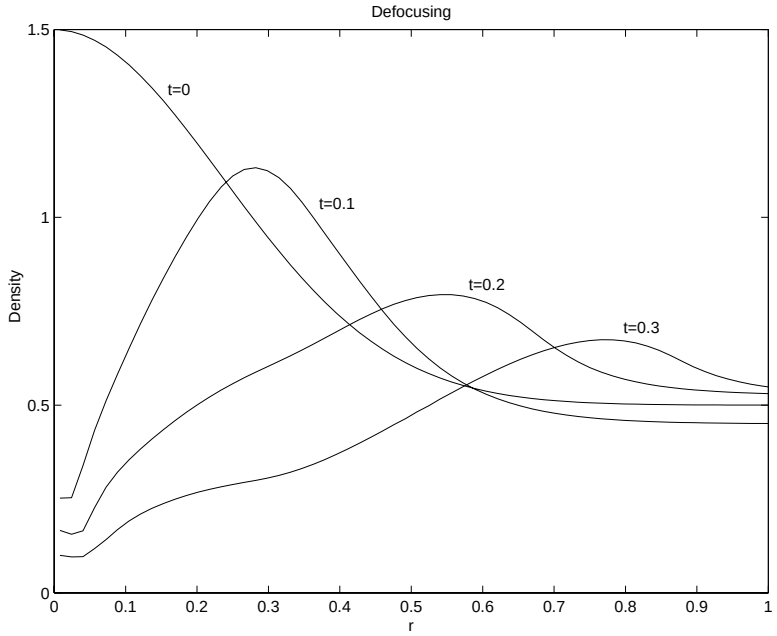


FIG. 3. *The density $\rho(t,r)$ versus r for the defocusing case with no randomness ($\alpha = 0$, $\beta = 1$) and $\lambda = 3$. The wave energy propagates away from the origin much faster than in the linear case.*

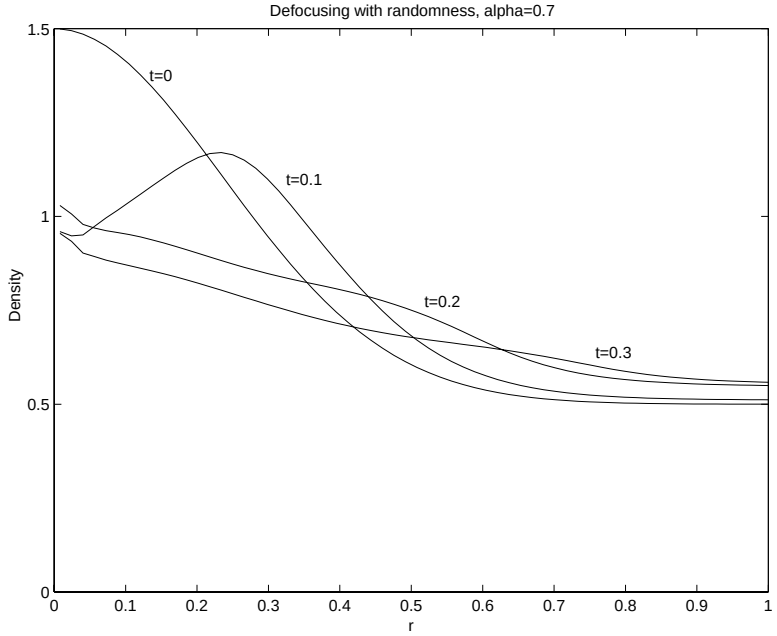


FIG. 4. *The energy density $\rho(t,r)$ versus r in the defocusing case with randomness ($\alpha = 0.7$, $\beta = 1$) and $\lambda = 3$. The density is more spread out.*

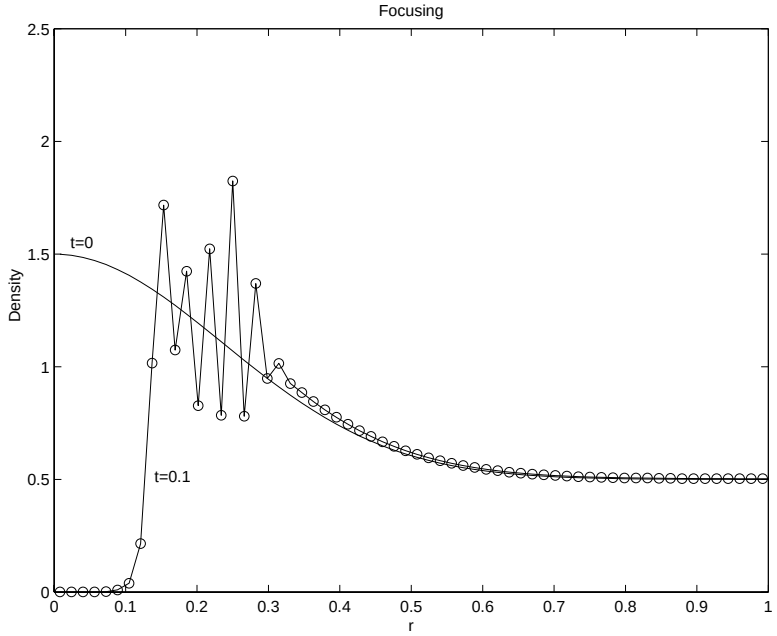


FIG. 5. *The energy density $\rho(t,r)$ versus r for the focusing case with no randomness. Here $\beta = -0.2$ and $\lambda = 3$.*

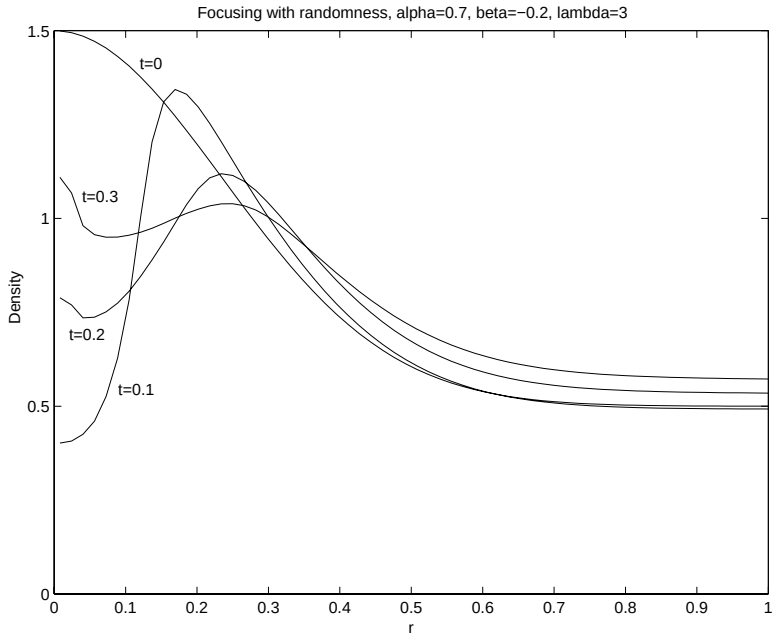


FIG. 6. *The energy density $\rho(t,r)$ versus r for the focusing case with randomness. Here $\alpha = 0.7$, $\beta = -0.2$, and $\lambda = 3$.*

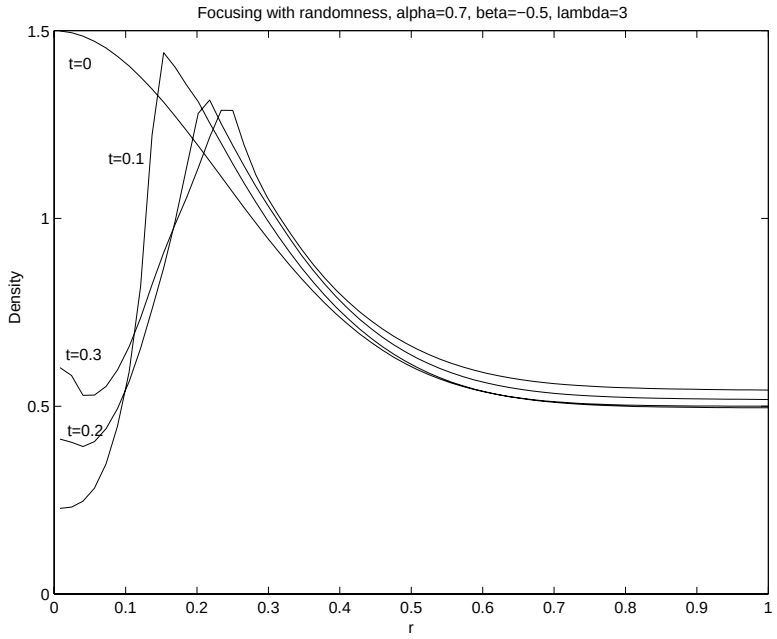


FIG. 7. *The energy density $\rho(t,r)$ versus r for the focusing case with randomness. Here $\alpha = 0.7$, $\beta = -0.5$, and $\lambda = 3$.*

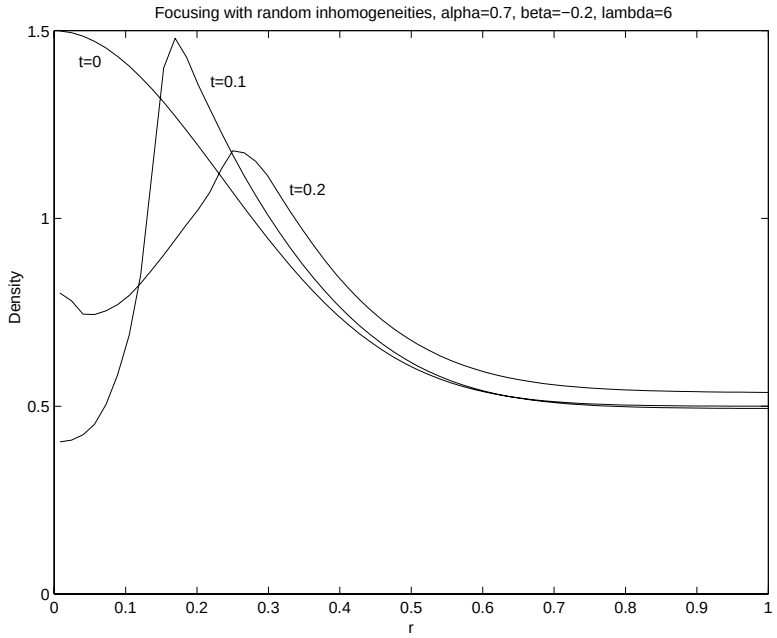


FIG. 8. *The energy density $\rho(t,r)$ versus r for the focusing case with randomness. Here $\alpha = 0.7$, $\beta = -0.2$, and $\lambda = 6$.*

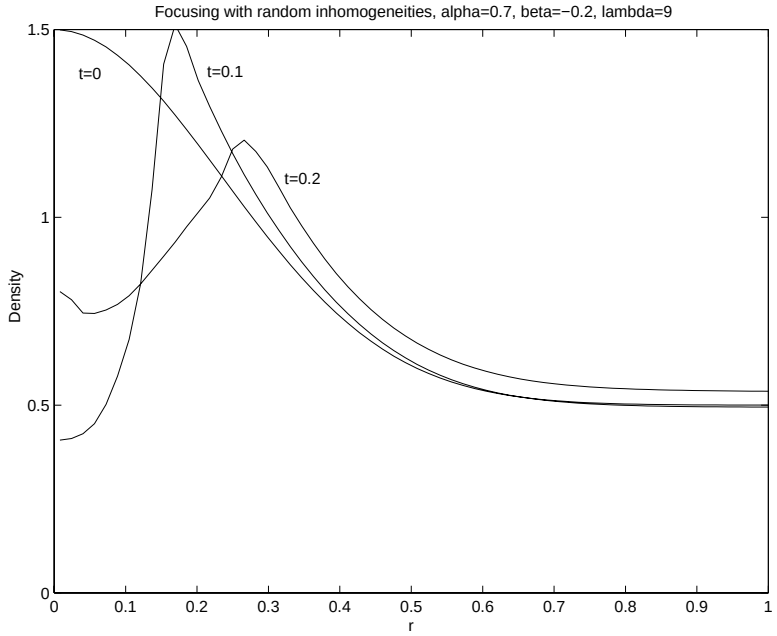


FIG. 9. *The energy density $\rho(t,r)$ versus r for the focusing case with randomness. Here $\alpha = 0.7$, $\beta = -0.2$, and $\lambda = 9$.*

It still remains an important issue to investigate the quantitative relation among α, β, and λ and how it effects the stability of the physical problem. This will be a topic for future research.

10. Numerical solution of the nonlinear diffusion equation. We also numerically solve the nonlinear diffusion equation (75). We rescale the time variable so that α disappears. We have four independent variables, but we will assume spherical symmetry to reduce them to two. Let $r = |\mathbf{x}|$ and $W(t, \mathbf{x}, k) = W(t, r, k)$. The diffusion equation (75) in polar coordinates is

$$\begin{aligned}&\frac{\partial W_0}{\partial t} - \frac{1}{12\pi}\frac{k^2}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial W_0}{\partial r}\right) + \frac{k}{12\pi}\frac{\partial W_0}{\partial k}\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial V}{\partial r}\right) + \frac{k}{6\pi}\frac{\partial^2 W_0}{\partial r\partial k}\frac{\partial V}{\partial r}\\ &\quad + \frac{1}{4\pi}\frac{\partial V}{\partial r}\frac{\partial W_0}{\partial r} - \frac{1}{12\pi}\left(\frac{\partial V}{\partial r}\right)^2\frac{\partial^2 W_0}{\partial k^2} - \frac{1}{6\pi k}\left(\frac{\partial V}{\partial r}\right)^2\frac{\partial W_0}{\partial k} = 0,\\ &\qquad V(\rho) = \beta\rho,\end{aligned} \tag{118}$$

with the initial condition

$$W_0(0, r, k) = \frac{1}{k^2}\delta(k - u_0(r))\rho_0(r). \tag{119}$$

We use a second order centered difference scheme for spatial derivatives and a composite midpoint rule to approximate $\rho_0(r)$. Let $W_{i,j}$ denote the spatial discretization of W_0. The spatially discretized form of (118) is

$$\begin{aligned}&\frac{\partial W_{i,j}}{\partial t} - \frac{1}{12\pi}\frac{k_j^2}{r_i^2(\Delta r)^2}(r_{i+1/2}^2(W_{i+1,j} - W_{i,j}) - r_{i-1/2}^2(W_{i,j} - W_{i-1,j}))\\ &\quad + \frac{k}{12\pi r_i^2}\left(\frac{W_{i,j+1} - W_{i,j-1}}{2\Delta k}\right)\frac{1}{(\Delta r)^2}[r_{i+1/2}^2(V_{i+1} - V_i) - r_{i-1/2}^2(V_i - V_{i-1})]\\ &\quad + \frac{k_j}{6\pi}\left(\frac{V_{i+1} - V_{i-1}}{2\Delta r}\right)\frac{1}{4\Delta r\Delta k}(W_{i+1,j+1} - W_{i+1,j-1} - W_{i-1,j+1} + W_{i-1,j-1})\\ &\quad + \frac{1}{4\pi}\left(\frac{V_{i+1} - V_{i-1}}{2\Delta r}\right)\frac{W_{i+1,j} - W_{i-1,j}}{2\Delta r}\\ &\quad - \frac{1}{12\pi}\left(\frac{V_{i+1} - V_{i-1}}{2\Delta r}\right)^2\frac{1}{\Delta k^2}(W_{i,j+1} - 2W_{i,j} + W_{i,j-1})\\ &\quad - \frac{1}{6\pi k_j}\left(\frac{V_{i+1} - V_{i-1}}{2\Delta r}\right)^2\frac{W_{i,j+1} - W_{i,j-1}}{2\Delta k} = 0.\end{aligned} \tag{120}$$

Here we define $r_{i+1/2} = (r_i + r_{i+1})/2$ and $k_{j+1/2} = (k_j + k_{j+1})/2$. To get V_i from W we use the composite midpoint rule. For time discretization we use the second order Runge–Kutta method. At $r = 0$ or $k = 0$ we use reflecting boundary conditions. At the outer boundaries we use vanishing flux conditions. It is important to have a domain in k that is large enough so that W is very small at the outer boundary of k. We choose the initial data

$$W_0(0, r, k) = \frac{1}{k^2}e^{-9((k-1)^2 + r^2)}. \tag{121}$$

The domain of integration is $[0, 1] \times [0, 2]$. We take 32 points in r, 64 points in k, and $\Delta t = 10^{-4}$. The numerical results for linear, defocusing, and focusing (with

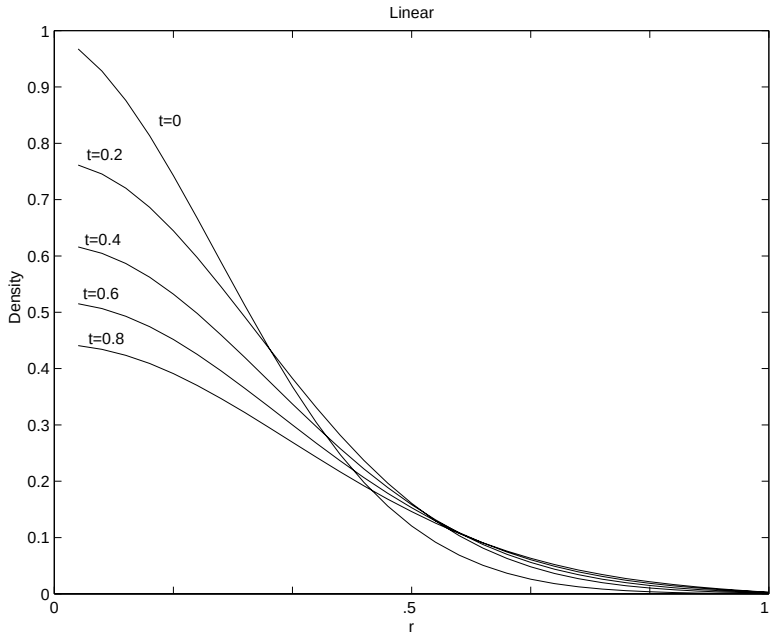


FIG. 10. *The energy density $\rho(t,r)$ in the diffusion approximation in the linear case, plotted as a function of r with $\Delta t = 0.0001$.*

$\beta = -1$ and $\beta = -0.5$) cases are shown in Figures 10, 11, 12, 13, respectively. In the linear case there is less diffusion than in the defocusing case, while in the focusing case diffusion is quite reduced. In fact, in the focusing case with $\beta = -1$ the stability condition (94) is not satisfied in part of the domain, and $\rho(t,r)$ grows locally as the solution tends to focus. In Figure 12 we show the solution until it is about to blow up (at $t \approx 0.51$). By decreasing Δt we may compute the solution to a slightly longer time, but it still breaks down numerically. For a stable solution we have to take a smaller $|\beta|$. In Figure 13 we show ρ with $\beta = -0.5$ so that the stability condition (94) is satisfied in the whole domain. The solution is stable and diffuses, at a slower rate than in the linear case. Our numerical results support the linear stability analysis that leads to (94). If (94) holds at every point of the domain initially, then the nonlinear diffusion equation is stable in time.

In Figure 14 we plot the diffusivity

$$\sigma(t) = \int_0^\infty \int_0^\infty W_0 r^4 k^2 \, dr \, dk \tag{122}$$

as a function of time, where the integrals are computed numerically with the composite midpoint rule. We see more clearly what we observed in the previous figures. The slope of $\sigma(t)$, the rate of diffusion, decreases as we go from the defocusing to the linear and to the focusing ($\beta = -0.5$) case.

11. Summary and conclusions. We have studied the interaction of nonlinear waves, solutions of the nonlinear Schrödinger equation (NLS), and random inhomogeneities, which have mean zero, are stationary, and have correlation length comparable to the wavelength. Using the Wigner phase space form of the Schrödinger equation, we formally derive a nonlinear mean field transport approximation in the

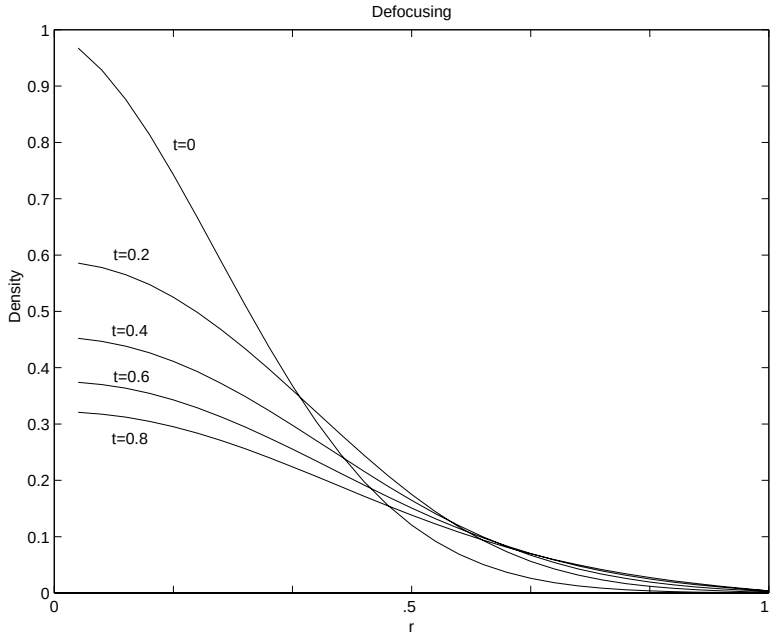


FIG. 11. *The energy density $\rho(t,r)$ in the diffusion approximation in the defocusing case ($\beta = 1$), plotted as a function of r with $\Delta t = 0.0001$.*

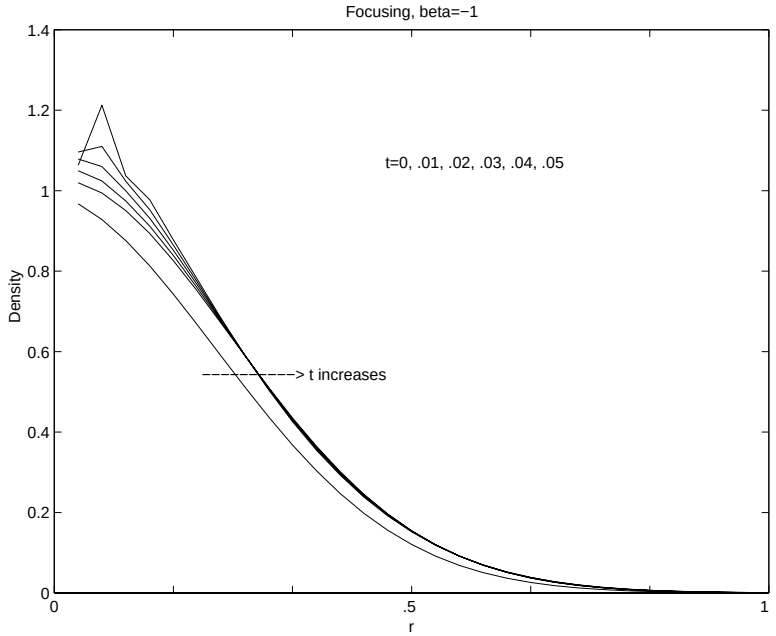


FIG. 12. *The energy density $\rho(t,r)$ in the diffusion approximation in the focusing case with $\beta = -1$, plotted as a function of r with $\Delta t = 0.0001$. Note the onset of instability where condition* (94) *is violated.*

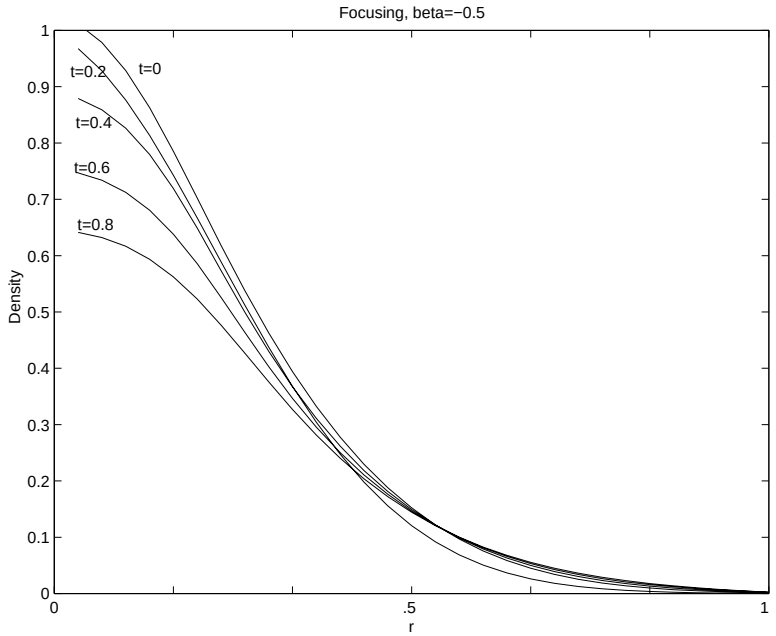


FIG. 13. *The energy density $\rho(t,r)$ in the diffusion approximation in the focusing case with $\beta = -0.5$, plotted as a function of r with $\Delta t = 0.0001$. Now the condition* (94) *holds everywhere so there is no instability.*

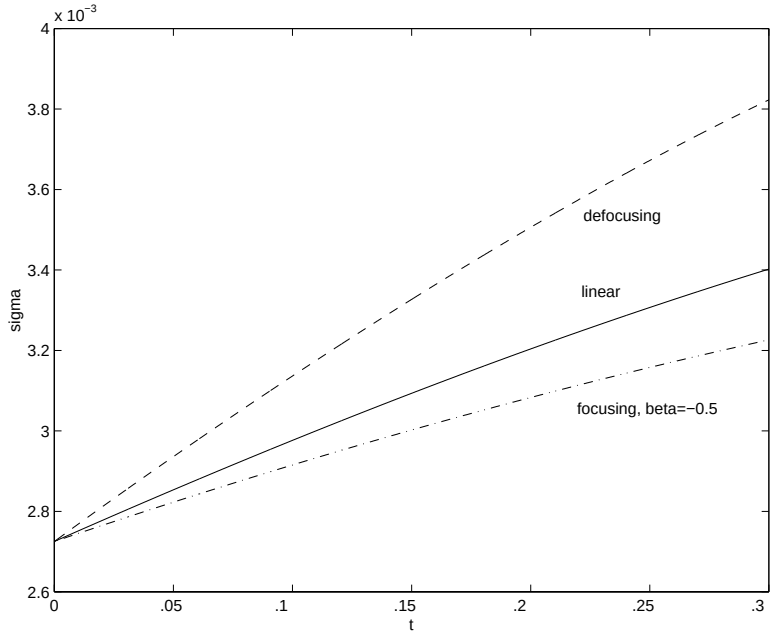


FIG. 14. *A comparison of the diffusivity $\sigma(t)$ defined in* (122) *for the linear, defocusing, and focusing cases.*

high frequency limit, and then get the diffusion approximation of this nonlinear transport equation. A linear stability analysis of the nonlinear diffusion equation shows in a simplified way how the nonlinearity and randomness interact. The focusing nonlinearity has an antidiffusive effect (see (94) with $\beta < 0$), but as long as it is not very strong the diffusion equation is linearly stable. The linear stability condition (94) has a surprising connection with the variance identity of the NLS (4): it is the right-hand side of this identity in the high frequency limit.

We then use suitable numerical schemes for both the mean field transport equation and its nonlinear diffusion approximation, and obtain numerical solutions for these two equations. Our results indicate that in the high frequency regime the random inhomogeneities prevent the wave energy from propagating in the linear and defocusing cases, but they are not strong enough to interact fully with the focusing nonlinearity. However, in the diffusive regime, randomness and nonlinearity interact fully, in a diffusive way, in all cases, defocusing and focusing. More precisely, we find that the following hold:

1. In the high frequency regime, for the linear and defocusing Schrödinger equation, the presence of random inhomogeneities prevents the wave energy from propagating and damps its amplitude. In the focusing case, the random inhomogeneities can stabilize the focusing nonlinearity if the nonlinearity is not too strong.
2. In the diffusive regime the defocusing nonlinearity enhances the overall diffusivity. The focusing nonlinearity is antidiffusive. However, when the strength of the nonlinearity is within a stability threshold given by (94), the random diffusivity dominates and the overall solution is diffusive. Thus, if the original focusing NLS does not blow up because the right-hand side of the variance identity (4) is negative, then the random inhomogeneities can stabilize it in the high frequency and diffusive regimes.

Appendix A. The mean field approximation and the correctors. In this appendix, we provide some evidence supporting the mean field approximation invoked in section 5 by analyzing the behavior of the correctors $W^{(1)}, W^{(2)}, \ldots$ in the multiscale expansion (49).

In the linear case, Spohn [17] has proved local-in-time convergence of the solution of the linear Schrödinger equation to that of the linear transport equation. Erdös and Yau [3] have improved the result so that the convergence is global in time. Both results involve detailed analysis of graphs corresponding to multiple scattering.

There are no rigorous results for the nonlinear case. The fact that the corrector analysis predicts the same limiting scattering kernel and the dimension ($d \geq 3$) required for convergence as proved by Erdös and Yau [3] is probably not a coincidence. The extent to which the correctors tell us about the *fluctuation* around the mean field in the limit $\epsilon \to 0$ remains to be tested and needs further investigation.

The terms in the expansion (49) are determined by substituting into the Wigner equation (47) and collecting terms of same orders in ϵ:

$$O\left(\frac{1}{\epsilon}\right): \qquad \mathbf{k} \cdot \nabla_{\mathbf{y}} W = 0, \tag{123}$$

$$O\left(\frac{1}{\sqrt{\epsilon}}\right): \qquad \mathbf{k} \cdot \nabla_{\mathbf{y}} W^{(1)} + \mathcal{L}_{\frac{\mathbf{x}}{\epsilon}} W = 0, \tag{124}$$

$$O(1): \qquad \frac{\partial W}{\partial t} + \mathbf{k} \cdot \nabla_{\mathbf{x}} W + \mathcal{L}_{\frac{\mathbf{x}}{\epsilon}} W^{(1)} = -\mathbf{k} \cdot \nabla_{\mathbf{y}} W^{(2)}. \tag{125}$$

Equation (123) means that $W = W(t, \mathbf{x}, \mathbf{k})$ does not depend on the fast variable $\mathbf{y}$. Equation (124) is the corrector equation for $W^{(1)}$, which is degenerate. With a standard regularization, (124) becomes

$$\epsilon W_\epsilon^{(1)} + \mathbf{k} \cdot \nabla_{\mathbf{y}} W_\epsilon^{(1)} + \mathcal{L}_{\frac{\mathbf{x}}{\epsilon}} W = 0, \tag{126}$$

which has the solution

$$W_\epsilon^{(1)} = i \int d\mathbf{p}\, \hat{V}(\mathbf{p}) \frac{e^{-i\mathbf{p}\cdot\mathbf{y}}}{\epsilon - i\mathbf{k}\cdot\mathbf{p}} \left[W\left(t, \mathbf{x}, \mathbf{k} - \frac{\mathbf{p}}{2}\right) - W\left(t, \mathbf{x}, \mathbf{k} + \frac{\mathbf{p}}{2}\right) \right]. \tag{127}$$

Substituting (127) into (125), taking expectation, and passing to the limit $\epsilon \to 0$, we get the transport equation (50) for W. We see from (125) that the second order corrector $W^{(2)}$ satisfies the equation

$$\mathbf{k} \cdot \nabla_{\mathbf{y}} W^{(2)} + \mathcal{L}_{\frac{\mathbf{x}}{\epsilon}} W^{(1)} - \langle \mathcal{L}_{\frac{\mathbf{x}}{\epsilon}} W^{(1)} \rangle = 0,$$

which again should be regularized as in (127). Higher order correctors can be determined similarly. We focus on the first corrector $W^{(1)}$.

As pointed out in section 4, the initial data for the Wigner function does not converge strongly; thus neither the solution W^ϵ of the Wigner equation nor the corrector $W_\epsilon^{(1)}$ is expected to converge strongly. This is, indeed, the case, as stated next.

PROPOSITION A.1. *Suppose that $W(t, \mathbf{x}, \mathbf{k})$ is such that the function*

$$f(\mathbf{k}, \mathbf{p}) = \hat{R}(\mathbf{p}) \int d\mathbf{x} \left[W\left(\mathbf{x}, \mathbf{k} - \frac{\mathbf{p}}{2}\right) - W\left(\mathbf{x}, \mathbf{k} + \frac{\mathbf{p}}{2}\right) \right]^2$$

is continuous and its zero set does not contain the set $\{(\mathbf{k}, \mathbf{p}) : \mathbf{k} \cdot \mathbf{p} = 0\}$. Then, $\epsilon \iint \langle |W_\epsilon^{(1)}|^2 \rangle \, d\mathbf{x}\, d\mathbf{k}$ does not vanish as $\epsilon \to 0$, in any dimension.

Proof. A straightforward calculation leads to

$$\begin{aligned}
&\epsilon \iint \langle |W_\epsilon^{(1)}|^2 \rangle \, d\mathbf{x}\, d\mathbf{k} \\
&= \int d\mathbf{p}\, \hat{R}(\mathbf{p}) \int d\mathbf{k}\, \frac{\epsilon}{\epsilon^2 + (\mathbf{k}\cdot\mathbf{p})^2} \int d\mathbf{x} \left[W\left(\mathbf{x}, \mathbf{k} - \frac{\mathbf{p}}{2}\right) - W\left(\mathbf{x}, \mathbf{k} + \frac{\mathbf{p}}{2}\right) \right]^2 \\
&\geq \frac{1}{2\epsilon} \int d\mathbf{k} \int_{|\mathbf{k}\cdot\mathbf{p}| \leq \epsilon} d\mathbf{p}\, \hat{R}(\mathbf{p}) \int d\mathbf{x} \left[W\left(\mathbf{x}, \mathbf{k} - \frac{\mathbf{p}}{2}\right) - W\left(\mathbf{x}, \mathbf{k} + \frac{\mathbf{p}}{2}\right) \right]^2 \\
&\geq \frac{1}{2\epsilon} \int d\mathbf{k} \int_{|\mathbf{k}\cdot\mathbf{p}| \leq \epsilon} d\mathbf{p}\, f(\mathbf{k}, \mathbf{p}). \qquad (128)
\end{aligned}$$

As the set $\{\mathbf{k} \in R^d : |\mathbf{k} \cdot \mathbf{p}| \leq \epsilon\} \bigcap \mathrm{supp}\{f(\mathbf{k}, \mathbf{p})\}$ has a measure of order ϵ for W satisfying the stated assumption, the expression (128) does not vanish in the limit $\epsilon \to 0$, as we wanted to show.

We note that "generic" functions W that are *not* spherically symmetric in $\mathbf{k}$ satisfy the condition stated in Proposition A.1.

However, $\sqrt{\epsilon}\, W_\epsilon^{(1)}$ does vanish strongly (in $\mathbf{x}$) if it is first integrated against a test function of $\mathbf{k}$, as stated in the next proposition. This provides some reason for using the mean field hypothesis in the derivation of the transport equation.

PROPOSITION A.2. *For $d \geq 3$ and any differentiable $W(\mathbf{x}, \mathbf{k})$ with a compact support, we have*

$$\lim_{\epsilon \to 0} \epsilon \int d\mathbf{x} \left\langle \left| \int d\mathbf{k}\, W_\epsilon^{(1)} \phi(\mathbf{k}) \right|^2 \right\rangle = 0 \quad \forall \phi(\mathbf{k}) \in C_c^\infty. \tag{129}$$

Proof. After taking the expectation, the expression on the left-hand side of (129) becomes

$$\epsilon \int d\mathbf{p}\, \frac{\hat{R}(\mathbf{p})}{|\mathbf{p}|^2} \int d\mathbf{x} \left| \int d\mathbf{k}\, \frac{\phi(\mathbf{k})}{\epsilon - i\mathbf{k}\cdot\hat{\mathbf{p}}} \left[W\left(\mathbf{x}, \mathbf{k} - \frac{\mathbf{p}}{2}\right) - W\left(\mathbf{x}, \mathbf{k} + \frac{\mathbf{p}}{2}\right) \right] \right|^2 , \tag{130}$$

where $\hat{\mathbf{p}} = \mathbf{p}/|\mathbf{p}|$. Since $\phi(\mathbf{k})[W(\mathbf{x}, \mathbf{k} - \mathbf{p}/2) - W(\mathbf{x}, \mathbf{k} + \mathbf{p}/2)]$ is differentiable, we have

$$\begin{aligned} &\lim_{\epsilon\to 0} \int d\mathbf{k}\, \frac{\phi(\mathbf{k})}{\epsilon - i\mathbf{k}\cdot\hat{\mathbf{p}}} \left[W\left(\mathbf{x}, \mathbf{k} - \frac{\mathbf{p}}{2}\right) - W\left(\mathbf{x}, \mathbf{k} + \frac{\mathbf{p}}{2}\right) \right] \\ &= \lim_{\epsilon\to 0} \int_{|\mathbf{k}\cdot\hat{\mathbf{p}}|>\epsilon} d\mathbf{k}\, \frac{\phi(\mathbf{k})}{-i\mathbf{k}\cdot\hat{\mathbf{p}}} \left[W\left(\mathbf{x}, \mathbf{k} - \frac{\mathbf{p}}{2}\right) - W\left(\mathbf{x}, \mathbf{k} + \frac{\mathbf{p}}{2}\right) \right] \\ &= i \int d\mathbf{k}_\perp(\hat{\mathbf{p}}) \fint d(\mathbf{k}\cdot\hat{\mathbf{p}})\, \frac{\phi(\mathbf{k})}{\mathbf{k}\cdot\hat{\mathbf{p}}} \left[W\left(\mathbf{x}, \mathbf{k} - \frac{\mathbf{p}}{2}\right) - W\left(\mathbf{x}, \mathbf{k} + \frac{\mathbf{p}}{2}\right) \right], \end{aligned} \tag{131}$$

with $\mathbf{k}_\perp(\hat{\mathbf{p}})$ the orthogonal projection of $\mathbf{k}$ onto the plane normal to $\mathbf{p}$. Here $\fint$ stands for the Cauchy principal value integral. Since $|\mathbf{p}|^{-2}$ in (130) is an integrable singularity in three or more dimensions, the expression in (129) is $O(\epsilon)$ as $\epsilon \to 0$, as we wanted to show.

Acknowledgments. S. J. thanks the Institute for Advanced Study in Princeton and the Department of Mathematics at Stanford University for their hospitality during his extended visits there.

REFERENCES

[1] N. Ben Abdallah and P. Degond, *On a hierarchy of macroscopic models for semiconductors*, J. Math. Phys., 37 (1996), pp. 3306–3333.
[2] G. Dell'Antonio, *Large time small coupling behavior of a quantum particle in a random potential*, Ann. Inst. H. Poincaré Sect. A (N.S.), 39 (1983), pp. 339–384.
[3] L. Erdös and H.T. Yau, *Linear Boltzmann equation as the weak coupling limit of a random Schrödinger equation*, Comm. Pure Appl. Math., 53 (2000), pp. 667–735.
[4] J. Froelich and T. Spencer, *Absence of diffusion in the Anderson tight binding model for large disorder or low energy*, Comm. Math. Phys., 88 (1983), pp. 151–184.
[5] P. Gérard, P. Markovich, N. Mauser, and F. Poupaud, *Homogenization limits and Wigner transforms*, Comm. Pure Appl. Math., 50 (1997), pp. 323–380.
[6] R.T. Glassey, *On the blowing-up of solutions to the Cauchy problem for the nonlinear Schrödinger equation*, J. Math. Phys., 18 (1977), pp. 1794–1797.
[7] E. Grenier, *Limite semi-classique de l'equation de Schrödinger non linéaire en temps petit*, C. R. Acad. Sci. Paris Sér. I Math., 320 (1995), pp. 691–694.
[8] T.G. Ho, L.G. Landau, and A.J. Wilkins, *On the weak coupling limit for a Fermi gas in a random potential*, Rev. Math. Phys., 5 (1993), pp. 209–298.
[9] S. Jin, C.D. Levermore, and D.W. McLaughlin, *The behavior of solutions of the NLS equation in the semiclassical limit*, in Singular Limits of Dispersive Waves, Lyon, 1991, NATO Adv. Sci. Inst. Ser. B Phys. 320, Plenum, New York, 1994, pp. 235–255.
[10] S. Jin, C.D. Levermore, and D.W. McLaughlin, *The semiclassical limit of the defocusing NLS hierarchy*, Comm. Pure Appl. Math., 52 (1999), pp. 613–654.
[11] J.B. Keller and R. Lewis, *Asymptotic methods for partial differential equations: The reduced wave equation and Maxwell's equations*, in Surveys in Applied Mathematics, J.B. Keller, D. McLaughlin, and G. Papanicolaou, eds., Plenum Press, New York, 1995.
[12] S. Kamvissis, K.T.-R. McLaughlin, and P.D. Miller, *Semiclassical Soliton Ensembles for the Focusing Nonlinear Schroedinger Equation*, Annals of Math., Studies Series, Princeton University Press, Princeton, NJ, 2002 (to appear).
[13] R.J. LeVeque, *Numerical Methods for Conservation Laws*, Birkhäuser, Basel, 1992.
[14] P.L. Lions and T. Paul, *Sur les mesures de Wigner*, Rev. Mat. Iberoamericana, 9 (1993), pp. 553–618.

[15] L. Ryzhik, G. Papanicolaou, and J. Keller, *Transport equations for elastic and other waves in random media*, Wave Motion, 24 (1996), pp. 327–370.
[16] P. Sheng, *Introduction to Wave Scattering, Localization, and Mesoscopic Phenomena*, Academic Press, San Diego, 1995.
[17] H. Spohn, *Derivation of the transport equation for electrons moving through random impurities*, J. Statist. Phys., 17 (1977), pp. 385–412.
[18] C. Sulem and P.-L. Sulem, *Nonlinear Schrödinger Equation: Self-Focusing and Wave Collapse*, Springer-Verlag, New York, 1999.
[19] E. Wigner, *On the quantum correction for thermodynamic equilibrium*, Phys. Rev., 40 (1932), pp. 749–759.
[20] V.E. Zakharov, *Handbook of Plasma Physics*, Vol. 2, M.N. Rosenbluth and R.Z. Sagdeev, eds., Elsevier, New York, 1984.
[21] P. Zhang, Y. Zheng, and N.J. Mauser, *The limit from the Schrödinger–Poisson to the Vlasov–Poisson equations with general data in one dimension*, Comm. Pure Appl. Math., 55 (2002), pp. 582–632.

SIAM J. APPL. MATH.
Vol. 63, No. 4, pp. 1359–1377

APPROXIMATION OF THE INTEGRAL BOUNDARY LAYER EQUATION BY THE KURAMOTO–SIVASHINSKY EQUATION*

HANNES UECKER†

Abstract. In suitable parameter regimes the integral boundary layer equation (IBLe) can be formally derived as a long wave approximation for the flow of a viscous incompressible fluid down an inclined plane. For very long waves with small amplitude, the IBLe can be further reduced to the Kuramoto–Sivashinsky equation (KSe). Here we justify this reduction of the IBLe to the KSe. Using energy estimates, we show that solutions of the KSe approximate solutions of the IBLe over sufficiently long time scales. This is a step towards understanding the approximation properties of the KSe for the full Navier–Stokes system describing the inclined film flow.

Key words. inclined film flow, integral boundary layer equation, Kuramoto–Sivashinsky equation, energy estimates

AMS subject classifications. 35K55, 35B45, 76E17

PII. S0036139902405900

1. Introduction and statement of result.

For typical flow conditions the so-called Nusselt flow of a viscous incompressible fluid down an inclined plane is subject to long wave surface instabilities, and trains of solitary waves develop on the free surface. Starting from the Navier–Stokes equations (NSe), a number of reduced equations have been formally derived to describe the evolution of the free surface and in particular to understand the formation of these wavetrains. See [3] for an extensive review and [12] for experiments on inclined film flows, and the comprehensive monograph [4] for a wealth of further information.

Here we study analytically the relationship between two of the approximate equations. The first one is the so-called integral boundary layer equation (IBLe), which is derived from the NSe using a long wave expansion followed by an averaging over the film height. In Appendix B we briefly review this derivation of the IBLe.

By a small amplitude and second long wave expansion in the IBLe, corresponding to a small amplitude and very long wave expansion of the NSe, the IBLe can be further reduced to the Kuramoto–Sivashinsky equation (KSe). This second reduction is justified in this paper; we show that the Kuramoto–Sivashinsky (KS) dynamics can be observed in the IBLe; see Theorem 1.1.

Using the time and space scales of the NSe, the IBLe we consider reads

(1.1)
$$\begin{aligned} h_t &= -q_x, \\ q_t &= -\frac{6}{5}\partial_x\left(\frac{q^2}{h}\right)+\frac{2}{\mathrm{R}}\left(h-\frac{3q}{2h^2}-h_x h\cot\theta\right)+\mathrm{W}\varepsilon^{-2}h\left(\partial_x^3 h-\frac{3}{2}\partial_x^3 h h_x^2-3h_{xx}^2 h_x\right) \\ &\quad +\frac{1}{\mathrm{R}}\left(\frac{7}{2}q_{xx}-\frac{9q_x h_x}{h}+\frac{6qh_x^2}{h^2}-\frac{9qh_{xx}}{2h}\right), \end{aligned}$$

*Received by the editors April 18, 2002; accepted for publication (in revised form) November 11, 2002; published electronically May 22, 2003. This research was supported by the Deutsche Forschungsgemeinschaft under grant UE60/1.
http://www.siam.org/journals/siap/63-4/40590.html
†Department of Mathematics, University of Maryland, College Park, MD 20740. Current address: Math. Institut I, Universität Karlsruhe, 76128 Karlsruhe, Germany (hannes.uecker@math.uni-karlsruhe.de).

where $x \in \mathbb{R}$ and $t > 0$. Here we remark that, in contrast to the Shkadov model [23], (1.1) is a parabolic system due the dissipative term $\frac{7}{2\mathrm{R}}\partial_x^2 q$ on the right-hand side; see Remarks 1.4 and Appendix B.3 for further discussion.

In (1.1), h is the film height, q describes the flow, $0 < \theta \leq \pi/2$ is the inclination angle, R is the Reynolds number, W is a normalized Weber number, and $0 < \varepsilon \ll 1$ is a small parameter. In the derivation of (1.1) it is assumed that the Weber number $\mathrm{W}_e = \mathrm{W}\varepsilon^{-2} = \mathcal{O}(\varepsilon^{-2})$, while $\mathrm{R} = \mathcal{O}(1)$ and $\cot\theta = \mathcal{O}(1)$. The latter means that the plane may not be close to horizontal. However, a vertical plane, i.e., $\cot\theta = 0$, is allowed. The parameter W could be adsorbed into ε, but we think the analysis becomes more transparent by keeping W. As noted, see Appendix B for the underlying scalings.

In the IBLe the Nusselt solution of the inclined film problem corresponds to $(h, q) = (1, 2/3)$. Since we are interested in the instability of this solution, we will assume throughout that R is larger than the critical Reynolds number (see [2]), i.e.,

$$\mathrm{R} > \mathrm{R}_c = \frac{5}{4}\cot\theta. \tag{1.2}$$

With an abuse of notation we set $h = 1 + \eta, q = 2/3 + q$, and expand (1.1) up to quadratic terms, since from previous work, e.g., [10], it is well known and can also readily be seen in the analysis below that cubic and higher order terms play no role in the justification of the long wave/small amplitude approximation for (1.1). See, however, Remarks 1.3, 3.2, and A.5 for changes in the function spaces if cubic and higher order terms are included.

We write the quadratic expansion as

$$\eta_t = -q_x, \tag{1.3}$$

$$q_t = a_0(\eta)\eta + a_1(\eta, q)\eta_x + a_2(\eta, q)\eta_{xx} + \varepsilon^{-2}a_3(\eta)\eta_{xxx} - b_0(\eta)q - b_1(\eta, q)q_x + b_2 q_{xx}, \tag{1.4}$$

where

$$\begin{aligned} &a_0(\eta) = \frac{6-6\eta}{\mathrm{R}}, \quad a_1(\eta, q) = \left(\frac{4}{5} - \frac{2}{\mathrm{R}}\cot\theta - \frac{8}{5}\eta + \frac{8}{5}q + \frac{6}{\mathrm{R}}\eta_x\right), \\ &a_2(\eta, q) = \frac{1}{\mathrm{R}}\left(-3 + \frac{9}{2}\eta - \frac{9}{2}q\right), \quad a_3(\eta) = \mathrm{W}(1+\eta), \\ &b_0(\eta) = \frac{3}{\mathrm{R}}(1-2\eta), \quad b_1(\eta, q) = \frac{8}{5} - \frac{8}{5}\eta + \frac{12}{5}q + \frac{9}{\mathrm{R}}\eta_x, \quad b_2 = \frac{7}{2\mathrm{R}}. \end{aligned} \tag{1.5}$$

Splitting (1.3) and (1.4) into linear and nonlinear terms, we write

$$U_t = A_0 U + F(U), \tag{1.6}$$

$$U = \begin{pmatrix} \eta \\ q \end{pmatrix}, \quad A_0 = A_0(\partial_x) = \begin{pmatrix} 0 & -\partial_x \\ a_{00} + a_{10}\partial_x + a_{20}\partial_x^2 + \varepsilon^{-2}a_{30}\partial_x^3 & -b_{00} - b_{10}\partial_x + b_{20}\partial_x^2 \end{pmatrix},$$

where $a_{00} = a_0(0), a_{10} = a_1(0, 0), \ldots$, and where F contains the quadratic terms.

Inserting $U = \mathrm{e}^{\mu t + \mathrm{i}kx}U(k)$ into (1.6), we obtain the dispersion relation

$$\begin{aligned} \mu_{1,2}(k) = &-\frac{1}{2}\left(\frac{7}{2\mathrm{R}}k^2 + \frac{8}{5}\mathrm{i}k + \frac{3}{\mathrm{R}}\right) \\ &\pm\sqrt{\frac{1}{4}\left(\frac{7}{2\mathrm{R}}k^2 + \frac{8}{5}\mathrm{i}k + \frac{3}{\mathrm{R}}\right)^2 - \frac{6}{\mathrm{R}}\mathrm{i}k - \left(\frac{4}{5} - \frac{2}{\mathrm{R}}\cot\theta\right)k^2 + \frac{3}{\mathrm{R}}\mathrm{i}k^3 - \mathrm{W}\varepsilon^{-2}k^4}. \end{aligned} \tag{1.7}$$

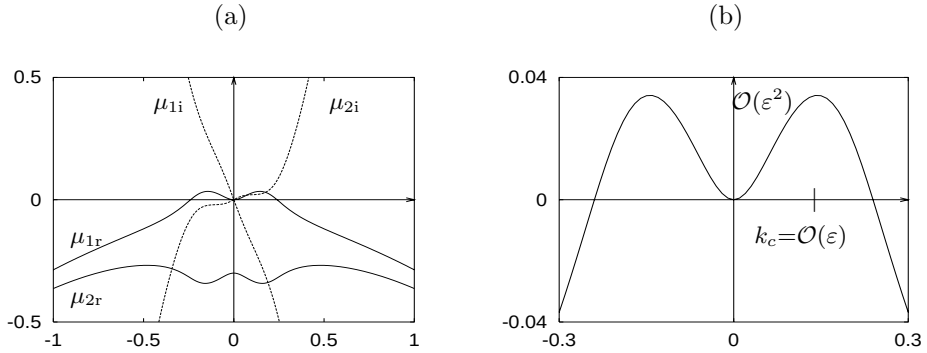


FIG. 1.1. *The dispersion relation* $\mu_j = \mu_{j\mathrm{r}} + \mathrm{i}\mu_{j\mathrm{i}}$, $j = 1, 2$, *for the IBLe;* $\varepsilon = 0.2$, $W = 1$, $\mathrm{R} = 10$, $\theta = \pi/2$. (a) *The two curves of eigenvalues* $\mu_{1,2}$; (b) *blowup of* $\mu_{1\mathrm{r}}(k)$ *near* $k = 0$.

This spectrum of the operator $A_0(\mathrm{i}k)$ is sketched in Figure 1.1. In particular, from μ_1 we recover the instability criterion (1.2): for $\mathrm{R} > \mathrm{R}_c$ we have a long wave instability with maximum growth rate $\mathrm{Re}\mu_1(k_c) = \mathcal{O}(\varepsilon^2)$, $k_c = \mathcal{O}(\varepsilon)$. Moreover, for $|k| \to \infty$ we have

$$\begin{aligned}\mu_{1,2}(k) &= \left(-\frac{7}{4\mathrm{R}} \pm \sqrt{\left(\frac{7}{4\mathrm{R}}\right)^2 - \mathrm{W}\varepsilon^{-2}}\right) k^2 + \mathcal{O}(|k|^{3/2}) \\ &= \left(-\frac{7}{4\mathrm{R}} \pm \mathrm{i}(\varepsilon^{-1}\sqrt{\mathrm{W}} + \mathcal{O}(\varepsilon))\right) k^2 + \mathcal{O}(|k|^{3/2}),\end{aligned} \tag{1.8}$$

where in the second equality we have assumed for simplicity that $(\frac{7}{4\mathrm{R}})^2 < \mathrm{W}\varepsilon^{-2}$. This shows the parabolic damping (and the very fast oscillations) of the high wavenumber modes.

In fact, A_0 generates an analytic semigroup e^{tA_0} with

$$\|e^{tA_0}U\|_Y \le C\mathrm{e}^{C\varepsilon^2 t}\|U\|_Y, \tag{1.9}$$

where as phase-space Y we choose, for instance, the Hilbert space $Y{=}H^2(\mathbb{R}){\times}H^1(\mathbb{R})$ equipped with the norm

$$\|U\|_Y^2 = \frac{1}{2}\int_{\mathbb{R}} \left\{q^2{+}c_1\eta^2{-}2c_2q\eta{-}2c_3\eta_x q{+}c_4q_x^2{+}\varepsilon^{-2}\mathrm{W}(\eta_x^2{+}c_4\eta_{xx}^2)\right\}\mathrm{d}x. \tag{1.10}$$

Here we must choose $c_2 = 2$, and c_1, c_3, c_4 can be chosen as

$$c_1 = 9, \quad c_3 = -\frac{11\mathrm{R}}{5} + \frac{2\cot\theta}{3}, \quad c_4 = \mathrm{R}^2; \tag{1.11}$$

see section 2.1, where we also motivate the choice of $\|\cdot\|_Y$. The strong weighting of derivatives of η in (1.10) reflects the fact that in (1.6) the small parameter ε appears in a rather unusual way, namely as an inverse power in front of the damping by the surface tension. This is inherited from the fact that in the underlying Navier–Stokes equations we consider the limit of large surface tension; see Appendix B.2.

Assuming very long waves with a small amplitude, the KSe for the film height η can be formally derived from the NSe. Accordingly, the KSe can also be derived from the IBLe, namely by the ansatz

$$\eta(t,x) = \varepsilon\eta_1(T,X) + \mathcal{O}(\varepsilon^3), \qquad q(t,x) = \varepsilon q_1(T,X) + \varepsilon^2 q_2(T,X) + \mathcal{O}(\varepsilon^3), \tag{1.12}$$

where $T = \varepsilon^2 t$ and $X = x - ct$ are the very slow timescale and the very long space scale in a frame moving with the speed c. These time and space scalings follow directly from the dispersion relation (1.7) for A_0. Plugging (1.12) into (1.6), we obtain the hierarchy of equations

$$\begin{aligned}
&\mathcal{O}(\varepsilon(1.4)): \quad q_1 = 2\eta_1,\\
&\mathcal{O}(\varepsilon^2(1.3)): \quad -c\eta_{1X} = -q_{1X} = -2\eta_{1X} \quad \Rightarrow \quad c = 2,\\
&\mathcal{O}(\varepsilon^2(1.4)): \quad -cq_{1X} = -\frac{8}{5}q_{1X} + \left(\frac{4}{5} - \frac{2}{\mathrm{R}}\cot\theta\right)\eta_{1X} + \mathrm{W}\partial_X^3\eta_1 - \frac{3}{\mathrm{R}}q_2 + \frac{6}{\mathrm{R}}(\eta_1 q_1 - \eta_1^2),\\
&\qquad\qquad \Rightarrow q_2 = \left(\frac{8\mathrm{R}}{15} - \frac{2}{3}\cot\theta\right)\eta_{1X} + \frac{1}{3}\mathrm{RW}\partial_X^3\eta_1 + 2\eta_1^2; \qquad (1.13)
\end{aligned}$$

that is, q_1, q_2 are given as functions of η_1. At $\mathcal{O}(\varepsilon^3(1.3))$ we find $\partial_T\eta_1 = -\partial_X q_2$, which gives the KSe

$$\partial_T\eta_1 = -\left(\frac{8\mathrm{R}}{15} - \frac{2}{3}\cot\theta\right)\partial_X^2\eta_1 - \frac{1}{3}\mathrm{RW}\partial_X^4\eta_1 - 4\eta_1\partial_X\eta_1 \qquad (1.14)$$

for η_1. Note that the coefficient of $\partial_X^2\eta_1$ is less than zero iff $\mathrm{R} > \mathrm{R}_c$.

Obviously (1.14) is a much simpler equation than (1.6) since it is a semilinear scalar parabolic equation, while the IBLe is a quasilinear system. Moreover, the KSe is a generic long wave equation; see, e.g., [13] for a basic review and, e.g., [14] for the existence and smoothness of solutions $\eta_1 \in C([0, T_0], H^m(\mathbb{R}))$ to initial conditions $\eta_1(0) \in H^m(\mathbb{R})$.

We define the approximation

$$\varepsilon\psi(t,x) = \begin{pmatrix} \varepsilon q_1(T,X) + \varepsilon^2 q_2(T,X) \\ \varepsilon\eta_1(T,X) \end{pmatrix}, \qquad (1.15)$$

with q_1, q_2 given by (1.13), and the spaces

$$H^{r,s}((0,t_0)\times\mathbb{R}) = L^2((0,t_0), H^r(\mathbb{R})) \cap H^s((0,t_0), L^2(\mathbb{R}))$$

and show the following result.

THEOREM 1.1. *Assume that $\eta_1 \in C([0,T_0], H^9(\mathbb{R}))$ is a solution of the KSe. Then for all $C_1 > 0$ there exist $\varepsilon_0, C_2 > 0$ such that for all $\varepsilon \in (0, \varepsilon_0)$ the following holds. If*

$$\|U(0,\cdot) - \varepsilon\psi(0,\cdot)\|_Y \le C_1\varepsilon^{3/2}, \qquad (1.16)$$

then there exists a unique solution $U = (\eta, q)$ of the IBLe up to large time $t_0 = T_0/\varepsilon^2$,

$$\eta \in H^{3,3/2}((0,t_0)\times\mathbb{R}), \quad q \in H^{2,1}((0,t_0)\times\mathbb{R}). \qquad (1.17)$$

For $t > 0$ the solution is smooth, and it fulfills

$$\sup_{0\le t\le t_0} \|U(t,\cdot) - \varepsilon\psi(t,\cdot)\|_Y \le C_2\varepsilon^{3/2}. \qquad (1.18)$$

Remark 1.2. The properties of the spaces $H^{r,s}((0,t_0)\times\mathbb{R})$ are reviewed in Appendix A. Here we remark that $H^{3,3/2}((0,t_0)\times\mathbb{R})\times H^{2,1}((0,t_0)\times\mathbb{R}) \subset C([0,t_0], Y)$ such that (1.18) makes sense. Note that C_1, C_2 in Theorem 1.1 do not depend on

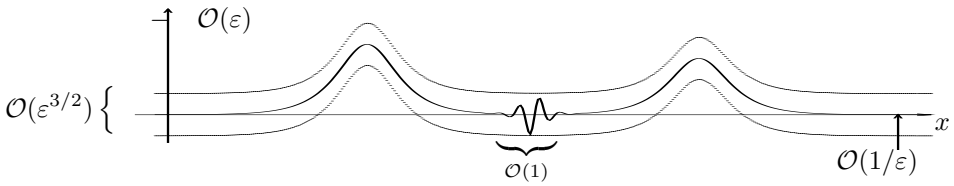


FIG. 1.2. *Sketch of the initial data. For q_0 we may allow small oscillations on the original scale.*

$\varepsilon \in (0, \varepsilon_0)$. However, ε_0, C_1, C_2 do depend on R and W in such a way that, for instance, $\mathrm{R}_2 \geq \mathrm{R}_1$ gives $\varepsilon_0(\mathrm{R}_2) \leq \varepsilon_0(\mathrm{R}_1)$ and similarly for C_1, C_2, but we do not work this out in detail.

From (1.18) we obtain $\sup_{0\leq t\leq t_0} \|U(t,\cdot) - \varepsilon\psi(t,\cdot)\|_{L^\infty} \leq C_2\varepsilon^{3/2}$. Thus, the error is small compared to the size of the solution. Moreover, the error for η_x is much smaller, i.e., $\sup_{0\leq t\leq t_0} \|\partial_x\eta(t,\cdot) - \varepsilon\partial_x\psi_1(t,\cdot)\|_{C^0} \leq C_2\varepsilon^{7/2}$. On the other hand, we must impose the same condition on the initial condition. This means that η_0 must be a long wave in a much stricter sense than q_0. For q_0 we may allow small perturbations of $\varepsilon\psi_1$ on the original scale. Such "fast" oscillations in η_0 would violate (1.16). This situation is sketched in Figure 1.2.

Remark 1.3. Theorem 1.1 also holds in higher order Sobolev spaces; see Remark 3.2. In fact, if cubic terms were included in (1.6), then in order to control the nonlinearity we would *have* to work in $H^3 \times H^2$, due to the term $\eta_{xx}^2\eta_x$; cf. Remark A.5.

Remark 1.4. As already said, our IBLe differs from previously derived *hyperbolic* IBLe, IBL_h, also called the Shkadov model [23] (see Appendix B.3), in that the linearization of (1.1) around Nusselt's solution is *sectorial*. Heuristically, this can be seen from the dispersion relation (1.7) (cf.(1.8)): the spectrum of A_0 lies in a sector around the negative real axis; the additionally needed resolvent estimates are given in Lemma A.3.

A parabolic IBLe has also been derived in [16] (see also [4, section 3.3]), under the assumption of $\mathrm{W} = \mathcal{O}(\varepsilon^{-2})$, $\mathrm{R} = \mathcal{O}(\varepsilon^{-1})$, and small θ.

Remark 1.5. Numerical simulations of the free boundary problem for the governing NSe of IBL_h and of the KSe suggest that for high but finite Weber numbers, corresponding to finite $\varepsilon > 0$, IBL_h is valid as an approximation for NSe up to intermediate Reynolds numbers. We conclude that this also holds true for our IBLe (1.1), since (1.1) is derived as a higher order approximation of the NSe than IBL_h. On the other hand, the KSe gives good results only for smaller Reynolds numbers; for details, see [3] and the references therein. See also [17] for comparison of solutions of the IBLe with experimental results and extensive numerical studies of the NSe.

Remark 1.6. One major reason for the reduction of the inclined film problem to the IBLe or to the KSe is to gain understanding of the formation of solitary and periodic waves on the free surface of the film. For the existence and properties of solitary waves for IBL_h and the KSe, see, again [3, 4] and the references therein. In a somewhat different scaling, a generalized KSe or KdV-KS equation with third order dispersion can be derived from the NSe (see [25]). For this KdV-KSe there are analytical results and extensive numerical studies concerning the stability and dynamics of solitary waves [15, 7, 5]. Our result does not contribute to the understanding of these waves but says that the (periodic or solitary) waves of the KSe can be seen as small amplitude waves in the IBLe over long times.

Remark 1.7. The IBLe itself is only formally derived from the NSe. Error

estimates in the sense of Theorem 1.1 seem to be very difficult for this reduction. Instead, we suggest studying directly the reduction of the NSe to the KSe. This will be the subject of future work. Since the linearization around Nusselt's solution in the NSe is also sectorial [1, 24], we can use similar methods as in the present paper.

Finally, we remark that a result like Theorem 1.1 is not obvious. There are counterexamples where formally derived amplitude equations make wrong predictions about the dynamics in the original system [20, 8]. Moreover, the question of which simplified equation, dependent on the parameter regime, still describes the inclined film problem is not settled. Here we contribute to the answer in the sense that for $\mathcal{O}(1)$ Reynolds numbers and in the limit of (very) large Weber number the KSe accurately captures the IBLe dynamics for long small amplitude waves over the right time scale.

Similarly to our result, the validity of multiple scale approximations to the NSe in a fixed domain, where the instability is located at a finite nonzero wavenumber, has been shown in [18, 21]. See also, e.g., [22] for the water wave problem, and [19] for such approximation results in simpler settings, i.e., for scalar semilinear parabolic problems.

To explain the difficulty for the proof of Theorem 1.1 we write the IBLe (1.6) as

$$U_t = A_0 U + B(U,U), \tag{1.19}$$

where $B(U,V)$ is a symmetric bilinear form representing the quadratic terms in (1.6). For a $\beta > 1$ we set

$$U = \varepsilon\psi + \varepsilon^\beta R, \tag{1.20}$$

and obtain the equation

$$\partial_t R = A_0 R + 2\varepsilon B(\psi, R) + \varepsilon^\beta B(R,R) + \varepsilon^{-\beta}\mathrm{Res}(\varepsilon\psi) \tag{1.21}$$

for the error R, where the so-called residual

$$\mathrm{Res}(\varepsilon\psi) = \varepsilon(-\partial_t\psi + A_0\psi + \varepsilon B(\psi,\psi)) \tag{1.22}$$

contains the terms that do not vanish after inserting (1.12) into (1.6). We essentially have to show (a) that solutions to (1.21) for initial conditions $R_0 = R|_{t=0}$ of order $\mathcal{O}(1)$ exist locally, and (b) that the solutions exist and stay $\mathcal{O}(1)$-bounded up to times $t = T_0/\varepsilon^2$. In order to show (a) for the quasilinear parabolic system (1.21), we use the maximal regularity techniques from [11]. To achieve (b), we first define an improved approximation $\varepsilon\tilde{\psi}$ such that the residual is sufficiently small and then derive an energy estimate similar to (1.9). Note that a priori we would expect a growth rate like $\mathrm{e}^{C\varepsilon t}$ for solutions of (1.21) due to the term $2\varepsilon B(\tilde{\psi}, R)$ in (1.21). Moreover, because of the term $\varepsilon^{-\beta}\mathrm{Res}(\varepsilon\tilde{\psi})$, we would like to choose β small, while in order to handle the nonlinear term $\varepsilon^\beta B(R,R)$, we would like to have β large. The approach turns out to work with $\beta = 3/2$.

In section 2 we give the calculation leading to the energy estimate (1.9) for the linearized problem and define the improved approximation. The proof of Theorem 1.1 follows in section 3. The local existence of solutions to (1.6) is shown in Appendix A, which also yields the local existence of solutions to the error equation. In Appendix B we give a brief review of the underlying physical problem, show how the parabolic IBLe (1.1) can be formally derived from the governing NSe, and end with a brief discussion of the hyperbolic Shkadov model.

2. Preparations.

2.1. The linearized energy estimate. Here we show the straightforward calculations leading to the energy estimate (1.9). In section 3 we extend these to the quasilinear problem (1.21). We fix $c_1 = 9$, $c_2 = 2$, $c_4 = \mathrm{R}^2$ in (1.10) and show how $c_3 = -11\mathrm{R}/5 + 2\cot\theta/3$ yields (1.9). Using

$$|ab| \le \delta a^2 + \frac{1}{4\delta} b^2, \quad \delta > 0, \tag{2.1}$$

it is clear that, for ε sufficiently small, $\|\cdot\|_Y$ is equivalent to $\|\cdot\|_{Y_0}$ with

$$\|(\eta, q)\|_{Y_0}^2 = \frac{1}{2}\int_{\mathbb{R}} q^2 + q_x^2 + \eta^2 + \varepsilon^{-2}(\eta_x^2 + \eta_{xx}^2)\,\mathrm{d}x,$$

and hence a norm on $H^2(\mathbb{R}) \times H^1(\mathbb{R})$. By Fourier transform it is obvious that the solution U of the linearized equation $U_t = A_0 U$, with A_0 from (1.6), exists and is smooth. We then obtain

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\|U\|_Y^2 = \int_{\mathbb{R}} \Big\{ & (q - 2\eta - c_3\eta_x - \mathrm{R}^2 q_{xx})\Big[\frac{6}{\mathrm{R}}\eta + \Big(\frac{4}{5} - \frac{2}{\mathrm{R}}\cot\theta\Big)\eta_x - \frac{3}{\mathrm{R}}\eta_{xx} \\
& \qquad + \varepsilon^{-2}\mathrm{W}\eta_{xxx} - \frac{3}{\mathrm{R}}q - \frac{8}{5}q_x + \frac{7}{2\mathrm{R}}q_{xx}\Big] \\
& - 9\eta q_x + \varepsilon^{-2}\mathrm{W}\big(\eta_{xx}q_x + \mathrm{R}^2\eta_{xxx}q_{xx}\big) - c_3 q_x^2\Big\}\,\mathrm{d}x \\
= \int_{\mathbb{R}} \Big\{ & -\frac{3}{\mathrm{R}}q^2 + \frac{12}{\mathrm{R}}q\eta - \frac{12}{\mathrm{R}}\eta^2 + \Big(\frac{4}{5} - \frac{2}{\mathrm{R}}\cot\theta - \frac{16}{5} + 9 + \frac{3}{\mathrm{R}}c_3\Big)\eta_x q \\
& + \Big(-c_3 - 3\mathrm{R} - \frac{7}{2\mathrm{R}}\Big)q_x^2 + \Big(6\mathrm{R} + \frac{8}{5}c_3 + \frac{10}{\mathrm{R}}\Big)q_x\eta_x \\
& + \Big(-c_3\Big(\tfrac{4}{5} - \tfrac{2}{3}\cot\theta\Big) - \tfrac{6}{\mathrm{R}}\Big)\eta_x^2 + \Big(\mathrm{R}^2\Big(\tfrac{4}{5} - \tfrac{2}{3}\cot\theta\Big) + \tfrac{7c_3}{2\mathrm{R}}\Big)q_x\eta_{xx} \\
& - \frac{7\mathrm{R}}{2}q_{xx}^2 + 3\mathrm{R}q_{xx}\eta_{xx} + \varepsilon^{-2}\mathrm{W}c_3\eta_{xx}^2\Big\}\,\mathrm{d}x.
\end{aligned}$$

The quadratic form in η, q without derivatives is nonpositive. For $c_3 = -11\mathrm{R}/5 + 2\cot\theta/3$ the coefficient of $\eta_x q$ vanishes. Moreover, the coefficients of q_x^2, q_{xx}^2 are negative definite, and the coefficient of η_{xx}^2 is negative definite with strong weight ε^{-2}. Note that $c_3 < 0$ due to (1.2). The terms with η_x, q_x yield $\frac{\mathrm{d}}{\mathrm{d}t}\|U\|_Y^2 \le C_1\|\eta_x\|_{L^2}^2$, but since η_x^2 appears in $\|U\|_Y$ with weight ε^{-2}, we nevertheless obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\|U\|_Y^2 \le C_1\varepsilon^2\|U\|_Y^2 - C\|q_x\|_{H^1}^2. \tag{2.2}$$

On the other hand, the coefficient of $q\eta_x$ has to vanish identically, since we have no negative definite term in q^2 and cannot have one, as is clear from the dispersion relation. Therefore we have to introduce c_3 in (1.10). From (2.2) we get (1.9) using Gronwall's lemma. The dissipation in $\partial_x q$ in (2.2) will be important for the quasilinear problem (1.21).

2.2. The residual. For notational convenience and without loss of generality for our purposes, we assume in the following that we have a vertically falling film such

that $\cot\theta = 0$. Then the critical Reynolds number is $\mathrm{R}_c = 0$, and we may further assume without loss of generality that

$$\mathrm{R} = W = 1. \tag{2.3}$$

In order to get a small residual in (1.21), we define an improved approximation by

$$\varepsilon\tilde{\psi}(t,x) = \begin{pmatrix} \varepsilon\eta_1(T,X) \\ \sum_{j=1}^3 \varepsilon^j q_j(T,X) \end{pmatrix}, \quad T = \varepsilon^2, \quad X = \varepsilon(x-2t). \tag{2.4}$$

Plugging (2.4) into (1.3), (1.4), we first obtain (1.13) and (1.14) as before, and then

$$\begin{aligned} \mathcal{O}(\varepsilon^3(1.4)): \quad q_{1T} - 2q_{2X} &= \frac{8}{5}\eta_{1X}\eta_1 + \frac{7}{2}q_{1XX} + \eta_1\eta_{1XXX} - 3\eta_{1XX} - 3q_3 \\ &\quad + 6\eta_1 q_2 - \frac{8}{5}q_{2X} - \frac{12}{5}q_1 q_{1X} + \frac{8}{5}\eta_1\eta_{1X} \\ \Rightarrow q_3 &= \frac{1}{3}\left(\frac{4}{5}\partial_X^4\eta_1 + 3\eta_{1X}\eta_1 + \frac{52}{25}\eta_{1XX}\eta_1 + \frac{112}{5}\eta_{1X}\eta_1 + 12\eta_1^3\right). \end{aligned} \tag{2.5}$$

With q_3 given by (2.5), all terms up to order $\mathcal{O}(\varepsilon^3)$ vanish in the residual

$$\mathrm{Res}(\varepsilon\tilde{\psi}) = \varepsilon^4 f = \varepsilon^4 \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}. \tag{2.6}$$

To leading order in derivatives we have

$$\begin{aligned} f_1 &= -q_{3X} = -\frac{4}{15}\partial_X^5\eta_1 + \tilde{f}_1, \\ f_2 &= -\varepsilon q_{3T} + \tilde{f}_2 = -\frac{4\varepsilon}{15}\partial_X^4\partial_T\eta_1 + \tilde{f}_2 = \frac{4\varepsilon}{45}\partial_X^8\eta_1 + \tilde{f}_2. \end{aligned} \tag{2.7}$$

Later we need

$$(f_1, f_2) \in C\left(\left[0, \varepsilon^2 T_0\right], H^2(\mathbb{R}) \times H^1(\mathbb{R})\right) \tag{2.8}$$

and therefore

$$\eta_1 \in C([0,T_0], H^9(\mathbb{R})) \tag{2.9}$$

in Theorem 1.1. The (nonlinear) functions $\tilde{f}_{1,2}$ in (2.7) contain lower order derivatives of η_1, and it can be easily checked that (2.8) holds if (2.9) does.

In order to estimate the residual in Y, we finally need to take care of how scaling affects the L^2 norm, i.e., $\|u(\varepsilon\cdot)\|_{L^2} = \varepsilon^{-1/2}\|u(\cdot)\|_{L^2}$. This loss of $\varepsilon^{-1/2}$ is the reason that we cannot choose $\beta = 2$ in (1.20), which would be more convenient in order to control the nonlinear terms in (1.21).

We summarize our results as follows.

LEMMA 2.1. *Assume that $\eta_1 \in C([0,T_0], H^9(\mathbb{R}))$. Then*

$$\sup_{0\le t<T_0/\varepsilon^2} \|\varepsilon\tilde{\psi} - \varepsilon\psi\|_Y \le C\varepsilon^{5/2} \quad \text{and} \quad \sup_{0\le t\le T_0/\varepsilon^2} \|\mathrm{Res}(\varepsilon\tilde{\psi})\|_Y \le C\varepsilon^{7/2}.$$

Due to the first estimate in Lemma 2.1 we can use $\varepsilon\tilde{\psi}$ instead of $\varepsilon\psi$ in the proof of Theorem 1.1, and in order to simplify symbols we drop the $\tilde{}$ in the following. Also we write $\psi = (\psi_1, \psi_2)$ and $\psi_j' = \partial_X\psi_j$.

3. Proof of Theorem 1.1. In the following many constants which are independent of ε and t are denoted by C, and C_1, C_2 are the constants from Theorem 1.1.

From the local existence of solutions to the IBLe in Theorem A.1 we directly obtain the following local existence of solutions to (1.21).

COROLLARY 3.1. *Let $R_0 \in H^2(\mathbb{R}) \times H^1(\mathbb{R})$ and $0 < t_1 \le T_0/\varepsilon^2$. Then there exists an $\varepsilon_1 > 0$ such that for all $\varepsilon \in (0, \varepsilon_1)$ there exists a unique solution $R \in H^{3,3/2}((0,t_1)\times\mathbb{R}) \times H^{2,1}((0,t_1)\times\mathbb{R})$ of the error equation (1.21) with $R(0) = R_0$.*

Proof. For ε_1 sufficiently small we have

$$\|U_0\|_{H^2\times H^1} = \|(\varepsilon\psi + \varepsilon^{3/2}R)|_{t=0}\|_{H^2\times H^1} \le \rho$$

for all $\varepsilon \in (0, \varepsilon_1)$, with $\rho > 0$ from Theorem A.1. Therefore there exists a unique solution $U \in H^{3,3/2}((0,t_1)\times\mathbb{R}) \times H^{2,1}((0,t_1)\times\mathbb{R})$ of (1.6). Using the smoothness of η_1 we find that the solution $R = \varepsilon^{-3/2}(U - \varepsilon\psi)$ of (1.21) has the same regularity. ∎

The proof of Theorem 1.1 now works as follows: due to Corollary 3.1 we have a local solution $R \in C([0,t_1), H^2 \times H^1)$ of (1.21). Thus we may choose t_1 so small that $\sup_{0\le t\le t_1} \|R\|_Y \le 2\|R_0\|_Y \le 2C_1$. This implies

$$\sup_{0\le t\le t_1} (\|r\|_\infty + \|\xi\|_\infty + \varepsilon^{-1}\|\partial_x\xi\|_\infty) \le 2CC_1, \tag{3.1}$$

where $C > 0$ comes from Sobolev embedding. Using (3.1), we derive an energy estimate that implies $\|R(t_1)\|_Y \le \mathrm{e}^{C\varepsilon^2 t_1}\|R_0\|_Y$. Thus, using Corollary 3.1 again, the solution can be continued and stays $\mathcal{O}(1)$-bounded in Y until $t_1 = t_0 = T_0/\varepsilon^2$.

It will be convenient to write (1.21) as

$$R_t = A(t,R)R + \varepsilon^2 f, \tag{3.2}$$

where, with $a_0, \dots, b_1$ from (1.5),

$$R = \begin{pmatrix}\xi\\ r\end{pmatrix}, \quad A(t,R) = \begin{pmatrix} 0 & -\partial_x \\ \tilde a_0 + \tilde a_1\partial_x + \tilde a_2\partial_x^2 + \varepsilon^{-2}\tilde a_3\partial_x^3 & -\tilde b_0 - \tilde b_1\partial_x + \tilde b_2\partial_x^2 \end{pmatrix},$$

$$\tilde a_0 = \tilde a_0(t,\xi) = a_0(\varepsilon\psi_1 + \varepsilon^{3/2}\xi) - \frac{8}{5}\varepsilon^2\psi_1' + \frac{9}{2}\varepsilon^3\psi_1'' + \varepsilon^4\psi_1''' + 6\varepsilon\psi_2 + \frac{8}{5}\varepsilon^2\psi_2' + 3\varepsilon^3\psi_2''',$$
$$\tilde a_1 = \tilde a_1(t,R) = a_1(\varepsilon\psi + \varepsilon^{3/2}R) - 2\varepsilon^2\psi_1', \quad \tilde a_2 = \tilde a_2(t,\xi) = a_2(\varepsilon\psi_1 + \varepsilon^{3/2}\xi),$$
$$\tilde a_3 = \tilde a_3(t,\xi) = a_3(\varepsilon\psi_1 + \varepsilon^{3/2}\xi),$$
$$\tilde b_0 = \tilde b_0(t,\xi) = b_0(\varepsilon\psi_1 + \varepsilon^{3/2}\xi) - \frac{8}{5}\varepsilon^2\psi_1' + \frac{12}{5}\varepsilon^2\psi_2' + \frac{9}{2}\varepsilon^3\psi_1'',$$
$$\tilde b_1 = \tilde b_1(t,R) = b_1(\varepsilon\psi + \varepsilon^{3/2}R) - \varepsilon^2\psi_1', \quad \tilde b_2 = b_2 = \frac{7}{2}.$$

The main idea for obtaining the energy estimate is to define an equivalent norm $N_Y(R,t)$ on Y that depends on time and the solution itself in such a way that high order and strongly weighted mixed products like $\varepsilon^{-2}\partial_x^2 q\partial_x^3\eta$ still cancel after integration by parts in $\frac{\mathrm{d}}{\mathrm{d}t}\|R\|^2_{N_Y(R,t)}$. This can be achieved by dividing all terms in (1.10) involving r by $\tilde a_3$. Moreover, we need correction terms that eliminate terms of order $\mathcal{O}(\varepsilon)$ and $\mathcal{O}(\varepsilon^{3/2})$ in $\frac{\mathrm{d}}{\mathrm{d}t}\|R\|_{N(t,R)}$ without derivatives that come from $2\varepsilon B(\tilde\psi, R) + \varepsilon^{3/2}B(R,R)$ in (1.21).

Thus, with coefficients $\gamma_1, \dots, \gamma_4 \in \mathbb{R}$ to be determined, we define

$$\|R\|^2_{N_Y(t,R)} = E + F_1 + F_2, \tag{3.3}$$

$$E = \frac{1}{2}\int_{\mathbb{R}} \left\{ \frac{1}{\tilde a_3}\left[r^2 - 4r\xi - 2c_3 r\xi_x + r_x^2\right] + 9\xi^2 + \varepsilon^{-2}\left[\xi_x^2 + \xi_{xx}^2\right] \right\} \mathrm{d}x,$$

$$F_1 = \int \frac{1}{\tilde a_3}\varepsilon\eta_1\left[\gamma_1 r^2 + \gamma_2 \xi r\right] \mathrm{d}x, \quad F_2 = \int \frac{1}{\tilde a_3}\varepsilon^{3/2} r\left[\gamma_3 r\xi + \gamma_4 \xi^2\right] \mathrm{d}x,$$

where for notational convenience we keep writing c_3 for $-11/5$. Due to (3.1) we have

$$1 - C\varepsilon \le \sup_{0\le t\le t_0, x\in\mathbb{R}} |a_3| = \sup_{0\le t\le t_0, x\in\mathbb{R}} |1 + \varepsilon\psi_1 + \varepsilon^{3/2}\xi| \le 1 + C\varepsilon. \tag{3.4}$$

Therefore $N_Y(t,R)$ is still an equivalent norm on Y if ε is sufficiently small. Moreover, space and time derivatives of $\tilde a_0, \tilde a_1, \dots, \tilde b_2$ produce terms of order $\mathcal{O}(\varepsilon^{3/2})$, and in particular we have

$$\frac{\mathrm{d}}{\mathrm{d}x}\tilde a_0 = 6\varepsilon^{3/2}\xi_x + \varepsilon^2 h_1, \quad \frac{\mathrm{d}}{\mathrm{d}x}\tilde b_0 = 3\varepsilon^{3/2}\xi_x + \varepsilon^2 h_2, \tag{3.5}$$

$$\frac{\mathrm{d}}{\mathrm{d}x}\tilde a_3 = \varepsilon^{3/2}\xi_x + \varepsilon^2 h_3, \quad \frac{\mathrm{d}}{\mathrm{d}t}\tilde a_3 = \varepsilon^{3/2}\xi_t + \varepsilon\frac{\mathrm{d}}{\mathrm{d}t}\psi_1 = -\varepsilon^{3/2} r_x + h_4, \tag{3.6}$$

with $\|h_j\|_{L^\infty} = \mathcal{O}(\varepsilon^2)$, $j = 1,2,3,4$. Hence we can estimate terms like, for instance, $(\frac{\mathrm{d}}{\mathrm{d}x}\frac{\tilde b_0}{\tilde a_3})r\xi$ that show up during integration by parts in $\frac{\mathrm{d}}{\mathrm{d}t}\|R\|^2_{N(t,R)}$ as

$$\begin{aligned}\int \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\tilde b_0}{\tilde a_3}\right) r\xi \,\mathrm{d}x &= \int \frac{1}{3\tilde a_3}(\varepsilon^{3/2}\xi_x r\xi) - 6\frac{\tilde b_0}{\tilde a_3^2}(\varepsilon^{3/2} r_x r\xi) + \mathcal{O}(\varepsilon^2)|r\xi| \,\mathrm{d}x \\ &\le C\|\xi\|_\infty \int \varepsilon\xi_x^2 + \varepsilon r_x^2 + \varepsilon^2(\xi^2 + r^2)\,\mathrm{d}x + C\varepsilon^2 \int r^2 + \xi^2 \,\mathrm{d}x,\end{aligned} \tag{3.7}$$

and similarly for $(\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{\tilde a_3})(r^2 - 4r\xi)$; see (3.10). The first term on the right-hand side of (3.7) is estimated by $\varepsilon^3\varepsilon^{-2}\xi_x^2$, and the second term is well behaved since we will have an $\mathcal{O}(1)$ negative definite term $-Cr_x^2$ in $\frac{\mathrm{d}}{\mathrm{d}t}\|R\|^2_{N_Y(t,R)}$. This is essentially the first reason why the estimate (2.2) can be carried over to the quasilinear problem (3.2). The second reason is that the coefficients $\gamma_1, \dots, \gamma_4$ can be chosen in such a way that the terms r^2, ξ^2 without derivatives in $\frac{\mathrm{d}}{\mathrm{d}t}\|R\|^2_{N_Y(t,R)}$ have $\mathcal{O}(\varepsilon^2)$ coefficients. This is possible again due to the fact that the small parameter ε does not as usual enter (1.6) as a coefficient of the low order terms but in inverse power as coefficient of the high order damping term.

We start with $\frac{\mathrm{d}}{\mathrm{d}t}E$. Using $2\int g f_x f \,\mathrm{d}x = -\int g_x f^2 \,\mathrm{d}x$, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}E = d_1 + d_2 + d_3 + d_4, \tag{3.8}$$

$$\begin{aligned} d_1 = \int \Bigg\{ &\left[r - 2\xi \; - c_3\xi_{xx} - r_{xx} - \tilde a_3 \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{1}{\tilde a_3}\right) r_x\right] \\ &\left[\frac{\tilde a_0}{\tilde a_3}\xi + \frac{\tilde a_1}{\tilde a_3}\xi_x + \frac{\tilde a_2}{\tilde a_3}\xi_{xx} + \varepsilon^{-2}\xi_{xxx} - \frac{\tilde b_0}{\tilde a_3} r - \frac{\tilde b_1}{\tilde a_3} r_x + \frac{\tilde b_2}{\tilde a_3} r_{xx} + \varepsilon^2 \frac{f_2}{\tilde a_3}\right] \Bigg\} \mathrm{d}x, \\ d_2 = \int &-9\xi r_x + \varepsilon^{-2}(\xi_{xx} r_x + \xi_{xxx} r_{xx}) + 9\xi\varepsilon^2 f_1 + \varepsilon^{-2}\varepsilon^2\left(\xi_x f_{1x} + \xi_{xx} f_{1xx}\right) \mathrm{d}x, \\ d_3 = \int \frac{1}{\tilde a_3}\left[2r r_x + c_3 r r_{xx}\right] \mathrm{d}x, \quad &d_4 = \int \left(\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{\tilde a_3}\right)\left[r^2 - 4r\xi - 2c_3 r\xi_x + r_x^2\right] \mathrm{d}x. \end{aligned}$$

Integration by parts yields

(3.9)
$$\begin{aligned} d_1+d_2+d_3 = \int\Big\{ & \frac{1}{\tilde a_3}\big[-\tilde b_0 r^2+(\tilde a_0+2\tilde b_0)r\xi-2\tilde a_0\xi^2\big]+\mathcal{O}(\varepsilon^2)(r^2+\xi^2) \\ & + \frac{1}{\tilde a_3}\big[\tilde a_1-2\tilde b_1+9+c_3\tilde b_0+\mathcal{O}(\varepsilon)\big]r\xi_x+\frac{1}{\tilde a_3}\big[-c_3-\tilde b_0-\tilde b_2+\mathcal{O}(\varepsilon)\big]r_x^2 \\ & + \frac{1}{\tilde a_3}\big[\tilde a_0+c_3\tilde b_1-\tilde a_2+2\tilde b_2+\mathcal{O}(\varepsilon)\big]r_x\xi_x+\frac{1}{\tilde a_3}\big[-c_3\tilde a_1+2\tilde a_2+\mathcal{O}(\varepsilon)\big]\xi_x^2 \\ & + \big[c_3\varepsilon^{-2}+\mathcal{O}(\varepsilon)\big]\xi_{xx}^2+\frac{1}{\tilde a_3}\big[\tilde a_1+c_3\tilde b_2+\mathcal{O}(\varepsilon)\big]r_x\xi_{xx}-\frac{\tilde b_2}{\tilde a_3}r_{xx}^2-\frac{2\tilde a_2}{\tilde a_3}r_{xx}\xi_{xx} \\ & + \varepsilon^2 f_2[r-2\xi-c_3\xi_{xx}-r_{xx}+\mathcal{O}(\varepsilon)r_x]+9\varepsilon^2 f_1\xi+f_{1x}\xi_x+f_{1xx}\xi_{xx}\Big\}\,\mathrm{d}x, \end{aligned}$$

where the order symbol $\mathcal{O}(\varepsilon)$ always refers to terms estimated in L^∞. The coefficient of $r\xi_x$ in (3.9) is $\mathcal{O}(\varepsilon)$ due to the choice of c_3 and (3.1). Similarly to (3.7), d_4 can be estimated as

$$\begin{aligned} d_4 \le{}& C\|r\|_\infty\int\varepsilon^{3/2}(|r_x r|+|r_x\xi|)\,\mathrm{d}x+C\varepsilon^2\int r^2+\xi^2\,\mathrm{d}x \\ & + C\varepsilon^{3/2}\|r_x\|_\infty\int\big[-2c_3 r\xi_x+r_x^2\big]\,\mathrm{d}x, \end{aligned} \tag{3.10}$$

and therefore (3.9) and (3.10), except for the first term on the right-hand side of (3.9), can be estimated by $C\varepsilon^2 E+C_{\mathrm{res}}\varepsilon^2$.

Thus, we now have

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\|R\|_{N_Y(t,R)}^2 &= \frac{\mathrm{d}}{\mathrm{d}t}E+\frac{\mathrm{d}}{\mathrm{d}t}F_1+\frac{\mathrm{d}}{\mathrm{d}t}F_2 \\ &\le \int\frac{1}{\tilde a_3}\big[-\tilde b_0 r^2+(\tilde a_0+2\tilde b_0)r\xi-2\tilde a_0\xi^2\big]\,\mathrm{d}x+\frac{\mathrm{d}}{\mathrm{d}t}F_1+\frac{\mathrm{d}}{\mathrm{d}t}F_2+C\varepsilon^2E+\varepsilon^2C_{\mathrm{res}}. \end{aligned} \tag{3.11}$$

To control the first three terms on the right-hand side of (3.11), we calculate

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}F_1 &= \int\frac{\varepsilon\eta_1}{\tilde a_3}\big[2\gamma_1 r(6\xi-3r)+\gamma_2\xi(6\xi-3r)\big]\,\mathrm{d}x+h_1, \\ \frac{\mathrm{d}}{\mathrm{d}t}F_2 &= \int\frac{\varepsilon^{3/2}\xi}{\tilde a_3}\big[2\gamma_2 r(6\xi-3r)+\gamma_4\xi(6\xi-3r)\big]\,\mathrm{d}x+h_2, \end{aligned}$$

where h_1 and h_2 contain terms like, for instance, $h_1=-\varepsilon\eta_1 r_x r^2/\tilde a_3+\cdots$, that can be controlled by the negative definite terms in $\frac{\mathrm{d}}{\mathrm{d}t}E$ as in (3.7) and (3.10). Since

$$\tilde b_0 = 3-6\varepsilon\eta_1-6\varepsilon^{3/2}\xi+\mathcal{O}(\varepsilon^2),\qquad \tilde a_0 = 6-18\varepsilon\eta_1-6\varepsilon^{3/2}\xi+\mathcal{O}(\varepsilon^2), \tag{3.12}$$

we thus obtain

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\|R\|_{N_Y(t,R)}^2 \le{}& \int\frac{1}{\tilde a_3}\Big\{\big(-3+6\varepsilon(1-\gamma_1)\eta_1+6\varepsilon^{3/2}\xi(1-\gamma_3)\varepsilon^{3/2}\xi\big)r^2 \\ & + \big(12-3(10-4\gamma_1+\gamma_2)\varepsilon\eta_1-3(6-4\gamma_3+\gamma_4)\varepsilon^{3/2}\xi\big)r\xi \\ & - \big(12-6(6+\gamma_2)\varepsilon\eta_1-6(2+\gamma_4)\varepsilon^{3/2}\xi\big)\xi^2\Big\}\,\mathrm{d}x \\ & + C\varepsilon^2E+C_{\mathrm{res}}\varepsilon^2. \end{aligned} \tag{3.13}$$

Choosing $\gamma_1 = 1, \gamma_2 = -6, \gamma_3 = 1, \gamma_4 = -2$, the $\mathcal{O}(\varepsilon)$ and $\mathcal{O}(\varepsilon^{3/2})$ coefficients in the integral vanish, and since $\int \frac{1}{\bar{a}_3}\left[-3r^2 + 12r\xi - 12\xi^2\right] \mathrm{d}x \le 0$, we finally obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\|R\|^2_{N_Y(t,R)} \le C\varepsilon^2 E + C_{\mathrm{res}}\varepsilon^2 \le C\varepsilon^2 \|R\|^2_{N_Y(t,R)} + C_{\mathrm{res}}\varepsilon^2. \tag{3.14}$$

This gives $\|R\|^2_{N_Y(t,R)} \le \mathrm{e}^{C\varepsilon^2 t}\|R|_{t=0}\|^2_{N_Y(t,R)} + C(\mathrm{e}^{C\varepsilon^{2t}-1})$, using Gronwall's lemma. The proof of Theorem 1.1 is now complete. □

Remark 3.2. Theorem 1.1 also holds in higher order Sobolev spaces. For $m \ge 3$ we can define $Y_m = H^m(\mathbb{R}) \times H^{m-1}(\mathbb{R})$, with $\|\cdot\|_{Y_m}$ defined in a way similar to $\|\cdot\|_Y$; i.e., for $\cot\theta = 0$, R = W = 1,

$$\|U\|^2_{Y_3} = \frac{1}{2}\int_{\mathbb{R}} \Big\{ q^2 + 9\eta^2 - 4q\eta - \frac{22}{5}\eta_x q + 2q_x\eta_{xx} + q_x^2 + q_{xx}^2 + 2q_x\eta_{xx} + \varepsilon^{-2}(\eta_x^2 + \eta_{xx}^2 + \eta_{xxx}^2) \Big\} \mathrm{d}x.$$

Then for $\|U(0,\cdot) - \varepsilon\psi(0,\cdot)\|_{Y_m} \le C_1\varepsilon^{3/2}$ and $\eta_1 \in H^{m+6}(\mathbb{R})$ we obtain a solution $U \in H^{m+1,(m+1)/2} \times H^{m,m/2}$ with $\sup_{0\le t\le t_0}\|U(t,\cdot) - \varepsilon\psi(t,\cdot)\|_{Y_m} \le C_2\varepsilon^{3/2}$. The local existence of solutions in these higher order spaces is already shown in Theorem A.1, and from the above proof of Theorem 1.1 it can be seen that the high order terms are uncritical in the energy estimates.

Appendix A. Local existence of solutions for the IBLe. To treat the initial value problem for the IBLe (1.6), we use the spaces

$$H^{r,s} = H^{r,s}((0,t_0)\times\mathbb{R}) = L^2((0,t_0), H^r(\mathbb{R})) \cap H^s((0,t_0), L^2(\mathbb{R})),$$

defined for $r, s \ge 0$. Because we have a parabolic system, we will always have $s = r/2$, and therefore we introduce the notation $K^r = K^r((0,t_0)\times\mathbb{R}) = H^{r,r/2}((0,t_0)\times\mathbb{R})$. We recall a few facts on the spaces $H^{r,r/2}((0,t_0)\times\mathbb{R})$, mainly from [11].

If $u \in H^{r,s}$ and $j,k \in \mathbb{N}$ with $1 - (j/r + k/s) \ge 0$, then $\partial_t^k\partial_x^j u \in H^{\mu,\nu}$ with $\mu/r = \nu/s = 1-(j/r+k/s)$; see [11, Proposition 4.2.3]. Especially, if $u \in K^r$ and $1-(j/r+2k/r) \ge 0$, then $\partial_t^k\partial_x^j u \in K^{r-j-2k}$. For $k < r/2 - 1/2$ we have traces $\partial_t^k u(0,\cdot) \in H^{r-2k-1}(\mathbb{R})$; see [11, Proposition 4.2.1]. Conversely, if these traces are given at $t = 0$, then there exists a bounded extension operator such that $u \in K^r$; see [11, Theorem 4.2.3]. Similarly, there exists a bounded extension operator from $K^r = K^r((0,t_0)\times\mathbb{R})$ into $K^r(\mathbb{R}\times\mathbb{R})$; see [24, Lemma 3.1].

For $u \in K^r(\mathbb{R}\times\mathbb{R}^n)$ let $\hat{\hat{u}}(\tau,k) = \iint \mathrm{e}^{-\mathrm{i}(\tau t + k\cdot x)}u(t,x)\,\mathrm{d}k\,\mathrm{d}t$ be the Fourier transform in time and space of u. Then we have the equivalence of norms

$$\|u\|^2_{K^r(\mathbb{R}\times\mathbb{R}^n)} \sim \iint |\hat{\hat{u}}(\tau,\xi)|^2(1+|k|^2+|\tau|)^r\,\mathrm{d}k\,\mathrm{d}\tau. \tag{A.1}$$

From this it follows easily that if $u \in K^r(\mathbb{R}\times\mathbb{R}^n)$ with $r > (n+2)/2$, then u is bounded and continuous. Finally, we need the special subspace

$$K_0^r = K_0^r((0,t_0)\times\mathbb{R}) = \{u \in K^r((0,t_0)\times\mathbb{R}) : \partial_t^k u(0,\cdot) = 0 \text{ for } k\in\mathbb{N}, k < r/2 - 1/2\}.$$

For $u \in K_0^r((0,\infty)\times\mathbb{R})$ the continuation by $u(t) = 0$ for $t < 0$ is in $K^r(\mathbb{R}\times\mathbb{R})$; see [11, Theorem 1.11.5]. In addition to the full space-time transform of $u \in K^r$ we also

use the Fourier transform in time only, denoted by $\hat{u}(\tau,x)=\int \mathrm{e}^{-\mathrm{i}\tau t}u(t,x)\,\mathrm{d}t$. For $u\in K_0^r((0,\infty)\times\mathbb{R})$ we then obtain the equivalence

$$\|u\|^2_{K^r((0,\infty)\times\mathbb{R})}\sim\int\|\hat{u}(\tau,\cdot)\|^2_{H^r}+|\tau|^r\|\hat{u}(\tau,\cdot)\|^2_{L^2}\,\mathrm{d}\tau. \tag{A.2}$$

We introduce the shorthand $\mathcal{K}^r=K^r\times K^{r-1}$. Also, in this section we write $|u|_r$ for the Sobolev norm in the spacial variable x (or its dual k), i.e., $|u|_r{=}\|u\|_{H^r(\mathbb{R})}$, and, e.g., $|\hat{u}(1{+}k^2)|_0$ for the L^2 norm of the function $k\mapsto\hat{u}(1{+}k^2)$. In the proof of Theorem 1.1 we use the following local existence theorem for the solutions of the IBLe (1.6) with $r=2$; however, here we state a more general case.

THEOREM A.1. *Let $2\le r<4$, $\varepsilon>0$, and $t_0>0$ be fixed. Then there exists a $\rho>0$ such that for all $U_0=(\eta_0,q_0)\in H^r(\mathbb{R})\times H^{r-1}(\mathbb{R})$ with $|U_0|_{H^r\times H^{r-1}}\le\rho$ there exists a unique solution*

$$U=(\eta,q)\in\mathcal{K}^{r+1} \tag{A.3}$$

of the IBLe (1.6) with $U|_{t=0}=U_0$ and $\|U\|_{\mathcal{K}^{r+1}}\le C|U_0|_{H^r\times H^{r-1}}$, where the constant $C>0$ depends only on ε and t_0. Moreover, for all $0<t_1<t_0$ and all $k>0$ we have $U\in\mathcal{K}^{r+k+1}((t_1,t_0)\times\mathbb{R})$; i.e., U is smooth for $t>0$.

Remark A.2. Examining the proof of Theorem A.1, we obtain that ρ may be chosen independently of $\varepsilon\in(0,\varepsilon_0)$. Theorem A.1 is used in Corollary 3.1 in this sense, but for simplicity we do not keep track of this here. Also, the upper bound $r<4$ is only for notational convenience, i.e., to avoid the formulation of higher order trace conditions at $t=0$; see (A.12).

The proof of Theorem A.1 consists of two steps. First we consider the linear inhomogeneous version of (1.6) with zero initial data, i.e., the equation

$$LU=F(t),\quad U(0)=0,\quad LU=U_t-A_0U,\quad F\in\mathcal{K}_0^{r-1}, \tag{A.4}$$

and estimate its solutions in $\mathcal{K}_0^{r+1}$. Then we write the solution U of (1.6) as $U=\tilde{U}{+}U^{(1)}$, where $\tilde{U}\in\mathcal{K}^{r+1}$ fulfills $\tilde{U}(0)=U_0$ and some (further) trace conditions at $t=0$; see (A.12). Then $U^{(1)}$ has to solve the equation

$$LU^{(1)}=G(U^{(1)}),\quad U^{(1)}(0)=0,\quad G(U^{(1)})=F(\tilde{U}+U^{(1)})-L\tilde{U}. \tag{A.5}$$

We show that for $U^{(1)}\in\mathcal{K}_0^{r+1}$ we have $G(U^{(1)})\in\mathcal{K}_0^{r-1}$, and use the estimates for (A.4), estimates for the nonlinearity, and the contraction mapping theorem to solve (A.5).

LEMMA A.3. *Let $r\ge2$, $\varepsilon>0$, and $t_0>0$. For every $F\in\mathcal{K}_0^{r-1}$ there exists a unique solution $U\in\mathcal{K}_0^{r+1}$ of (A.4) with $\|U\|_{\mathcal{K}^{r+1}}\le C\|F\|_{\mathcal{K}^{r-1}}$, where $C>0$ depends only on ε,t_0.*

Proof. We identify F with its extension to $\mathcal{K}^{r-1}(\mathbb{R}\times\mathbb{R})$, with $F(t)=0$ for $t\le0$. Then $\mathrm{e}^{-\sigma t}F\in L^1(H^{r-1})\cap L^2(H^{r-1})$ for $\mathrm{Re}\sigma>0$, and therefore $\hat{F}(\tau)$ has an analytic extension into $\mathrm{Im}\tau<0$. We write $\lambda=\sigma+\mathrm{i}\tau$ and consider the Fourier transform in t (i.e., the Laplace transform) of (A.4),

$$\begin{aligned}&\lambda\hat{\eta}=-\hat{q}_x+\hat{f}_1\quad\Leftrightarrow\quad\hat{\eta}=\frac{-\hat{q}_x+\hat{f}_1}{\lambda},\\&\lambda\hat{q}=\frac{1}{\lambda}\big(a_{00}{+}a_{10}\partial_x{+}a_{20}\partial_x^2{+}\varepsilon^{-2}a_{30}\partial_x^3\big)(-\hat{q}_x{+}\hat{f}_1){-}b_{00}\hat{q}{-}b_{10}\hat{q}_x{+}b_{20}\hat{q}_{xx}{+}\hat{f}_2.\end{aligned} \tag{A.6}$$

Now choose $\sigma_0 > 0$ such that $\mathrm{Re}\mu_{1,2}(k) < \sigma_0$ for all $k \in \mathbb{R}$. For $\mathrm{Re}\lambda = \sigma > \sigma_0$ we obtain

$$|\hat\eta|_{r+1} + |\lambda|^{(r+1)/2}|\hat\eta|_0 \le C\big(|\hat f_1|_{r-1} + |\lambda|^{(r-1)/2}|\hat f_1|_0 + |\hat f_2|_{r-2} + |\lambda|^{(r-2)/2}|\hat f_2|_0\big), \tag{A.7}$$

$$|\hat q|_r + |\lambda|^{r/2}|\hat q|_0 \le C\big(|\hat f_1|_{r-1} + |\lambda|^{(r-1)/2}|\hat f_1|_0 + |\hat f_2|_{r-2} + |\lambda|^{(r-2)/2}|\hat f_2|_0\big); \tag{A.8}$$

see below. Moreover, since $\hat F$ is analytic in λ, so is $\hat U = (\hat\eta, \hat q)$ for $\mathrm{Re}\lambda > \sigma_0$. Let

$$U(t) = \frac{1}{2\pi}\int \mathrm{e}^{\sigma_0 t}\mathrm{e}^{\mathrm{i}\tau t}\hat U(\sigma_0 + \mathrm{i}\tau)\,\mathrm{d}\tau.$$

Then $\mathrm{e}^{-\sigma_0 t}U$ is the inverse Fourier transform of the function $\tilde\lambda \mapsto \hat U(\sigma_0 + \tilde\lambda)$, which is analytic for $\mathrm{Re}\tilde\lambda > 0$. Thus, by the Paley–Wiener theorem [26, Theorem 6.4.2], we have $U(t) = 0$ for $t < 0$, and from (A.2), (A.7), and (A.8) we obtain $\mathrm{e}^{-\sigma_0 t}U \in \mathcal{K}_0^{r+1}(\mathbb{R}_+ \times \mathbb{R})$. Since t_0 is finite, we thus have $U \in \mathcal{K}_0^{r+1} = \mathcal{K}_0^{r+1}((0,t_0)\times\mathbb{R})$ with $\|U\|_{\mathcal{K}^{r+1}} \le C\|F\|_{\mathcal{K}^{r-1}}$, where C obviously depends only on t_0 and σ_0, and hence on t_0 and ε.

It remains to show (A.7), (A.8). This is essentially a direct consequence of the parabolic shape of the spectrum. After Fourier transform in x and sorting terms, (A.6) becomes

$$\hat{\hat\eta} = \frac{-\mathrm{i}k\hat{\hat q} + \hat{\hat f}_1}{\lambda}, \qquad g(\lambda,k)\hat{\hat q} = g_0(k)\hat{\hat f}_1 + \lambda\hat{\hat f}_2, \tag{A.9}$$

where

$$g(\lambda,k) = \lambda^2 + \lambda g_1(k) + \mathrm{i}k g_0(k), \quad g_1(k) = b_{20}k^2 + b_{10}\mathrm{i}k + b_{00},$$
$$g_0(k) = a_{00} + a_{10}\mathrm{i}k - a_{20}k^2 - a_{30}\mathrm{i}k^3.$$

Since $g(\lambda,k) = \det(\lambda\mathrm{Id} - A_0(\mathrm{i}k)) = (\lambda-\mu_1(k))(\lambda-\mu_2(k))$, with $\mu_{1,2}$ from (1.7), we have

$$|g(\lambda,k)| \ge C(|\lambda|^2 + (1+k^2)^2).$$

Thus we can estimate

$$\begin{aligned}
|\hat q|_r &\le C|\hat{\hat q}(1+k^2)^{r/2}|_0 \le C\left(\left|\frac{\hat{\hat f}_1 g_0(k)(1+k^2)^{r/2}}{g(\lambda,k)}\right|_0 + \left|\frac{\hat{\hat f}_2\lambda(1+k^2)^{r/2}}{g(\lambda,k)}\right|_0\right)\\
&\le C\left(|\hat{\hat f}_1(1+k^2)^{(r-1)/2}|_0 + |\hat{\hat f}_2(1+k^2)^{(r-2)/2}|_0\right) \le C(|\hat f_1|_{r-1} + |\hat f_2|_{r-2}),
\end{aligned}$$

$$|\lambda|^{r/2}|\hat q|_0 \le C|\lambda|^{r/2}\left|\frac{\hat{\hat f}_1 g_0(k)}{g(\lambda,k)}\right|_0 + \left|\frac{\hat{\hat f}_2\lambda}{g(\lambda,k)}\right|_0 \le C|\lambda|^{(r-2)/2}(|\hat{\hat f}_1|_0 + |\hat{\hat f}_2|_0),$$

$$\begin{aligned}
|\eta|_{r+1} &\le C\left|\frac{(-\mathrm{i}k\hat{\hat q} + \hat{\hat f}_1)(1+k^2)^{(r+1)/2}}{\lambda}\right|_0\\
&= C\left|\frac{(-\mathrm{i}k g_0(k) + g(\lambda,k))\hat{\hat f}_1 - \mathrm{i}k\lambda\hat{\hat f}_2}{\lambda g(\lambda,k)}(1+k^2)^{(r+1)/2}\right|_0\\
&= C\left|\frac{(\lambda + g_1(k))\hat{\hat f}_1 - \mathrm{i}k\hat{\hat f}_2}{g(\lambda,k)}(1+k^2)^{(r+1)/2}\right|_0 \le C(|\hat f_1|_{r-1} + |\hat f_2|_{r-2}),
\end{aligned}$$

$$|\lambda|^{(r+1)/2}|\eta|_{r+1} \le C\left|\frac{(-\mathrm{i}k\hat{\hat q} + \hat{\hat f}_1)}{\lambda}\right|_0 \le C(|\lambda|^{(r-1)/2}|\hat{\hat f}_1|_0 + |\lambda|^{(r-2)/2}|\hat{\hat f}_2|_0).$$

Here we have used the typical parabolic splitting of the domain, for instance,

$$\left|\frac{\hat{\hat{f}}_2 k}{g(\lambda,k)}\right|_0^2 = \left|\frac{\hat{\hat{f}}_2 k}{g(\lambda,k)}\right|^2_{L^2(\{k^2\le|\lambda|\})} + \left|\frac{\hat{\hat{f}}_2 k}{g(\lambda,k)}\right|^2_{L^2(\{k^2\ge|\lambda|\})}$$
$$\le \left|\frac{k\hat{\hat{f}}_2}{|\lambda|^2}\right|^2_{L^2(\{k^2\le|\lambda|\})} + \left||\lambda|^{-3/2}\frac{\hat{\hat{f}}_2 k|\lambda|^{3/2}}{(|\lambda|^2+(1+k^2)^2)}\right|^2_{L^2(\{k^2\ge|\lambda|\})} \le C|\lambda|^{-3}|\hat{f}_2|_0^2.$$

The proof of Lemma A.3 is now complete. ∎

The nonlinear terms in (A.5) can be controlled using the following result, the proof of which follows via extension from $\|uv\|_{K^s(\mathbb{R}\times\mathbb{R}^n)} \le C\|u\|_{K^r(\mathbb{R}\times\mathbb{R}^n)}\,\|v\|_{K^s(\mathbb{R}\times\mathbb{R}^n)}$ if $r > (n+2)/2$.

LEMMA A.4. *Let $r > 3/2$, $0 \le s \le r$. Then there exists a $C > 0$ such that for all $u \in K^r, v \in K^s$ we have $uv \in K^s$ and*

$$\|uv\|_{K^s} \le C\|u\|_{K^r}\|v\|_{K^s}. \tag{A.10}$$

Proof of Theorem A.1. Lemma A.4 applied to F gives

$$\begin{aligned}\|F(U)\|_{\mathcal{K}^{r-1}} &\le C\|U\|^2_{\mathcal{K}^{r+1}},\\ \|F(U)-F(V)\|_{\mathcal{K}^{r-1}} &\le C\|U-V\|_{\mathcal{K}^{r+1}}(\|U\|_{\mathcal{K}^{r-1}} + \|V\|_{\mathcal{K}^{r-1}}).\end{aligned} \tag{A.11}$$

Due to [11, Theorem 4.2.3] there exists an extension $\tilde{U} \in \mathcal{K}^{r+1}$ of $U_0 \in H^r \times H^{r-1}$. We have to choose $\tilde{U} = (\tilde{\eta}, \tilde{q})$ in such a way, that for $U^{(1)} \in \mathcal{K}_0^{r+1}$ the right-hand side $G = F(\tilde{U} + U^{(1)}) - L\tilde{U}$ of (A.5) is in $\mathcal{K}_0^{r-1}$, i.e.,

$$\begin{aligned}\partial_t^k G_1|_{t=0} &= 0 \quad \text{for} \quad 0 \le k < \frac{r}{2} - 1,\\ \partial_t^k G_2|_{t=0} &= 0 \quad \text{for} \quad 0 \le k < \frac{r}{2} - \frac{3}{2}.\end{aligned} \tag{A.12}$$

For $r = 2$ these conditions are trivially true. For $2 < r \le 3$, again due to [11, Theorem 4.2.3], we may choose $\tilde{\eta}$ in such a way that $\partial_t\tilde{\eta}|_{t=0} = -\partial_x q_0 \in H^{r-2}(\mathbb{R})$. Similarly, for $3 < r \le 4$ we additionally choose $\tilde{q}$ such that $\partial_t\tilde{U}|_{t=0} = A(U_0)U_0 \in H^{r-2} \times H^{r-3}$. Thus, in each case, $L\tilde{U} = F(\tilde{U}) = F(\tilde{U} + U^{(1)})$ at $t = 0$, and so $G \in \mathcal{K}_0^{r-1}$.

Thus, we finally consider the mapping

$$\Phi(U^{(1)}) = L_0^{-1}(F(\tilde{U} + U^{(1)}) - L\tilde{U}), \tag{A.13}$$

where $L_0^{-1} : \mathcal{K}_0^{r-1} \to \mathcal{K}_0^{r+1}$ is the solution operator of (A.4). If ρ is sufficiently small, it is easy to see via Lemma A.3, (A.11), and the contraction mapping theorem that Φ has a fixed point $U^{(1)}$ with $\|U^{(1)}\|_{\mathcal{K}^{r+1}} \le C|U_0|_{H^r\times H^{r-1}}$, which gives us the solution $U = \tilde{U} + U^{(1)}$ of the IBLe.

The proof of the regularity result is standard: $U \in L^2((0,t_0), H^{r+1}\times H^r)$ implies $U \in H^{r+1}\times H^r$ for almost every $t > 0$, and, starting again at some such t_1, we obtain $U \in \mathcal{K}^{r+2}((t_1,t_0)\times\mathbb{R})$. The necessary trace conditions at $t = t_1$ are automatically fulfilled. ∎

Remark A.5. (A.11) holds for $r \ge 2$ due to the special form of F, namely, due to the absence of terms of the form $\eta_{xx}(\eta_{xx} + \eta_{xxx})$ and $q_x(q_x + q_{xx})$. If, for instance, (1.1) is expanded to cubic terms, then we obtain a term $-3\varepsilon^{-2}\eta_{xx}^2\eta_x$ in (1.6), and then we would need $r > 5/2$ in Theorem A.1 and therefore $m = 3$ in Theorem 1.1.

Appendix B. Formal derivation of the IBLe. In order to make the paper sufficiently self-contained, and to point out the influence of dissipation, we give here a brief overview of the physical problem underlying (1.1) and describe how (1.1) is formally derived. Details of this calculus can be found in, e.g., [4] and the references therein.

B.1. The inclined film problem. We consider a two dimensional viscous liquid film flowing down an inclined "plane" with inclination angle θ. Using the thickness h_0 of the flat film as the characteristic length and the surface velocity $u_N = u_N(h_0) = gh_0^2 \sin\theta/2\nu$ of the basic Nusselt solution

$$(u, v, p) = (u_N, 0, p_N), \quad u_N(y) = \frac{g\sin\theta}{2\nu}(2h_0 y - y^2), \quad p_N = \rho g \cos\theta(h_0 - y)$$

as characteristic velocity, the governing dimensionless NSe and the continuity equation read

$$\mathbf{u}_t + (\mathbf{u}\cdot\nabla)\mathbf{u} = -\nabla p + \frac{1}{\mathrm{R}}\Delta\mathbf{u} + \frac{2}{\mathrm{R}}\mathbf{g}, \qquad \operatorname{div}\mathbf{u} = 0. \tag{B.1a}$$

Here $\mathbf{u} = (u, v)$ is the velocity field; $\mathrm{R} = u_N h_0/\nu$ is the Reynolds number; ν, ρ, g are the viscosity, density, and gravitational constant; and $\mathbf{g} = (1, -\cot\theta)$. At the free surface $y = h(t, x)$ we have the kinematic condition

$$h_t + h_x u = v \tag{B.1b}$$

and the tangential and normal stress conditions

$$4h_x u_x + (h_x^2 - 1)(u_y + v_x) = 0, \qquad p - \frac{2}{\mathrm{R}}\frac{h_x u_x - h_x(u_y + v_x) + v_y}{1 + h_x^2} = -\mathrm{W}_e K(h), \tag{B.1c}$$

where $\mathrm{W}_e = \sigma/(\rho u_N^2 h_0)$ is the Weber number, σ is the coefficient of surface tension, and $K(h) = h_{xx}(1 + h_x^2)^{-3/2}$ is the interfacial curvature. A constant atmospheric pressure p_a has been adsorbed into p. Finally, at the rigid wall we prescribe the no-slip condition

$$\mathbf{u} = 0 \quad \text{at} \quad y = 0. \tag{B.1d}$$

For $\mathrm{R} > \mathrm{R}_c = \frac{5}{4}\cot\theta$ Nusselt's solution is unstable to long wave perturbations [2], and in order to analyze this long wave instability a number of reduced equations for (B.1) have been derived. We briefly describe the derivation of (1.1).

B.2. Derivation of the IBLe. We assume that the Weber number is large, $\mathrm{W}_e = \mathrm{W}\varepsilon^{-2}$, where $0 < \varepsilon \ll 1$ is a small parameter, while the Reynolds number is $\mathcal{O}(1)$, and let

$$u(t, x, y) = \tilde{u}(\tau, \xi, y), \quad v(t, x, y) = \varepsilon^{2/3}\tilde{v}(\tau, \xi, y), \quad \tau = \varepsilon^{2/3}t, \quad \xi = \varepsilon^{2/3}x,$$
$$p(t, x, y) = \varepsilon^{-2/3}\tilde{p}(\tau, \xi, y), \quad h(t, x) = \tilde{h}(\tau, \xi).$$

Substituting this long wave ansatz into the free boundary value problem (B.1) and retaining terms up to order $\mathcal{O}(\varepsilon^{4/3})$, we obtain

$$\text{(B.2a)} \qquad \text{in } \Omega: \quad \varepsilon^{2/3}\big(\tilde{u}_\tau+\tilde{u}_\xi\tilde{u}+\tilde{u}_y\tilde{v}\big) = -\tilde{p}_\xi+\frac{\varepsilon^{4/3}\tilde{u}_{\xi\xi}+\tilde{u}_{yy}+2}{\mathrm{R}},$$

$$\text{(B.2b)} \qquad 0 = -\tilde{p}_y - \frac{2\varepsilon^{2/3}\cot\theta}{\mathrm{R}}+\frac{\varepsilon^{4/3}\tilde{v}_{yy}}{\mathrm{R}},$$

$$\text{(B.2c)} \qquad \tilde{u}_\xi = -\tilde{v}_y,$$

$$\text{(B.2d)} \qquad \text{at } y=\tilde{h}(t,\xi): \quad \tilde{h}_\tau+\tilde{h}_\xi\tilde{u} = \tilde{v},$$

$$\text{(B.2e)} \qquad -\tilde{u}_y+\varepsilon^{4/3}\big(\tilde{h}_\xi\tilde{u}_\xi+\tilde{h}_\xi^2\tilde{u}_y-\tilde{v}_\xi\big) = 0,$$

$$\text{(B.2f)} \qquad \tilde{p} - \frac{2\varepsilon^{4/3}(-\tilde{u}_\xi - \tilde{h}_\xi\tilde{u}_y)}{\mathrm{R}} = -\mathrm{W}\tilde{h}_{\xi\xi}\bigg(1-\frac{3}{2}\varepsilon^{4/3}\tilde{h}_\xi^2\bigg),$$

$$\text{(B.2g)} \qquad \text{at } y=0: \quad \tilde{u}=\tilde{v}=0.$$

In order to derive the IBLe, we define the flow rate

$$\text{(B.3)} \qquad \tilde{q}(\tau,\xi) = \int_0^{\tilde{h}(\tau,\xi)} \tilde{u}(\tau,\xi,y)\,\mathrm{d}y$$

such that $\partial_\tau\tilde{h} = -\partial_\xi\tilde{q}$. Following [23], we assume that the velocity field is slaved to the elevation $\tilde{h}$ and the flow $\tilde{q}$ in a Nusselt-like fashion, i.e.,

$$\text{(B.4)} \qquad \tilde{u} = \frac{3\tilde{q}}{2\tilde{h}^3}(2\tilde{h}y-y^2).$$

Substituting (B.3) and (B.4) into (B.2), we obtain

$$\begin{aligned}\tilde{q}_\tau = &-\frac{6}{5}\partial_\xi\bigg(\frac{\tilde{q}^2}{\tilde{h}}\bigg)+\varepsilon^{-2/3}\bigg[\tilde{h}\mathrm{W}\bigg(\tilde{h}_{\xi\xi\xi}\bigg(1-\frac{3\varepsilon^{4/3}\tilde{h}_\xi^2}{2}\bigg)-3\varepsilon^{4/3}\tilde{h}_{\xi\xi}^2\tilde{h}_\xi\bigg)+\frac{(2\tilde{h}-3\tilde{q}/\tilde{h}^2)}{\mathrm{R}}\bigg]\\ &-\frac{2\cot\theta\tilde{h}_\xi\tilde{h}}{\mathrm{R}}+\frac{\varepsilon^{2/3}}{\mathrm{R}}\bigg[\bigg(\frac{7}{2}\bigg)\tilde{q}_{\xi\xi}-\frac{9\tilde{q}_\xi\tilde{h}_\xi}{\tilde{h}}+\frac{6\tilde{q}\tilde{h}_\xi^2}{\tilde{h}^2}-\frac{9\tilde{q}\tilde{h}_{\xi\xi}}{2\tilde{h}}\bigg].\end{aligned}$$

This, together with $\partial_\tau\tilde{h} = -\partial_\xi\tilde{q}$, is (1.1) when scaling back to t,x, i.e., defining $h(t,x) = \tilde{h}(\varepsilon^{-2/3}\tau,\varepsilon^{-2/3}\xi)$ and $q(t,x) = \tilde{q}(\varepsilon^{-2/3}\tau,\varepsilon^{-2/3}\xi)$.

Remark B.1. Evaluating the assumption (B.4) mathematically seems rather difficult. Note that, with this assumption and defining $\tilde{v}(\tau,\xi,y) = -\int_0^y \tilde{u}_\xi(\tau,\xi,\tilde{y})\,\mathrm{d}\tilde{y}$, the no-slip boundary condition (B.2g) is fulfilled, but the condition (B.2e) for the tangential stress holds only up to order $\mathcal{O}(\varepsilon^{4/3})$. See also the following subsection.

B.3. Remarks on first order boundary layer theory. If in (B.2) we keep terms only up to order $\mathcal{O}(\varepsilon^{2/3})$, we obtain the so-called boundary layer equation (see [6])

$$\text{(B.5a)} \qquad \text{in } \Omega: \quad \tilde{u}_\tau+\tilde{u}_\xi\tilde{u}+\tilde{u}_y\tilde{u} = \frac{\varepsilon^{-2/3}}{\mathrm{R}}\big[\tilde{u}_{yy}+2+\mathrm{RW}\tilde{h}_{\xi\xi\xi}\big]-\frac{2\cot\theta\tilde{h}_\xi}{\mathrm{R}},$$

$$\text{(B.5b)} \qquad \tilde{u}_\xi = -\tilde{v}_y,$$

$$\text{(B.5c)} \qquad \text{at } y=\tilde{h}(t,\xi): \quad \tilde{h}_\tau+\tilde{h}_\xi\tilde{u} = \tilde{v}, \quad \tilde{u}_y = 0,$$

$$\text{(B.5d)} \qquad \text{at } y=0: \quad \tilde{u}=\tilde{v}=0.$$

In this case, the ansatz (B.3) and (B.4) gives

$$\begin{aligned} \tilde{h}_\tau &= -\tilde{q}_\xi, \\ \tilde{q}_\tau &= -\frac{6}{5}\partial_\xi\left(\frac{\tilde{q}^2}{\tilde{h}}\right) + \frac{\varepsilon^{-2/3}}{\mathrm{R}}\left[2\tilde{h} - \frac{3\tilde{q}}{\tilde{h}^2} + \mathrm{RW}\tilde{h}\partial_\xi^3\tilde{h}\right] - \frac{2\cot\theta\tilde{h}_\xi\tilde{h}}{\mathrm{R}}, \end{aligned} \tag{B.6}$$

and the reduction of (B.5) to (B.6) is exact; i.e., every solution of (B.6) gives an exact solution of (B.5) via (B.4). In other words, the solutions of (B.6) define an invariant manifold for (B.5). Moreover, (B.6) reduces to the KSe (1.14) in just the same way as (1.1) does.

However, (B.6) is a quasilinear *hyperbolic* system, as can be seen from the dispersion relation

$$\mu_{1,2}(k) = -\frac{1}{2}\left(\frac{3}{\mathrm{R}} + \frac{8}{5}\mathrm{i}k\right) \pm \sqrt{\frac{1}{4}\left(\frac{3}{\mathrm{R}} + \frac{8}{5}\mathrm{i}k\right)^2 - \frac{6}{\mathrm{R}}\mathrm{i}k - \left(\frac{4}{5} - \frac{2}{\mathrm{R}}\cot\theta\right)k^2 - \mathrm{W}\varepsilon^{-2}k^4} \tag{B.7}$$

for the linearization of (B.6) (after rescaling to t, x coordinates) around $(q, h) = (2/3, 1)$. In particular, for $|k| \to \infty$, this yields (in contrast to (1.8))

$$\mu_{1,2}(k) = -\frac{3}{2\mathrm{R}} \pm \mathrm{i}\left(\varepsilon^{-1}\sqrt{W}k^2 + \mathcal{O}(|k|)\right), \tag{B.8}$$

such that the high wave number modes are just uniformly damped but there is no dissipation.

The derivation of the KSe from (1.1) and from (B.6) is the same, since the dissipation terms first show up in the q-equation at order $\mathcal{O}(\varepsilon^3)$. In other words, the linear part of the KSe is determined from the local expansion of $\mu_1(k)$ at $k = 0$.

However, due to (B.8), the linearization of (B.6) around $(h, q) = (1, 2/3)$ generates only a strongly continuous semigroup. The local existence of solutions to (B.6) can still be shown, for instance, using the methods from [9], but with our method we cannot prove an approximation result like Theorem 1.1 for the reduction of (B.6) to the KSe. The reason for this is that, due to the lack of dissipation, our energy estimate for the quasilinear problem breaks down.

Acknowledgment. The author thanks R. L. Pego for helpful discussions and remarks.

REFERENCES

[1] J.T. Beale, *Large time regularity of viscous surface waves*, Arch. Ration. Mech. Anal., 84 (1984), pp. 307–352.

[2] T.B. Benjamin, *Wave formation in laminar flow down an inclined plane*, J. Fluid Mech., 2 (1957), pp. 554–574.

[3] H.-C. Chang and E.A. Demekhin, *Solitary wave formation and dynamics on falling films*, Adv. Appl. Mech., 32 (1996), pp. 1–58.

[4] H.-C. Chang and E.A. Demekhin, *Complex Wave Dynamics on Thin Films*, Stud. Interface Sci. 14, Elsevier, Amsterdam, 2002.

[5] H.-C. Chang, E.A. Demekhin, and E. Kalaidin, *Generation and suppression of radiation by solitary pulses*, SIAM J. Appl. Math., 58 (1998), pp. 1246–1277.

[6] H.-C. Chang, E.A. Demekhin, and D.I. Kopelevich, *Nonlinear evolution of waves on a vertically falling film*, J. Fluid Mech., 250 (1993), pp. 433–480.

[7] H.-C. Chang, E.A. Demekhin, and D.I. Kopelevich, *Local stability theory of solitary pulses in an active medium*, Phys. D, 97 (1996), pp. 353–375.

[8] Th. Gallay and G. Schneider, *KP-description of unidirectional dispersive waves—The model case*, Proc. Roy. Soc. Edinburgh A, 131 (2001), pp. 885–898.
[9] T. Kato, *Quasi-linear equations of evolution with applications to partial differential equations*, in Spectral Theory of Differential Equations, Lecture Notes in Math. 448, W. E. Everitt, ed., Springer, New York, 1975, pp. 25–70.
[10] P. Kirrmann, G. Schneider, and A. Mielke, *The validity of modulation equations for extended systems with cubic nonlinearities*, Proc. Roy. Soc. Edinburgh A, 122 (1992), pp. 85–91.
[11] J.L. Lions and E. Magenes, *Problèmes aux limites non homogènes*, Dunod, Paris, 1968.
[12] J. Liu and J.P. Gollub, *Solitary wave dynamics of film flows*, Phys. Fluids, 6 (1994), pp. 1702–1712.
[13] P. Manneville, *The Kuramoto–Sivashinsky equation: A progress report*, in Propagation in Systems Far from Equilibrium (Les Houches, 1987), Springer Ser. Synergetics 41, J.E. Wesfreid, H.R. Brand, P. Manneville, and G. Albinet, eds., Springer, Berlin, 1988, pp. 265–280.
[14] A. Mielke and G. Schneider, *Attractors for modulation equations on unbounded domains—Existence and comparison*, Nonlinearity, 8 (1995), pp. 743–768.
[15] T. Ogawa and H. Suzuki, *On the spectra of pulses in a nearly integrable system*, SIAM J. Appl. Math., 57 (1997), pp. 485–500.
[16] Th. Prokopiou, M. Cheng, and H.-C. Chang, *Long waves on inclined films at high Reynolds number*, J. Fluid Mech., 222 (1991), pp. 665–691.
[17] B. Ramaswamy, S. Chippada, and S. W. Joo, *A full scale numerical study of interfacial instabilities in thin film flows*, J. Fluid Mech., 325 (1996), pp. 163–194.
[18] G. Schneider, *Error estimates for the Ginzburg–Landau approximation*, Z. Angew. Math. Phys., 45 (1994), pp. 433–457.
[19] G. Schneider, *A new estimate for the Ginzburg–Landau approximation on the real axis*, J. Nonlinear Sci., 4 (1994), pp. 23–34.
[20] G. Schneider, *Validity and limitation of the Newell–Whitehead equation*, Math. Nachr., 176 (1995), pp. 249–263.
[21] G. Schneider, *Global existence results in pattern forming systems—Applications to* $3D$ *Navier–Stokes problems*, J. Math. Pures Appl. (9), 78 (1999), pp. 265–312.
[22] G. Schneider and C.E. Wayne, *The long wave limit for the water wave problem* I. *The case of zero surface tension*, Comm. Pure. Appl. Math., 53 (2000), pp. 1475–1535.
[23] W.Ya. Shkadov, *Wave conditions in the flow of a thin layer of a viscous liquid under the action of gravity*, Izv. Akad. Nauk SSSR Mekh. Zhidk. Gaza, 1 (1967), pp. 43–50.
[24] Y. Teramoto, *On the Navier–Stokes flow down an inclined plane*, J. Math. Kyoto University, 32 (1992), pp. 593–619.
[25] J. Topper and T. Kawahara, *Approximate equations for long nonlinear waves on a viscous fluid*, J. Phys. Soc. Japan, 44 (1978), pp. 663–666.
[26] K. Yosida, *Functional Analysis*, Springer, New York, 1971.

SIAM J. APPL. MATH.
Vol. 63, No. 4, pp. 1378–1391

BIFURCATION ANALYSIS OF A PREY-PREDATOR COEVOLUTION MODEL*

FABIO DERCOLE†, JEAN-OLIVIER IRISSON‡, AND SERGIO RINALDI§

Abstract. We show in this paper how numerical bifurcation analysis can be used to study the evolution of genetically transmitted phenotypic traits. For this, we consider the standard Rosenzweig–MacArthur prey-predator model [*The American Naturalist,* 97 (1963), pp. 209–223] and, following the so-called adaptive dynamics approach, we derive from it a second-order evolutionary model composed of two ODEs, one for the prey trait and one for the predator trait. Then, we perform a detailed bifurcation analysis of the evolutionary model with respect to various environmental and demographic parameters. Surprisingly, the evolutionary dynamics turn out to be much richer than the population dynamics. Up to three evolutionary attractors can be present, and the bifurcation diagrams contain numerous global bifurcations and codimension-2 bifurcation points. Interesting biological properties can be extracted from these bifurcation diagrams. In particular, one can conclude that evolution of the traits can be cyclic and easily promote prey species diversity.

Key words. bifurcation analysis, coevolution, evolution, evolutionary dynamics, Lotka–Volterra model, monomorphism, prey-predator model

AMS subject classifications. 92D15, 34C23, 65L07

PII. S0036139902411612

Introduction. One of the most important notions in biology, namely evolution, is now recognized to be of primary importance in many fields of science. Evolution of markets, institutions, technologies, languages, and social rules are relevant examples. Thus, a well-founded mathematical theory of evolution is now needed more than ever.

Evolving systems are in general composed of N homogeneous subsystems identified by two features: dimension n_i and characteristic trait x_i. For example, in ecology, the subsystems are interacting plant and animal populations, n_i is the number of individuals of each population or, equivalently, the density of the population, and x_i is a genetically transmitted phenotypic trait (e.g., body size). Both features vary in time, but densities can vary at much faster rates than traits. This means that an evolving system has two distinct timescales: one is fast and concerns only the densities, which vary while traits remain practically constant, and the other concerns the slow variation of the traits entraining slow variations of the densities. In some favorable cases, these slow variations of the traits can be described by a standard ODE model (called the evolutionary model).

The theoretical work developed so far has shown that evolutionary dynamics can be extremely complex. For example, cyclic regimes [7] (called *Red Queen dynamics*, as in [27]) and chaotic regimes [3] are possible, as well as evolutionary suicides and murders, which occur when the variation of the trait of a population entails the extinction of the same or another population [20, 10]. Moreover, an evolving system

*Received by the editors July 18, 2002; accepted for publication (in revised form) December 7, 2002; published electronically May 22, 2003.
http://www.siam.org/journals/siap/63-4/41161.html
†Dipartimento di Elettronica e Informazione, Politecnico di Milano, Via Ponzio 34/5, 20133 Milano, Italy (fabio.dercole@polimi.it).
‡Ecole Normale Supérieure, 45 rue d'Ulm, 75005 Paris, France (irisson@ens.fr).
§Dipartimento di Elettronica e Informazione, Politecnico di Milano, Via Ponzio 34/5, 20133 Milano, Italy, and Adaptive Dynamics Network, International Institute for Applied Systems Analysis, 2361 Laxenburg, Austria (sergio.rinaldi@polimi.it).

can also have alternative evolutionary attractors, in which case the fate of the system is determined by its ancestral conditions.

Once an evolutionary model is available, the powerful machinery of numerical bifurcation analysis can be applied to it. This is mandatory if the aim is to detect the impact of some strategic parameters on the evolution of the system. Systematic bifurcation analysis with respect to key environmental parameters could, for example, explain why ecosystems differ at various latitudes, altitudes, and depths. The few bifurcation studies performed to date on evolutionary models (see, for example, [19, 21, 7]) are far from satisfactory: they are inaccurate because they have been mainly carried out through simulation, and they are incomplete because they refer to non-generic cases or point out only some aspects of the full bifurcation diagram. In this article we therefore present an accurate and detailed bifurcation analysis of a typical evolutionary model. The problem we tackle is the coevolution of prey and predator traits, a subject that has received a great deal of attention in the last decade (see [1] for a review). We consider two populations (prey and predator) and two traits (one for each population), and the bifurcation analysis of the evolutionary model is performed with respect to pairs of parameters. The results we obtain are of rather limited biological value because they refer to a specific prey-predator coevolution model. However, the methodology is very general and could be applied to other models in order to obtain, through a suitable comparative analysis, general conclusions on the coevolution of prey-predator communities.

The paper is organized as follows. In the next section we recall how, under suitable assumptions on the mutation and selection processes, a canonical evolutionary model can be derived from a general population model [6, 5]. Then, we focus on the well known Rosenzweig–MacArthur prey-predator model [26], showing how the canonical evolutionary model can be explicitly derived from it. Finally, we present the bifurcation analysis of the evolutionary model and demonstrate how interesting biological conclusions can be extracted from it. Some comments and comparisons with the literature close the paper.

The canonical equation of monomorphic evolutionary dynamics. Consider two interacting populations, hereafter called *prey* and *predator* populations, with densities n_1 and n_2 and phenotypic traits x_1 and x_2.

At ecological timescale (fast dynamics), the traits are constant while the densities vary according to two ODEs of the form

$$\begin{aligned} \dot{n}_1 &= n_1 F_1(n_1, n_2, x_1, x_2), \\ \dot{n}_2 &= n_2 F_2(n_1, n_2, x_1, x_2), \end{aligned} \tag{1}$$

where F_i is the net per capita growth rate of the ith population. In the following, model (1), called the *resident model*, is assumed to have one strictly positive and globally stable equilibrium $\bar{n}(x_1, x_2)$ for each (x_1, x_2) belonging to a set of the trait space called the *stationary coexistence region*. This condition is not strictly necessary, but it simplifies the discussion.

At evolutionary timescale (slow dynamics), the traits vary according to two ODEs called the *evolutionary model:*

$$\begin{aligned} \dot{x}_1 &= k_1 G_1(x_1, x_2), \\ \dot{x}_2 &= k_2 G_2(x_1, x_2), \end{aligned} \tag{2}$$

where k_1 and k_2 are suitable constant parameters determined by size and frequency

of mutations. However, population densities vary slowly with the traits because, at evolutionary timescale, model (1) is always at the equilibrium $\bar{n}(x_1, x_2)$.

Some authors discuss evolutionary problems by assigning particular forms to the functions G_1 and G_2 in model (2) without connecting them with a population model (see [2] and references therein). More frequently, model (2) is derived from model (1) through various arguments [16, 2]. This is a little surprising, since the dynamics of the traits should reflect the characteristics of the mutation and selection processes, which, however, are not included in the resident model (1). In fact, the most transparent approach for deriving the evolutionary model (2) is the so-called *adaptive dynamics* approach [15, 22, 6, 13, 12] based on the *resident-mutant models*, which describe the interactions among three populations, namely, the two resident populations and a mutant population with trait x_1' or x_2'. (Notice that this approach rules out the possibility that prey and predator mutants are present at the same time.) When the prey population is split into two subpopulations (resident and mutant) with densities n_1 and n_1' and traits x_1 and x_1', the model is

$$\begin{aligned} \dot{n}_1 &= n_1 f_1(n_1, n_2, n_1', x_1, x_2, x_1'), \\ \dot{n}_2 &= n_2 f_2(n_1, n_2, n_1', x_1, x_2, x_1'), \\ \dot{n}_1' &= n_1' f_1'(n_1, n_2, n_1', x_1, x_2, x_1'). \end{aligned} \tag{3}$$

The initial value of n_1' in these equations is very small because a mutant population is initially composed of one or a few individuals. A similar third-order model involving the mutant trait x_2', the density n_2', and the function f_2' describes the case in which the mutant is a predator. In the ecological literature, models like model (3) are often called "competition models" because they describe the competition between two similar populations. Obviously, model (3), together with its companion model for the predator mutation, contains more information than the resident model (1). Indeed, the latter can be immediately derived from the former by disregarding the mutant equation and letting $n_1' = n_2' = 0$, thus obtaining

$$F_i(n_1, n_2, x_1, x_2) = f_i(n_1, n_2, 0, x_1, x_2, x_i'),$$

where the function $f_i(n_1, n_2, 0, x_1, x_2, x_i')$ does not depend on x_i'. The functions f_i and f_i', identifying the right-hand sides of the resident-mutant models, are called *fitness functions*, and they enjoy a number of structural properties. Function f_i, $i = 1, 2$, satisfies the condition

$$f_i(n_1, n_2, n_1', x_1, x_2, x_1) = F_i(n_1 + n_1', n_2, x_1, x_2)$$

because, when $x_1 = x_1'$, resident and mutant individuals do not differ, so that only the total number of prey $(n_1 + n_1')$ matters. Function f_1' is defined by

$$f_1'(n_1, n_2, n_1', x_1, x_2, x_1') = f_1(n_1', n_2, n_1, x_1', x_2, x_1)$$

because either one of the two prey subpopulations can be considered as mutant, provided the other is considered as resident. Of course, the same properties hold for the functions f_1, f_2, and f_2' appearing in the competition model for the predator mutation.

Now that we have defined the resident model (1) and the resident-mutant model (3), we can show how the evolutionary model (2) can be derived following the adaptive dynamics approach. For this, assume that the resident population model (1) with

traits x_1 and x_2 is at its equilibrium $\bar{n}(x_1, x_2)$ when a mutant appears. If mutations are rare at ecological timescale, the initial conditions $(\bar{n}_1(x_1, x_2), \bar{n}_2(x_1, x_2), n_1')$ can be used in model (3) to determine the fate of the mutant population. If the mutant population does not invade, i.e., if

$$f_1'(\bar{n}_1(x_1, x_2), \bar{n}_2(x_1, x_2), n_1', x_1, x_2, x_1') < 0 \tag{4}$$

for all small $n_1' > 0$, then it becomes extinct and the final result is still a pair of resident populations with traits x_1 and x_2 and densities $\bar{n}_1(x_1, x_2)$ and $\bar{n}_2(x_1, x_2)$. By contrast, it can be proved [11] that if (4) holds with the opposite inequality sign and if mutations are small (i.e., x_1' differs only slightly from x_1), then the resident population generically becomes extinct and is replaced by the mutant population with density $\bar{n}_1(x_1', x_2)$. In other words, each mutation brings a new trait into the system, but competition between resident and mutant populations selects the winner, namely, the trait that remains in the system. This kind of evolution of the traits is called *monomorphic evolution*. This process of mutation and selection can be further specified by making suitable assumptions on the frequency and distribution of small mutations [6, 5], and the conclusion is that the rate at which the trait x_i varies at evolutionary timescale is given by the following ODE, called the *canonical equation of adaptive dynamics*:

$$\dot{x}_i = k_i \bar{n}_i(x_1, x_2) \left. \frac{\partial f_i'}{\partial x_i'} \right|_{\substack{n_1=\bar{n}_1(x_1,x_2) \\ n_2=\bar{n}_2(x_1,x_2) \\ n_i'=0; x_i'=x_i}} , \tag{5}$$

where k_i is proportional to the frequency and variance of mutations, $\bar{n}_i(x_1, x_2)$ is the equilibrium density of the resident model, and $\partial f_i'/\partial x_i'$ is the derivative of the fitness of the mutant, called the *selective derivative*. Equation (5), written for the prey and for the predator, gives two ODEs that form the evolutionary model (2) with

$$G_i(x_1, x_2) = \bar{n}_i(x_1, x_2) \left. \frac{\partial f_i'}{\partial x_i'} \right|_{\substack{n_1=\bar{n}_1(x_1,x_2) \\ n_2=\bar{n}_2(x_1,x_2) \\ n_i'=0; x_i'=x_i}} . \tag{6}$$

Thus, model (5) describes the monomorphic coevolution of the traits under the assumption of rare and random mutations of small effects. Monomorphic evolutionary dynamics are usually presented by drawing a few trajectories of model (5) in the stationary coexistence region. This set of trajectories, called the *coevolutionary portrait*, points out, as sketched in Figure 1, all relevant invariant sets (equilibria, limit cycles, and saddle separatrices). Some trajectories of the coevolutionary portrait (see gray regions in Figure 1) reach the boundary of the stationary coexistence region, thus implying the extinction of one of the two populations.

Figure 2 schematically summarizes monomorphic evolution and highlights the different roles played by the three models we have introduced. The ecological literature mainly deals with the resident model (1) since ecologists are interested in the short-term dynamics of the populations and usually do not even consider the possibility of having a mutant population involved in the game. By contrast, theoretical studies on evolution are based on formal evolutionary models (2), or on verbal theories that can be considered as a sort of surrogate of these models. Figure 2 points out two facts. The first is that both the resident model (1) and the evolutionary model (2) are needed if one is interested in the population dynamics entrained at evolutionary timescale by the dynamics of the traits. The second is that the resident-mutant

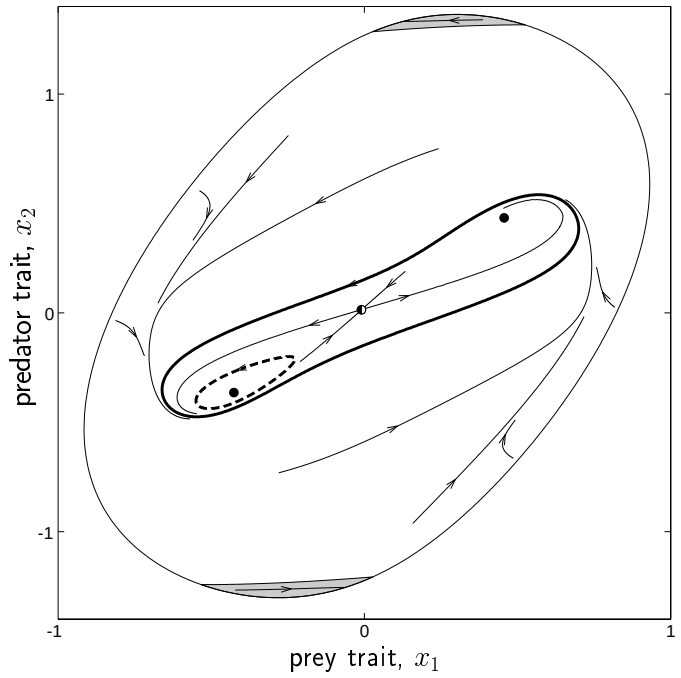


FIG. 1. *Example of a coevolutionary portrait in the stationary coexistence region. The portrait is characterized by three equilibria (two stable foci (filled circles) and one saddle (half-filled circle)) and two limit cycles (one stable (thick line) and one unstable (dashed line)).*

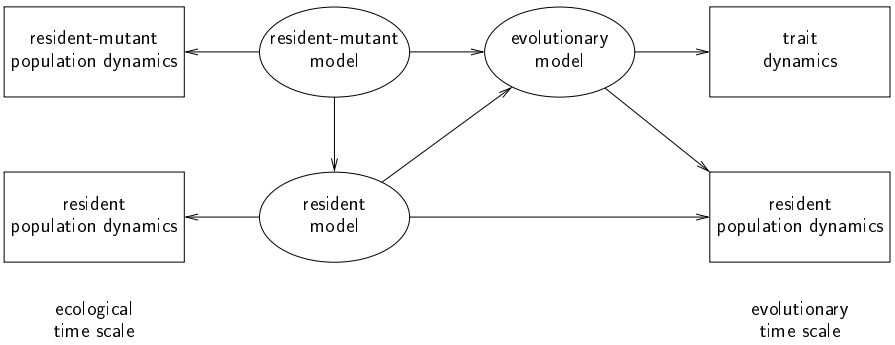


FIG. 2. *Flow chart demonstrating the relationships among resident-mutant models, the resident model, and the evolutionary model.*

model (3) is a "source" model, namely, a model that contains the information needed to answer all questions. Unfortunately, the scheme of Figure 2 is not always taken into account, and evolutionary models (2) are often derived directly from the resident model through arguments, which at best give the same result that a hidden equivalent resident-mutant model would give.

Once monomorphic dynamics has found a halt at a stable monomorphic equilibrium $\bar{x}$, one can look at the second-order terms in the Taylor expansion of the mutant fitness function to establish whether the equilibrium is a *branching point* [13] or not.

More precisely, a stable equilibrium $\bar{x}$ is said to be a branching point if

$$\left.\frac{\partial^2 f_i'}{\partial x_i'^2}\right|_{\substack{n_1=\bar{n}_1(\bar{x}_1,\bar{x}_2)\\ n_2=\bar{n}_2(\bar{x}_1,\bar{x}_2)\\ n_i'=0;x_i'=x_i;x=\bar{x}}} > 0 \tag{7}$$

and

$$\left.\frac{\partial^2 f_i'}{\partial x_i' \partial x_i}\right|_{\substack{n_1=\bar{n}_1(\bar{x}_1,\bar{x}_2)\\ n_2=\bar{n}_2(\bar{x}_1,\bar{x}_2)\\ n_i'=0;x_i'=x_i;x=\bar{x}}} < 0 \tag{8}$$

for i equal to 1 or 2, since small mutations of the ith population invade and coexist, at equilibrium, with the former resident [12]. Thus, branching points are the origin of dimorphism. Of course, after a branching has occurred, there are three resident populations, and one can continue the analysis by deriving the three corresponding canonical equations.

A model of prey-predator coevolution. In this section we specify the prey-predator coevolution problem, which is analyzed in what follows. First we present the resident prey-predator model that has most often been used in the last few decades to predict prey and predator abundances at ecological timescales in the absence of mutations. We then extend this model to a scenario in which a mutant population is also present, by adding a third ODE for the mutant population and by specifying the dependence of the demographic parameters upon the traits of the resident and mutant populations. This produces a resident-mutant population model from which, following the scheme described in the previous section, we finally derive an evolutionary model of the form (2) (details are relegated to the appendix).

The population model we consider is the well-known Rosenzweig–MacArthur prey-predator model [26]:

$$\begin{aligned} \dot{n}_1 &= n_1\left(r - cn_1 - \frac{a}{1+ahn_1}n_2\right),\\ \dot{n}_2 &= n_2\left(e\frac{an_1}{1+ahn_1} - d\right), \end{aligned} \tag{9}$$

where r is prey growth rate per capita, c is prey intraspecific competition, a is predator attack rate, h is predator handling time (namely, the time needed by each predator to handle and digest one unit of prey), e is efficiency (namely, a conversion factor transforming each unit of predated biomass into predator newborns), and d is predator death rate. The reader interested in more details on the biological interpretation of the parameters can refer to [23]. The six positive parameters of the model (r, c, a, h, e, d) could be reduced to three through rescaling. However, we do not follow this option because it would complicate the biological interpretation of the dependence of the parameters upon the prey and predator traits. In order to have a meaningful problem, one must assume that $e > dh$, because otherwise the predator population cannot grow even in the presence of an infinitely abundant prey population.

For any meaningful parameter setting, model (9) has a global attractor in $\mathbb{R}^2_+$, namely,

(a) the trivial equilibrium $(r/c, 0)$ if $d/a(e-dh) \geq r/c$,

(b) the strictly positive equilibrium

$$\bar{n}_1 = \frac{d}{a(e-dh)}, \qquad \bar{n}_2 = \frac{c}{a}\left(\frac{r}{c} - \frac{d}{a(e-dh)}\right)\left(1 + ah\frac{d}{a(e-dh)}\right) \tag{10}$$

if

$$\frac{rah-c}{2ahc} \le \frac{d}{a(e-dh)} < \frac{r}{c}, \tag{11}$$

(c) a strictly positive limit cycle if $d/a(e-dh) < (rah-c)/(2ahc)$.

The transition from (a) to (b) is a transcritical bifurcation (which is generic in the class of positive systems of the form (9)), while the transition from (b) to (c) is a supercritical Hopf bifurcation (see [17] for a proof).

If we now imagine that a mutant population is also present, we can enlarge model (9) by adding a third ODE and by slightly modifying the equations of the resident populations in order to take the mutant population into account. Of course we also need to specify how the parameters depend upon the traits x_1, x_2, x_1', x_2'. The number of possibilities is practically unlimited, because even for well-identified prey and predator species there are many meaningful options. Thus, at this level of abstraction, it is reasonable to limit the number of parameters sensitive to the traits, and to avoid trait dependencies that could give rise to biologically unrealistic evolutionary dynamics, like the unlimited growth of a trait (so-called runaway). Our choice has been to assume that the parameters r, e, and d are independent of the traits, because this will allow us to compare our results with those obtained in [7]. Thus, in the case of a mutation in the prey population, the resident-mutant model is

$$\begin{aligned}
\dot{n}_1 &= n_1\Bigg(r - c(x_1,x_1)n_1 - c(x_1,x_1')n_1' \\
&\qquad - \frac{a(x_1,x_2)}{1+a(x_1,x_2)h(x_1,x_2)n_1 + a(x_1',x_2)h(x_1',x_2)n_1'}n_2\Bigg), \\
\dot{n}_2 &= n_2\left(e\frac{a(x_1,x_2)n_1 + a(x_1',x_2)n_1'}{1+a(x_1,x_2)h(x_1,x_2)n_1 + a(x_1',x_2)h(x_1',x_2)n_1'} - d\right), \\
\dot{n}_1' &= n_1'\Bigg(r - c(x_1',x_1)n_1 - c(x_1',x_1')n_1' \\
&\qquad - \frac{a(x_1',x_2)}{1+a(x_1,x_2)h(x_1,x_2)n_1 + a(x_1',x_2)h(x_1',x_2)n_1'}n_2\Bigg).
\end{aligned} \tag{12}$$

The traits are assumed to be real variables obtained from the actual phenotypic traits through a suitable nonlinear scaling that maps the positive interval of the phenotype into the real axis. Thus, the maximum and minimum values of the prey (predator) phenotype correspond to the limit values ∞ and $-\infty$ of x_1 (resp., x_2). Similarly, in the case of a mutation in the predator population, the resident-mutant model is

$$\begin{aligned}
\dot{n}_1 &= n_1\Bigg(r - c(x_1,x_1)n_1 \\
&\qquad - \frac{a(x_1,x_2)}{1+a(x_1,x_2)h(x_1,x_2)n_1}n_2 - \frac{a(x_1,x_2')}{1+a(x_1,x_2')h(x_1,x_2')n_1}n_2'\Bigg), \\
\dot{n}_2 &= n_2\left(e\frac{a(x_1,x_2)n_1}{1+a(x_1,x_2)h(x_1,x_2)n_1} - d\right), \\
\dot{n}_2' &= n_2'\left(e\frac{a(x_1,x_2')n_1}{1+a(x_1,x_2')h(x_1,x_2')n_1} - d\right).
\end{aligned} \tag{13}$$

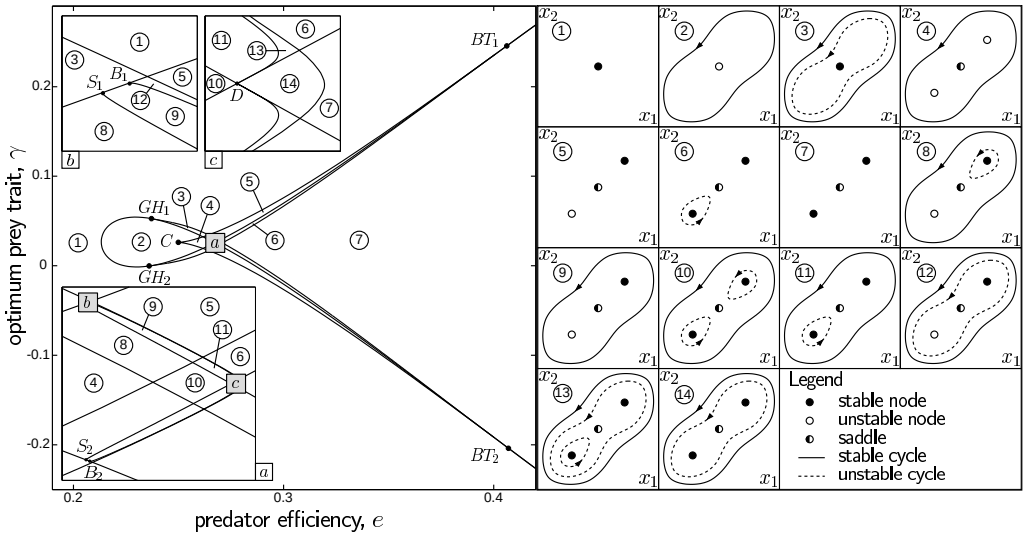


FIG. 3. *Bifurcation diagram of evolutionary model* (5) *with respect to predator efficiency* e *and optimum prey trait* γ *and corresponding sketches of coevolutionary state portraits. Panels a, b, and c are magnified views of the bifurcation diagram. Parameter values are* $r = 0.5$, $d = 0.05$, $k_1 = k_2 = 1$, $\gamma_0 = 0.01$, $\gamma_1 = 0.5$, $\gamma_2 = 1$, $\alpha = 1$, $\alpha_0 = 0.01$, $\alpha_1 = 1$, $\alpha_2 = 1$, $\alpha_3 = 0.6$, $\theta = 0.9$, $\theta_1 = \theta_2 = 0.5$, $\theta_3 = \theta_4 = 1$.

The functional forms specifying the parameters' dependence upon the traits are reported in the appendix and are such that the left-hand inequality of condition (11) (i.e., $(rah - c)/(2ahc) \leq d/(a(e - dh))$) is always satisfied. (This excludes the possibility of population cycles.) Thus, the boundary of the stationary coexistence region is simply the set of pairs (x_1, x_2) for which $d/(a(e-dh)) = r/c$ (see (11)). This means that on that boundary $\bar{n}_2(x_1, x_2) = 0$; i.e., the predator population becomes extinct if the traits reach the boundary of the stationary coexistence region.

At this point, (6) can be used to derive the evolutionary model (2), since the strictly positive equilibrium $\bar{n}(x_1, x_2)$ is known (see (10)). The analytic expressions of the selective derivatives and of the second-order derivatives needed for evaluating the branching conditions (7), (8) are not reported because they are very long. In any case, they can be easily derived by means of any software for symbolic computation.

Bifurcation analysis. The evolutionary model derived in the previous section has been studied through numerical bifurcation analysis. Local and global codimension-1 bifurcations with respect to various parameters have been obtained by means of specialized software based on continuation techniques [9, 8, 18]. Moreover, two-dimensional bifurcation diagrams have been produced by focusing on codimension-2 bifurcation points [17].

The first surprising result is that the evolutionary model is much richer than the resident population model. In fact, while the latter is characterized by two bifurcations, in the former twelve bifurcations have been detected. Figure 3 shows these bifurcations in the space (e, γ), where e is predator efficiency and γ is the prey trait value (called optimum) at which intraspecific competition is minimum. In general, both parameters are influenced by environmental factors. For example, the efficiency of an herbivore (predator) depends upon the caloric content of its prey (grass), which, in turn, is mainly fixed by humidity, temperature, and soil composition. Figure 3

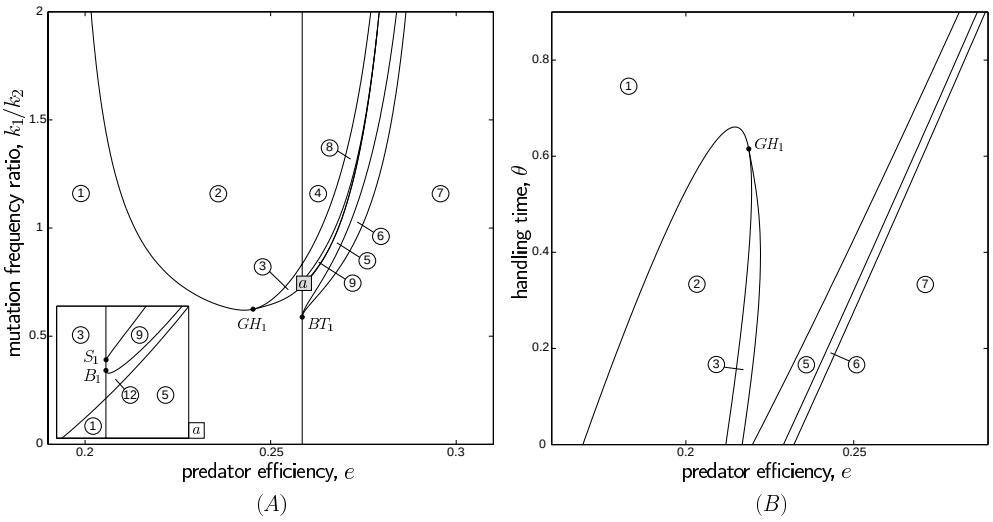


FIG. 4. *Bifurcation diagram of evolutionary model* (5) *with respect to predator efficiency* e *and mutation frequency ratio* k_1/k_2 (A) *and handling time* θ (B). *See Figure* 3 *for coevolutionary state portraits and parameter values.*

points out that there are fourteen subregions in the parameter space characterized by different coevolutionary portraits. In each one of them, for simplicity, the boundary of the stationary coexistence region, where the predator population becomes extinct, is not shown. This, however, fails to point out, graphically, that evolutionary extinction of the predator population occurs in all cases, as shown in Figure 1, which is actually the coevolutionary portrait corresponding to subregion 11. It is worth noticing that this form of evolutionary extinction is always an evolutionary murder. In fact, on the boundary of the stationary coexistence region $\dot{x}_2 = 0$, because $\bar{n}_2 = 0$ in (5); i.e., the predator trait is locally constant while the prey trait varies.

Coevolutionary attractors can be equilibria or limit cycles, and the existence of alternative attractors is rather common. When they exist, attracting cycles surround all equilibria. Actually, there can be up to three alternative attractors (two equilibria and one cycle), as shown by the coevolutionary portraits 10, 11, 13, and 14. There are ten codimension-2 bifurcation points, namely a cusp (C), two generalized Hopf (GH_1 and GH_2), two Bogdanov–Takens (BT_1 and BT_2), four noncentral saddle-node homoclinic loops (S_1, S_2, B_1, and B_2), and a double homoclinic loop (D) (see [17]).

No other bifurcation curves and codimension-2 bifurcation points are present in the two extra bifurcation diagrams presented in Figure 4, where the coevolutionary portraits are intentionally not shown to stress that they are exactly as in Figure 3. The parameter on the horizontal axis of these two bifurcation diagrams is still the efficiency of the predator, while the parameter on the vertical axis is related to two important characteristics of the mutation and predation processes, namely, the ratio k_1/k_2 between the frequencies of prey and predator mutations, and the predator handling time θ corresponding to the maximum attack rate (see the appendix).

The bifurcation diagrams are very useful for deriving interesting biological properties concerning the impact of various factors on coevolution. For example, one could be interested in identifying the factors favoring the so-called Red Queen dynamics, namely, the possibility of cyclic coevolution of the traits. For this, one should extract

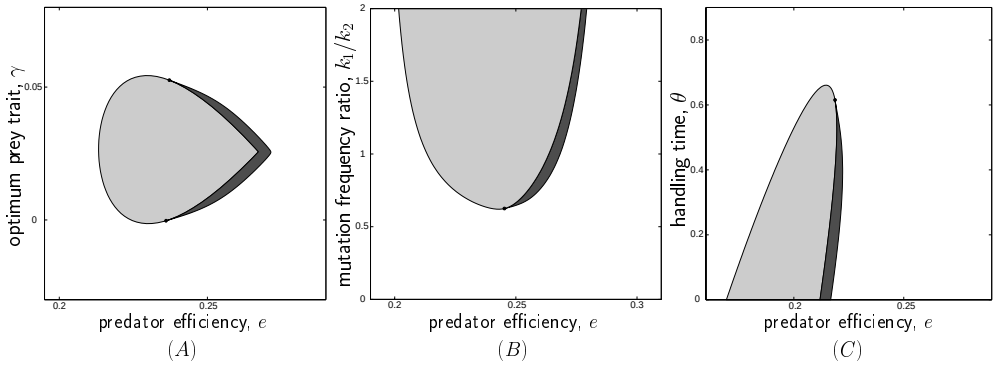


FIG. 5. *Red Queen dynamics: in the white regions cyclic coevolution is not possible, while in the gray regions it is the only long-term form of coevolution. In the black regions both stationary and cyclic coevolution are possible. Panels* (A), (B), *and* (C) *are extracted from the bifurcation diagrams of Figures* 3, 4(A), *and* 4(B), *respectively.*

from each bifurcation diagram the subregions 2–4, 8–14, where at least one of the coevolutionary attractors is a limit cycle. The result is Figure 5, which shows where cyclic coevolution is the only possible outcome (gray regions) and where stationary coevolution is also possible (black regions). Figure 5 indicates that Red Queen dynamics occur only for intermediate values of predator efficiency. Thus, slow environmental drifts entraining slow but continuous variations of predator efficiency can promote the disappearance of Red Queen dynamics. However, if efficiency decreases, Red Queen dynamics disappear smoothly through a supercritical Hopf bifurcation (where the attracting evolutionary cycle shrinks to a point). By contrast, if efficiency increases, Red Queen dynamics disappear discontinuously through a catastrophic bifurcation (tangent bifurcation of limit cycles). Figure 5 also indicates other biologically relevant properties, such as the fact that Red Queen dynamics are facilitated by high (low) frequency of prey (predator) mutation, and by low predator handling times. This last result shows that the highest chances for cyclic coevolution are obtained when $\theta = 0$, i.e., when the Rosenzweig–MacArthur model degenerates into the Lotka–Volterra model. This brings us to the following rather intriguing conclusion: the Lotka–Volterra assumptions (which do not give rise to population cycles) can easily explain coevolutionary cycles, while the Rosenzweig–MacArthur assumptions (which can easily give rise to population cycles) can hardly support Red Queen dynamics.

Extra information can be added to the bifurcation diagrams of Figures 3 and 4 by specifying whether the stable monomorphic equilibria $(\bar{x}_1, \bar{x}_2)$ are branching points (B) or not (NB). This can be easily done by computing (through continuation) the curves where conditions (7), (8) are critical. Thus, any region of parameter space characterized by only one stable monomorphic equilibrium can, in principle, be partitioned into four subregions: in one of these subregions monomorphism is the only form of coevolution, while in the other three regions dimorphism is possible through the branching of one of the two populations or of both. However, in all of the numerical experiments that we have performed, only prey branching occurred. This is consistent with the well-known principle of "competitive exclusion" [14]. In fact, if the predator population would branch, the system would converge to an equilibrium with two slightly different predators and one prey, in contrast with the competitive exclusion principle. In conclusion, there are only two possibilities: $(\bar{x}_1, \bar{x}_2)$ is not a

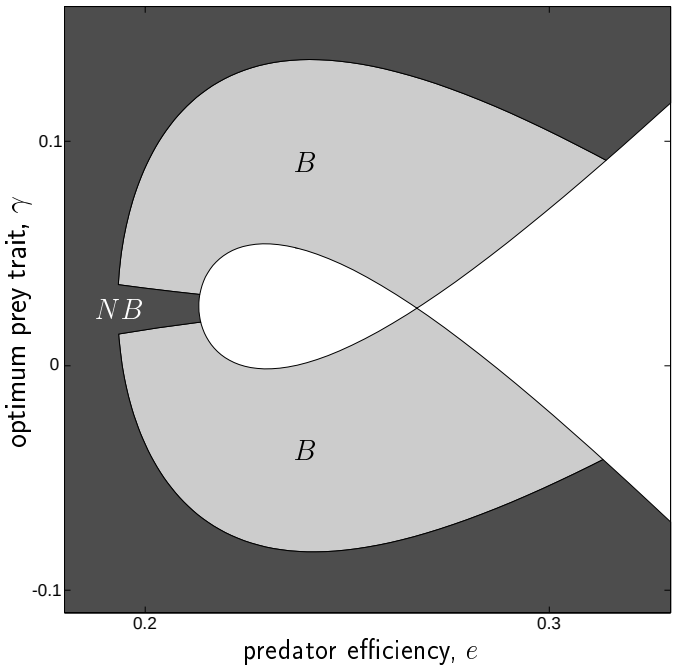


FIG. 6. *In the gray and black regions (extracted from Figure* 3*) the evolutionary model* (5) *has only one stable equilibrium, which is either a branching point (B) for the prey population or not (NB).*

branching point or it is a branching point for the prey population. In other words, our findings are in line with biological principles and support the idea [4] that predators are promoters of prey species diversity. Figure 6 shows how the region characterized by a unique stable equilibrium $(\bar{x}_1, \bar{x}_2)$ is partitioned into B and NB subregions. The result is rather interesting if it is complemented with what has already been discovered about the disappearance of Red Queen dynamics induced by variations of predator efficiency. In fact, the overall conclusion is that Red Queen dynamics disappear abruptly if predator efficiency increases, and smoothly if predator efficiency decreases. However, in the latter case, as soon as Red Queen dynamics disappear, dimorphism can occur in the prey population. Thus, environmental drifts of any sign can give rise to discontinuities in the dynamics of the traits. This observation proves once more that coevolution is an astonishingly complex dynamic process.

Discussion and conclusions. The problem of prey-predator coevolution has been investigated in this paper from a purely mathematical point of view. For this, the classic Rosenzweig–MacArthur model (logistic prey and predator with saturating functional response) has first been transformed into a resident-mutant model by adding a third equation for the mutant population. Then, an evolutionary model describing the slow dynamics of the traits has been derived from the resident-mutant model through the standard adaptive dynamics approach [15, 22, 6, 13, 12]. The bifurcation analysis of the evolutionary model has shown that the dynamics of the traits at evolutionary timescale are much more complex than the dynamics of the populations at ecological timescale. The numerically produced bifurcation diagrams have proved to be powerful tools for extracting qualitative information on the impact

of various factors on coevolution. Conclusions like those we have obtained on the impact of environmental drifts on evolutionary cycles (so-called Red Queen dynamics) could not have been derived without performing a detailed bifurcation analysis of an evolutionary model. A generally encouraging message emerging from this study is that other very important biological problems, such as the evolution of mutualism, cannibalism, and parasitism, could most likely be studied successfully through the bifurcation analysis of the canonical evolutionary model. But the same approach should also be very effective for studying relevant problems in social sciences and economics, where mechanisms somewhat similar to biological mutation and selection can sometimes be identified.

Limiting the discussion to the problem of prey-predator coevolution, we can say that the results presented in this paper are far more complete than those available in the literature. Indeed, the only comparable result is the bifurcation analysis presented in [7], where a bifurcation diagram similar to that of Figure 4(A) was obtained through simulation. That bifurcation diagram is incomplete and derived for a quite degenerate case, i.e., for a Lotka–Volterra model ($\theta = 0$ in our model) with a very special parameter combination reducing the number of bifurcation curves to six. However, despite this double degeneracy, the analysis in [7] points out Red Queen dynamics, multiple evolutionary attractors, and evolutionary murder. A comparison with the nonmathematical literature (see, for example, [24, 25] and [1]) neither contradicts nor supports our findings.

Even if what we have presented in this paper might seem rather general, the analysis should first be repeated for many other prey-predator models and for different assumptions on the trait dependence of the demographic parameters, and a comparative analysis should be performed in order to extract biologically significant results. Moreover, there are a number of possible interesting extensions. First, one could investigate the dynamics of dimorphism by applying the bifurcation approach followed in this paper to more complex population assemblies, composed, for example, of one resident predator population and two resident prey populations. The outcome of such a study could be that a predator branching generating a second resident predator population is possible, because this outcome is not in conflict with the principle of competitive exclusion. Second, while remaining in the simple context of monomorphism, one could be interested in detecting the prey-predator coevolutionary dynamics under the assumption that the two populations can coexist by cycling at ecological timescale. This extension is absolutely not trivial, because the derivation of the evolutionary model is rather difficult in this case. However, the problem is of great interest because its analysis could perhaps help to answer the very intriguing question: does coevolution destabilize populations? Third, one could be interested in extending the analysis to the coevolution of tritrophic food chains composed of a prey, a predator, and a superpredator population. From the results obtained in this paper, showing that evolutionary dynamics of ditrophic food chains (composed of a prey and a predator population) are much more complex than the corresponding population dynamics, one should naturally be inclined to conjecture that chaotic coevolutionary dynamics should be possible in tritrophic food chains. The proof of this conjecture would be a great result.

Appendix. In this appendix we specify how the prey intraspecific competition c, the predator attack rate a, and the predator handling time h, appearing in the resident-mutant models (12), (13), depend upon the resident and mutant traits. Due to our definition of the traits, which are scaled measures of the phenotypes, c, a, and

h are bounded functions of the traits. Unless otherwise stated, all the parameters appearing in these functions are assumed to be positive.

Prey intraspecific competition c is given by

$$c(x_1, x_1') = \frac{\gamma_1 + \gamma_2 (x_1 - \gamma)^2}{1 + \gamma_0(\gamma_1 + \gamma_2 (x_1 - \gamma)^2)}. \tag{A1}$$

Notice that c depends only upon its first argument. This means that resident prey individuals face the same competition when they are opposed to other resident individuals or to mutant individuals. The parameter γ, which can be either positive or negative, is the value of the prey trait x_1 (called *optimum prey trait*) at which intraspecific competition is minimum (and equal to $\gamma_1/(1+\gamma_0\gamma_1)$). For prey traits x_1 far from γ, intraspecific competition saturates at $1/\gamma_0$.

The predator attack rate a is the bell-shaped function

$$a(x_1, x_2) = \alpha_0 + \alpha \exp\left(- \left(\frac{x_1}{\alpha_1}\right)^2 + 2\alpha_3 \left(\frac{x_1}{\alpha_1}\right)\left(\frac{x_2}{\alpha_2}\right) - \left(\frac{x_2}{\alpha_2}\right)^2 \right), \tag{A2}$$

where $\alpha_3 < 1$. If prey and predator traits are tuned, i.e., if $x_1 = x_2 = 0$, the predator attack rate is maximum (and equal to $\alpha_0 + \alpha$). When prey and predator traits are far from being tuned, the predator attack rate drops to α_0.

The predator handling time is the product of an increasing sigmoidal function of the prey trait x_1 and of a decreasing sigmoidal function of the predator trait x_2,

$$h(x_1, x_2) = \theta \left[1 + \theta_1 - \frac{2\theta_1}{1 + \exp(\theta_3 x_1)}\right] \left[1 + \theta_2 - \frac{2\theta_2}{1 + \exp(-\theta_4 x_2)}\right], \tag{A3}$$

where θ is the handling time corresponding to the tuned situation ($(x_1, x_2) = (0, 0)$), referred to as handling time in Figures 4(B) and 5(C).

Finally, we have fixed r, d, and all the parameters of the functions c, a, and h at the values indicated in the caption of Figure 3, and we have limited θ from above and e from below so that the following two inequalities hold for all (x_1, x_2):

$$e - dh(x_1, x_2) > 0, \qquad \frac{r}{c(x_1)} \leq \frac{1}{a(x_1, x_2)h(x_1, x_2)}.$$

These conditions guarantee that the left-hand inequality of condition (11) holds. Thus, population cycles are ruled out from our study.

Acknowledgments. The authors are grateful to Ulf Dieckmann, who invited them to participate is the Adaptive Dynamics Network program at the International Institute for Applied Systems Analysis, Laxenburg, Austria. The support of CESTIA (CNR, Milano, Italy) to S. R. is also acknowledged.

REFERENCES

[1] P. A. Abrams, *The evolution of predator-prey interactions: Theory and evidence*, Ann. Rev. Ecology Systematics, 31 (2000), pp. 79–105.

[2] P. A. Abrams, *Modelling the adaptive dynamics of traits involved in inter and intraspecific interactions: An assessment of three methods*, Ecology Letters, 4 (2001), pp. 166–175.

[3] P. A. Abrams and H. Matsuda, *Prey evolution as a cause of predator-prey cycles*, Evolution, 51 (1997), pp. 1740–1748.

[4] J. S. Brown and T. L. Vincent, *Organization of predator-prey communities as an evolutionary game*, Evolution, 46 (1992), pp. 1269–1283.
[5] N. Champagnat, R. Ferrière, and G. Ben Arous, *The canonical equation of adaptive dynamics: A mathematical view*, Selection, 2 (2001), pp. 73–83.
[6] U. Dieckmann and R. Law, *The dynamical theory of coevolution: A derivation from stochastic ecological processes*, J. Math. Biol., 34 (1996), pp. 579–612.
[7] U. Dieckmann, U. Marrow, and R. Law, *Evolutionary cycling in predator-prey interactions: Population dynamics and the Red Queen*, J. Theoret. Biol., 176 (1995), pp. 91–102.
[8] E. Doedel, A. Champneys, T. Fairgrieve, Y. A. Kuznetsov, B. Sandstede, and X. Wang, *AUTO97: Continuation and bifurcation software for ordinary differential equations (with HOMCONT)*, Department of Computer Science, Concordia University, Montreal, QC, 1997.
[9] E. Doedel and J. P. Kernevez, *AUTO: Software for continuation problems in ordinary differential equations*, Department of Applied Mathematics, California Institute of Technology, Pasadena, CA, 1986.
[10] R. Ferrière, *Adaptive responses to environmental threats: Evolutionary suicide, insurance and rescue*, Options, Spring (2000), pp. 12–16.
[11] S. A. H. Geritz, M. Gyllenberg, F. J. A. Jacobs, and K. Parvinen, *Invasion dynamics and attractor inheritance*, J. Math. Biol., 44 (2002), pp. 548–560.
[12] S. A. H. Geritz, E. Kisdi, G. Meszéna, and J. A. J. Metz, *Evolutionarily singular strategies and the adaptive growth and branching of the evolutionary tree*, Ecology, 12 (1998), pp. 35–57.
[13] S. A. H. Geritz, J. A. J. Metz, E. Kisdi, and G. Meszéna, *The dynamics of adaptation and evolutionary branching*, Phys. Rev. Lett., 78 (1997), pp. 2024–2027.
[14] G. Hardin, *The competitive exclusion principle*, Science, 131 (1960), pp. 1292–1298.
[15] J. Hofbauer and K. Sigmund, *Adaptive dynamics and evolutionary stability*, Math. Lett., 3 (1990), pp. 75–79.
[16] A. I. Khibnik and A. S. Kondrashov, *Three mechanisms of Red Queen dynamics*, Proc. Roy. Soc. London B, 264 (1997), pp. 1049–1056.
[17] Y. A. Kuznetsov, *Elements of Applied Bifurcation Theory*, Springer-Verlag, Berlin, 1998.
[18] Y. A. Kuznetsov and V. V. Levitin, *CONTENT: A Multiplatform Environment for Analyzing Dynamical Systems*, Dynamic Systems Laboratory, Centrum voor Wiskunde en Informatica, Amsterdam, The Netherlands, 1997; available from `ftp.cwi.nl/pub/CONTENT`.
[19] P. Marrow, R. Law, and C. Cannings, *The coevolution of predator-prey interactions: ESSs and Red Queen dynamics*, Proc. Roy. Soc. London B, 250 (1992), pp. 133–141.
[20] H. Matsuda and P. A. Abrams, *Runaway evolution to self-extinction under asymmetrical competition*, Evolution, 48 (1994), pp. 1764–1772.
[21] H. Matsuda and P. A. Abrams, *Timid consumers: Self extinction due to adaptive change in foraging and anti-predator effort*, Theoret. Population Biology, 45 (1994), pp. 76–91.
[22] J. A. J. Metz, R. M. Nisbet, and S. A. H. Geritz, *How should we define fitness for general ecological scenarios?*, Trends in Ecology and Evolution, 7 (1992), pp. 198–202.
[23] S. Muratori and S. Rinaldi, *Low- and high-frequency oscillations in three-dimensional food chain systems*, SIAM J. Appl. Math., 52 (1992), pp. 1688–1706.
[24] D. Pimentel, *Animal population regulation by the genetic feedback mechanism*, The American Naturalist, 95 (1961), pp. 65–79.
[25] D. Pimentel, *Population regulation and genetic feedback*, Science, 159 (1968), pp. 1432–1437.
[26] M. L. Rosenzweig and R. H. MacArthur, *Graphical representation and stability conditions of predator-prey interactions*, The American Naturalist, 97 (1963), pp. 209–223.
[27] L. Van Valen, *A new evolutionary law*, Evolutionary Theory, 1 (1973), pp. 1–30.

SIAM J. APPL. MATH.
Vol. 63, No. 4, pp. 1392–1423

ANALYSIS OF A MODEL FOR MULTICOMPONENT MASS TRANSFER IN THE CATHODE OF A POLYMER ELECTROLYTE FUEL CELL*

M. VYNNYCKY† AND E. BIRGERSSON†

Abstract. A chief factor that is thought to limit the performance of polymer electrolyte fuel cells (PEFCs) is the hydrodynamics associated with the cathode. In this paper, a two-dimensional model for three-component (oxygen, nitrogen, water) gaseous flow in a PEFC cathode is derived, nondimensionalized, and analyzed. The fact that the geometry is slender allows the use of a narrow-gap approximation leading to a simplified formulation. In spite of the highly nonlinear coupling between the velocity variables and the mole fractions, an asymptotic treatment of the problem indicates that oxygen consumption and water production can be described rather simply in the classical lubrication theory limit with the reduced Reynolds number as a small parameter. In general, however, the reduced Reynolds number is $O(1)$, requiring a numerical treatment; this is done using the Keller–Box discretization scheme. The analytical and numerical results are compared in the limit mentioned above, and further results are generated for varying inlet velocity and gas composition, channel width and porous backing thickness, pressure and current density. Also, a novel, compact way to present fuel cell performance, which takes into account geometrical, hydrodynamical, and electrochemical features, is introduced.

Key words. proton-exchange membrane (PEM) fuel cells

AMS subject classifications. 35-04, 76S05

PII. S003613990139369X

1. Introduction.

There is at present a rapidly increasing interest in improving the design of fuel cells, that is, electrochemical devices that convert the chemical energy of a fuel with an oxidant directly into electricity. Fuel cells have a variety of applications; for instance, the alkaline fuel cell (AFC) was mainly used in space exploration, while the phosphoric acid fuel cell (PAFC), the solid oxide fuel cell (SOFC), and the molten carbonate fuel cell (MCFC) are most suited to stationary applications. Of the several types of fuel cells that are currently under development, perhaps the one that has received the most attention, particularly from the point of view of commercialization in the automotive industry, has been the polymer electrolyte fuel cell (PEFC), also often referred to as the proton-exchange membrane (PEM) fuel cell or the solid polymer fuel cell (SPFC); the merit of this type of fuel cell over others for this particular application is that it can generate the high current densities that are required to power a vehicle, as well as the fact that it operates at comparatively low temperatures (often no higher than 100°C).

A schematic diagram of a PEFC is given in Figure 1. Essentially, this entails a polymer membrane sandwiched between two gas-diffusion electrodes, which are each adjacent to flow channels contained within bipolar plates. The oxidant, usually oxygen from air which is either dry or humidified to some extent, is fed in at the inlet of the channel on the cathode side and is transported to the electrolyte/cathode interface; the fuel, on the other hand, normally hydrogen, is fed at the anode channel inlet and is transported to the electrolyte/anode interface. Both interfaces contain a catalyst,

*Received by the editors August 13, 2001; accepted for publication (in revised form) December 5, 2002; published electronically May 22, 2003. This work was supported by the Swedish Foundation for Strategic Environmental Research (MISTRA). The work was done within the framework of the Jungner Centre.

http://www.siam.org/journals/siap/63-4/39369.html

†FaxénLaboratoriet, KTH, 100 44 Stockholm, Sweden (michaelv@mech.kth.se, erikb@mech.kth.se).

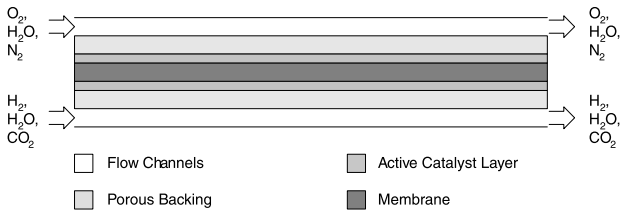


FIG. 1. 2*D PEFC.*

often platinum, to accelerate the reactions

$$2H_2 \rightarrow 4H^+ + 4e^- \quad \text{at the anode,} \tag{1.1}$$

$$O_2 + 4H^+ + 4e^- \rightarrow 2H_2O \quad \text{at the cathode,} \tag{1.2}$$

in the course of which an electric current is produced to drive a given load. In particular, the reaction at the cathode also produces both heat and water as by-products, the latter of which may be present throughout the system as either vapor or liquid or both; the production of the former can lead to temperatures at the catalytic layer in the order of 80–90°C. Optimal fuel cell performance is achieved at typical voltages of around 0.5 V at current densities of about 1 Acm^{-2}.

Recent years have seen the appearance of mathematical models for some or all of the parts of a typical fuel cell described above. Modeling proves necessary because of an, as yet, incomplete understanding of several important phenomena:

1. mass transport limitations, that is, to ensure that sufficient amounts of oxygen reach the catalytic layer at the cathode in order that a desired current is sustained;
2. water management, that is, to ensure that the water flow in the system is great enough to keep the membrane adequately hydrated but low enough to prevent flooding;
3. thermal management, that is, to ensure that the cell does not overheat, which may well occur as the result of the heat produced by electrochemical reactions in the catalyst layer.

Since the full problem is highly three-dimensional (3D), nonisothermal, multiphase, multicomponental, and most likely time-dependent in nature, numerous simplifications have been made in existing models to ensure some element of tractability. Perhaps the first one-dimensional (1D) models to provide a simplified treatment were developed by Bernardi and Verbrugge [5, 6] and Springer, Zawodzinski, and Gottesfeld [37]; a recent contribution is due to Gurau, Barbir, and Lui [20]. 1D treatments, whilst they are able to address some aspects of the three issues related to fuel cell performance mentioned above, are not able to address these questions at a local level: that is to say, where oxygen depletion occurs or where there is flooding or inadequate heat removal. Subsequent pseudo–two-dimensional (2D) models have tackled some of these issues [12, 17, 31, 44] with varying assumptions about the nature of the flow; in these so-called along-the-channel models, the resulting equations are ordinary differential equations with the coordinate along the fuel cell as the independent variable. Most recently, techniques of computational fluid dynamics have been used. Amongst models assuming single-phase gaseous flow, there are 2D isothermal models for the cathode [25, 45], 2D isothermal models for the whole cell [18, 19, 23, 36], 3D isother-

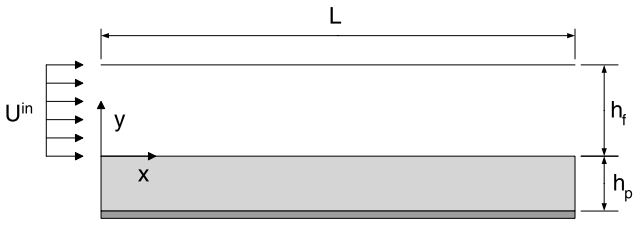


FIG. 2. *The cathode of a PEFC.*

mal models for the whole cell [14, 15], and 3D nonisothermal models for the whole cell [35]; generally speaking, there does not appear to be any experimental evidence that fuel cells are isothermal, although this assumption may indeed be valid for either small cells or large cells from which heat is removed at an adequate rate. In addition, two-phase flow at the cathode has also begun to receive attention [22, 31, 41].

This paper primarily addresses the first issue of the three given above. In addition, one of the goals is to steer between 1D models and full computational fluid dynamics to derive a 2D formulation that does not sacrifice too many geometrical features, yet on the other hand does not demand excessive computing time either. We focus here on the isothermal, three-component, gas-phase, 2D flow in a gas channel and adjacent porous gas backing of a PEFC cathode (Figure 2), although we note that the problem of multicomponent flow is a generic one, appearing not only in both electrodes of a PEFC (the gases are $(O_2,\ N_2,\ H_2O)$ at the cathode and (H_2, CO_2, H_2O) at the anode), but also in electrodes of other types of fuel cells [7, 21]. The geometry is assumed to be slender, as is typically the case in practice. Air, possibly humidified, is fed in at the inlet at the left (Figure 2); oxygen that reaches the catalytic layer reacts to produce water vapor, which is transported, along with oxygen and nitrogen, out at the outlet. The approach used here, however, differs from previous ones in that we use scaling arguments, nondimensionalization, and asymptotics to identify the main governing parameters and, subsequently, to obtain a reduced model. The benefits of this are the availability of closed-form analytical solutions in certain limits, as well as a model that is cheap to compute away from those limits; this feature is important from the point of view of extension to fuel cell stacks where transport in as many as 125 such assemblies may need to be computed (see, e.g., [27, 28, 29, 39]). The solution of this benchmark problem is useful from several other points of view:

- as a basis for later work and comparison when two-phase flow is introduced;
- to elucidate features that might not be obvious from simply solving the full equations.

Regarding the second point, it is clear from the majority of cathode studies that the mole fraction of O_2 decreases monotonically along the channel, while the mole fraction of H_2O increases, with the two slopes in some way dependent on physical and operating parameters. Among the results of the present treatment are closed-form expressions for these in certain limits.

The mathematical model is formulated in section 2. This consists of mass, momentum, and species transport equations and allows for the possibility of varying mixture density, as well as the crossed diffusion of species. A nondimensional analysis of the governing equations in section 3 provides an indication of the qualitative features one would expect in a multicomponent flow; there are found to be similarities with classical lubrication theory, in view of the slenderness of the geometry, except

that the reduced Reynolds number is typically $O(1)$. Furthermore, transport in the porous backing is found, to a reasonable approximation, to be 1D. In section 4, the first term in an asymptotic series in the reduced Reynolds number is derived: at leading order, the mole fractions are found to be solely a function of the distance along the fuel cell. Section 5 provides a description of a numerical scheme that is subsequently used when the reduced Reynolds number is $O(1)$; the scheme is verified in the lubrication theory limit for which closed-form solutions can be secured. Section 6 presents the results. A novel feature which we demonstrate here is that the traditional method for evaluating fuel cell performance, namely through the use of polarization curves, can be supplemented by the concept of a "polarization surface," whereby the average current density is plotted not as a function of cell potential but as a function of two dimensionless parameters which depend on cell potential, channel geometry, and inlet velocity; consequently, individual polarization curves are then paths along a "polarization surface." The implications of these results for a PEFC are also considered, in particular, as regards the limitations of the formulation with respect to liquid water formation, and conclusions are drawn in section 7.

2. Mathematical formulation.

2.1. Basics of multicomponent flow. We define the local mass average velocity, $\mathbf{v}$, of an n-component gas by

$$\mathbf{v} = \frac{\sum_{i=1}^{n} \rho_i \mathbf{v}_i}{\sum_{i=1}^{n} \rho_i},$$

where $\mathbf{v}_i$ denotes the velocity of species i with respect to stationary coordinate axes, and ρ_i is the mass concentration (the mass of species i per unit of volume of solution). For each component, the mass flux with respect to a coordinate system fixed in space is given by

$$\mathbf{n}_i = \rho\omega_i\mathbf{v} + \mathbf{j}_i, \quad i = 1, \ldots, n,$$

with

$$\rho = \sum_{i=1}^{n} \rho_i,$$

where ω_i is the mass fraction of species i, given by $\omega_i = \rho_i/\rho$, $\mathbf{j}_i$ is the mass diffusive flux relative to the mass-averaged velocity, and ρ_i denotes the density of species i. If we consider just concentration diffusion for an ideal gas mixture [8], we have

$$\mathbf{j}_i = \frac{c^2}{\rho} \sum_{j=1}^{n} M_i M_j D_{ij} \nabla x_j, \quad i = 1, \ldots, n; \tag{2.1}$$

here $(M_i)_{i=1,\ldots,n}$ are the molecular weights, $(D_{ij})_{i,j=1,\ldots,n}$ are the multicomponent diffusion coefficients, and $(x_i)_{i=1,\ldots,n}$ is the mole fraction of species i and is given by $x_i = c_i/c$, where c_i is the molar concentration of species i in moles per m^3 ($c_i = \rho_i/M_i$) and

$$c = \sum_{i=1}^{n} c_i.$$

Useful additional identities are $c = \rho/M$, where $M = \sum_{i=1}^{n} x_i M_i$ and a relation between the mass and mole fractions

$$\omega_i = x_i c M_i/\rho.$$

In general, $(D_{ij})_{i,j=1,\ldots,n}$ are strongly dependent on composition but can be expressed in terms of the Stefan–Maxwell diffusion coefficients, $(\mathcal{D}_{ij})_{i,j=1,\ldots,n}$, which are independent of composition. For a three-component system, as will be the case here, the relations are of the form [8]

$$D_{ij} = \mathcal{D}_{ij}\left\{1 + \frac{x_k\left[(M_k/M_j)\,\mathcal{D}_{ik} - \mathcal{D}_{ij}\right]}{x_i\mathcal{D}_{jk} + x_j\mathcal{D}_{ik} + x_k\mathcal{D}_{ij}}\right\}, \quad i,j,k = 1,2,3 \ \ (i \neq j)\,. \tag{2.2}$$

$(\mathcal{D}_{ij})_{i,j=1,\ldots,n}$ can in principle be measured experimentally [8, 38].

For the mixture density, we use the constitutive relation for an ideal gas,

$$\rho = \frac{pM}{RT}, \tag{2.3}$$

where p is the pressure, T is the temperature, and R is the universal gas constant (8.314 $\mathrm{kgm^2s^{-2}mol^{-1}K^{-1}}$). We note, in addition, the possibility that the mixture viscosity, μ_{mix}, will not necessarily be constant either, although we treat it to be so here.

2.2. Channel. Consider the 2D steady flow of a three-component gas in a channel of height h_f, adjacent to a porous medium of length L and height h_p (see Figure 2). The equations of continuity of mass and momentum for the mixture are taken as

$$\nabla.\,(\rho\mathbf{v}) = 0, \tag{2.4}$$

$$\nabla.\,(\rho\mathbf{v}\otimes\mathbf{v}) = -\nabla\left(p + \frac{2\mu}{3}\nabla.\mathbf{v}\right) + \mu\nabla^2\mathbf{v} - \rho g\mathbf{j}, \tag{2.5}$$

where g is the acceleration due to gravity and $\mathbf{j}$ is the unit vector in the positive y-direction; for later use, it is also convenient to define p', the modified pressure, given by

$$p' = p + \frac{2}{3}\mu\nabla.\mathbf{v}.$$

The continuity equation for each of the species,

$$\nabla.\mathbf{n}_i = 0, \quad i = 1,\ldots,3, \tag{2.6}$$

can be recast as, for the cathode of a fuel cell, with $n = 3$ in the form of two transport equations

$$\nabla.\left(\frac{\rho\mathbf{v}}{M}\begin{bmatrix} x_{O_2} \\ x_{H_2O}\end{bmatrix}\right) = \nabla.\left(\frac{\rho}{M^2}\mathbf{M}\begin{bmatrix}\nabla x_{O_2} \\ \nabla x_{H_2O}\end{bmatrix}\right), \tag{2.7}$$

where

$$\mathbf{M} = M_{N_2}\begin{pmatrix} D_{O_2,N_2} & D_{O_2,N_2} \\ D_{H_2O,N_2} & D_{H_2O,N_2}\end{pmatrix} - \begin{pmatrix} 0 & M_{H_2O}D_{O_2,H_2O} \\ M_{O_2}D_{H_2O,O_2} & 0\end{pmatrix}.$$

Here, use has been made of the relation $x_{O_2} + x_{N_2} + x_{H_2O} = 1$ to eliminate x_{N_2}, with the diffusion coefficients D_{O_2,H_2O}, D_{H_2O,O_2}, D_{H_2O,N_2}, and D_{O_2,N_2} given by (2.2).

2.3. Porous backing. For the porous region, volume-averaging of (2.4)–(2.7) along the lines of De Vidts and White [13] or Whitaker [43] is required. We present this at a moderate level of detail in order to sketch how the transport equations that are normally used, (2.23) below, can be arrived at; fuller details of analogous equations can be found elsewhere [13, 43]. First, let B be a quantity (either scalar, vector, or tensor) associated with the gas phase, and let the quantity $\langle B\rangle$ be the local volume (or superficial) average of B,

$$\langle B\rangle \equiv \frac{1}{\mathcal{V}}\int_{\mathcal{V}^{(\mathrm{g})}} B d\mathcal{V}, \tag{2.8}$$

and let $\langle B\rangle^{(\mathrm{g})}$ be the intrinsic volume average of B in the gas phase,

$$\langle B\rangle^{(\mathrm{g})} \equiv \frac{1}{\mathcal{V}^{(\mathrm{g})}}\int_{\mathcal{V}^{(\mathrm{g})}} B d\mathcal{V}. \tag{2.9}$$

Also, let γ be the porosity, given by $\gamma = \mathcal{V}^{(\mathrm{g})}/\mathcal{V}$. A comparison of (2.8) and (2.9) shows that the local and intrinsic volume average for the gas phase is given by

$$\langle B\rangle = \gamma \langle B\rangle^{(\mathrm{g})}. \tag{2.10}$$

Taking the superficial average of (2.4) gives

$$\langle \nabla.(\rho \mathbf{v})\rangle = 0, \tag{2.11}$$

whilst the superficial average of (2.5) gives (cf. [42])

$$\langle \mathbf{v}\rangle = -\frac{\mathbf{K}}{\mu}.\left(\nabla \langle p'\rangle^{(\mathrm{g})} + \langle \rho\rangle^{(\mathrm{g})} g\mathbf{j}\right) + \mathbf{K}.\nabla^2\left(\frac{\langle \mathbf{v}\rangle}{\gamma}\right) - \mathbf{F}.\langle \mathbf{v}\rangle, \tag{2.12}$$

where $\mathbf{K}$ is the Darcy Law permeability tensor and $\mathbf{F}$ is the Forchheimer correction tensor. Writing $\widetilde{D}_{ij} = c^2 D_{ij}/\rho$, we have

$$\nabla.\langle c_i \mathbf{v}\rangle + \frac{1}{\mathcal{V}}\int_{A_{gs}} c_i \mathbf{n}_{gs}.\mathbf{v} dA + \sum_{j=1}^{n} M_i M_j \nabla.\left(\left\langle \widetilde{D}_{ij}\right\rangle\left[\nabla\langle x_j\rangle + \frac{1}{\mathcal{V}}\int_{A_{gs}} x_j \mathbf{n}_{gs} dA\right]\right)$$

$$+\frac{1}{\mathcal{V}}\sum_{j=1}^{n} M_i M_j \widetilde{D}_{ij}\int_{A_{gs}} \mathbf{n}_{gs}.\nabla x_j dA = 0, \tag{2.13}$$

where $\mathbf{n}_{gs}$ represents the unit normal vector pointing from the gas phase to the solid phase, and A_{gs} represents the area of the gas-solid interface contained within $\mathcal{V}$. In the absence of surface reactions and zero normal velocity (passive dispersion), this reduces to

$$\nabla.\langle c_i \mathbf{v}\rangle + \sum_{j=1}^{n} M_i M_j \nabla.\left(\left\langle \widetilde{D}_{ij}\right\rangle\left[\nabla\langle x_j\rangle + \frac{1}{\mathcal{V}}\int_{A_{gs}} x_j \mathbf{n}_{gs} dA\right]\right) = 0, \tag{2.14}$$

and then

$$\nabla.\langle c_i \mathbf{v}\rangle + \sum_{j=1}^{n} M_i M_j \nabla.\left(\widetilde{D}_{ij}\left[\nabla\left(\gamma\langle x_j\rangle^{(\mathrm{g})}\right) + \frac{1}{\mathcal{V}}\int_{A_{gs}} x_j \mathbf{n}_{gs} dA\right]\right) = 0, \tag{2.15}$$

where we have used the fact that $\widetilde{D}_{ij}$ changes slowly with temperature and mole fraction within the representative elementary volume in order to be able to write

$$\nabla . \left\langle \widetilde{D}_{ij} \nabla x_j \right\rangle = \nabla . \left(\widetilde{D}_{ij} \left\langle \nabla x_j \right\rangle \right). \tag{2.16}$$

To ascertain this, we generalize the reasoning given by Whitaker [43] as follows. With

$$\begin{aligned} \widetilde{D}_{ij} &= \frac{pD_{ij}}{RTM} \\ &= \frac{p\mathcal{D}_{ij}}{RTM} \left\{ 1 + \frac{x_k \left[\left(M_k / M_j \right) \mathcal{D}_{ik} - \mathcal{D}_{ij} \right]}{x_i \mathcal{D}_{jk} + x_j \mathcal{D}_{ik} + x_k \mathcal{D}_{ij}} \right\}, \end{aligned}$$

we require

$$\frac{l_\gamma}{\widetilde{D}_{ij}} \left(\left. \frac{\partial \widetilde{D}_{ij}}{\partial p} \right]_{\langle p \rangle^{(\mathrm{g})}} \nabla \langle p \rangle^{(\mathrm{g})} + \left. \frac{\partial \widetilde{D}_{ij}}{\partial T} \right]_{\langle T \rangle^{(\mathrm{g})}} \nabla \langle T \rangle^{(\mathrm{g})} + \sum_{l=1}^{n} \left. \frac{\partial \widetilde{D}_{ij}}{\partial x_l} \right]_{\langle x_l \rangle^{(\mathrm{g})}} \nabla \langle x_l \rangle^{(\mathrm{g})} \right) \ll 1,$$

where l_γ is the pore length scale. Also, for later use, we need to be able to justify that within the representative elementary volume,

$$\begin{aligned} \nabla . \langle c_i \mathbf{v} \rangle &= \nabla . \left(\frac{p}{RT} \langle x_i \mathbf{v} \rangle \right) \\ &= \nabla . \left(\frac{\rho}{M} \langle x_i \mathbf{v} \rangle \right) \\ &= \nabla . \left(\frac{\langle \rho \rangle}{\langle M \rangle} \langle x_i \mathbf{v} \rangle \right); \end{aligned}$$

this would be justified if

$$\frac{l_\gamma}{D_\gamma} \left(\left. \frac{\partial D_\gamma}{\partial p} \right]_{\langle p \rangle^{(\mathrm{g})}} \nabla \langle p \rangle^{(\mathrm{g})} + \left. \frac{\partial D_\gamma}{\partial T} \right]_{\langle T \rangle^{(\mathrm{g})}} \nabla \langle T \rangle^{(\mathrm{g})} \right) \ll 1,$$

where $D_\gamma = p/RT$. Thus, we would require

$$l_\gamma \left(\frac{\nabla \langle p \rangle^{(\mathrm{g})}}{p} - \frac{\nabla \langle T \rangle^{(\mathrm{g})}}{T} \right) \ll 1; \tag{2.17}$$

we verify that this relation is indeed satisfied in section 3.2.

Now, decomposing according to

$$\phi = \langle \phi \rangle^{(\mathrm{g})} + \phi',$$

where $\phi = (x_j, c_j, \mathbf{v}, \rho)$ and the primed quantities denote spatial fluctuations, (2.11) and (2.15) can be shown to become, respectively,

$$\nabla . \left(\langle \rho \rangle^{(\mathrm{g})} \langle \mathbf{v} \rangle \right) = -\nabla . \left(\langle \rho' \mathbf{v}' \rangle \right), \tag{2.18}$$

$$\nabla.\left(\frac{\gamma\left\langle\rho\right\rangle^{(\mathrm{g})}\left\langle x_i\right\rangle^{(\mathrm{g})}\left\langle\mathbf{v}\right\rangle^{(\mathrm{g})}}{\left\langle M\right\rangle^{(\mathrm{g})}}\right)$$
$$+\sum_{j=1}^{n}M_iM_j\nabla.\left(\widetilde{D}_{ij}\left[\nabla\left(\gamma\left\langle x_j\right\rangle^{(\mathrm{g})}\right)+\frac{1}{\mathcal{V}}\int_{A_{gs}}x_j'\mathbf{n}_{gs}dA\right]\right)$$
$$-\nabla.\left(\frac{\left\langle\rho\right\rangle^{(\mathrm{g})}}{\left\langle M\right\rangle^{(\mathrm{g})}}\left\langle\mathbf{v}'x_i'\right\rangle\right)=0,$$

(2.19)

where we again use analysis due to Whitaker [43, pp. 14–20]. To keep the ongoing discussion simple, we assume henceforth that γ is constant. Eventually, we arrive at

(2.20)
$$\nabla.\left(\frac{\gamma\left\langle\rho\right\rangle^{(\mathrm{g})}\left\langle x_i\right\rangle^{(\mathrm{g})}\left\langle\mathbf{v}\right\rangle^{(\mathrm{g})}}{\left\langle M\right\rangle^{(\mathrm{g})}}\right)+\sum_{j=1}^{n}M_iM_j\nabla.\left(\left[\mathbf{D}_{ij}^{\mathrm{eff}}+\gamma\mathbf{D}_j^{\mathrm{hyd}}\delta_{ij}\right]\nabla\left\langle x_j\right\rangle^{(\mathrm{g})}\right)=0,$$

where $\mathbf{D}_{ij}^{\mathrm{eff}}$ is an effective diffusivity tensor given by

$$\mathbf{D}_{ij}^{\mathrm{eff}}=\gamma\widetilde{D}_{ij}\left(1+\frac{1}{\mathcal{V}^{(\mathrm{g})}}\int_{A_{gs}}\mathbf{n}_{gs}\mathbf{b}_gdA\right),$$

and δ_{ij} is the Kronecker delta; here, $\mathbf{b}_g$ is referred to as the closure variable and is found from the so-called closure problem. $\mathbf{D}_j^{\mathrm{hyd}}$ is called the hydrodynamic dispersion tensor and is defined by

$$\mathbf{D}_j^{\mathrm{hyd}}:=-\frac{\left\langle\rho\right\rangle^{(\mathrm{g})}}{\left\langle M\right\rangle^{(\mathrm{g})}}\left\langle\mathbf{v}'x_j'\right\rangle^{(\mathrm{g})}.$$

For gas diffusion electrodes, the following is often used [3, 5, 6, 45]:

$$\mathbf{D}_{ij}^{\mathrm{eff}}=\widetilde{D}_{ij}\gamma^{\frac{3}{2}};$$

this would imply

$$\left(1+\frac{1}{\mathcal{V}^{(\mathrm{g})}}\int_{A_{gs}}\mathbf{n}_{gs}\mathbf{b}_gdA\right)=\gamma^{\frac{1}{2}}.$$

For the cathode, with $i=O_2,N_2$, and H_2O, we have in more expedient form, on assuming the permeability to be isotropic and constant and neglecting the Forchheimer correction term in (2.12) and dispersion terms in (2.18) and (2.19) (see section 3.4),

$$\nabla.\left(\left\langle\rho\right\rangle^{(\mathrm{g})}\left\langle\mathbf{v}\right\rangle\right)=0, \tag{2.21}$$

$$\left\langle\mathbf{v}\right\rangle=-\frac{\kappa}{\mu}\left(\nabla\left\langle p'\right\rangle^{(\mathrm{g})}+\left\langle\rho\right\rangle^{(\mathrm{g})}g\mathbf{j}\right)+\frac{\kappa}{\gamma}\nabla^2\left\langle\mathbf{v}\right\rangle, \tag{2.22}$$

(2.23)
$$\nabla.\left(\frac{\left\langle\rho\right\rangle^{(\mathrm{g})}\left\langle\mathbf{v}\right\rangle}{\left\langle M\right\rangle^{(\mathrm{g})}}\begin{bmatrix}\left\langle x_{O_2}\right\rangle^{(\mathrm{g})}\\ \left\langle x_{H_2O}\right\rangle^{(\mathrm{g})}\end{bmatrix}\right)=\nabla.\left(\frac{\gamma^{\frac{3}{2}}\left\langle\rho\right\rangle^{(\mathrm{g})}}{\left(\left\langle M\right\rangle^{(\mathrm{g})}\right)^2}\left\langle\mathbf{M}\right\rangle^{(\mathrm{g})}\begin{bmatrix}\nabla\left\langle x_{O_2}\right\rangle^{(\mathrm{g})}\\ \nabla\left\langle x_{H_2O}\right\rangle^{(\mathrm{g})}\end{bmatrix}\right),$$

where

$$\langle \mathbf{M} \rangle^{(\mathrm{g})} = M_{N_2} \begin{pmatrix} \langle D_{O_2,N_2} \rangle^{(\mathrm{g})} & \langle D_{O_2,N_2} \rangle^{(\mathrm{g})} \\ \langle D_{H_2O,N_2} \rangle^{(\mathrm{g})} & \langle D_{H_2O,N_2} \rangle^{(\mathrm{g})} \end{pmatrix} - \begin{pmatrix} 0 & M_{H_2O} \langle D_{O_2,H_2O} \rangle^{(\mathrm{g})} \\ M_{O_2} \langle D_{H_2O,O_2} \rangle^{(\mathrm{g})} & 0 \end{pmatrix},$$

with

$$\langle D_{ij} \rangle^{(\mathrm{g})} = \mathcal{D}_{ij} \left\{ 1 + \frac{\langle x_k \rangle^{(\mathrm{g})} \left[(M_k/M_j)\, \mathcal{D}_{ik} - \mathcal{D}_{ij} \right]}{\langle x_i \rangle^{(\mathrm{g})} \mathcal{D}_{jk} + \langle x_j \rangle^{(\mathrm{g})} \mathcal{D}_{ik} + \langle x_k \rangle^{(\mathrm{g})} \mathcal{D}_{ij}} \right\};$$

(2.21)–(2.23) are then akin to the governing equations for the porous backing used by most authors, although with more attention having been paid here to the distinction between intrinsic and superficial variables, the possibility of nonconstant diffusion coefficients, and the inclusion of crossed diffusion terms.

2.4. Boundary conditions.

2.4.1. Inlet, outlet, upper wall, vertical walls. For boundary conditions in the channel, we prescribe inlet velocity and gas composition at $x = 0$, $0 \le y \le h_f$, so that

$$u = U^{in}, \quad v = 0, \quad x_{O_2} = x^{in}_{O_2}, \quad x_{H_2O} = x^{in}_{H_2O}, \tag{2.24}$$

where $\mathbf{v} = (u, v)$. At the upper channel wall ($0 \le x \le L$, $y = h_f$), there is no slip, no normal flow, and no componental flux, so that

$$u = v = \frac{\partial x_{O_2}}{\partial y} = \frac{\partial x_{H_2O}}{\partial y} = 0. \tag{2.25}$$

At the outlet at $x = L, 0 \le y \le h_f$, we have constant pressure, and no diffusive componental flux, so that

$$p = p^{out}, \quad \frac{\partial v}{\partial x} = \frac{\partial x_{O_2}}{\partial x} = \frac{\partial x_{H_2O}}{\partial x} = 0. \tag{2.26}$$

At the vertical walls of the porous electrode ($x = 0, L, -h_p \le y \le 0$), we prescribe no normal flow, no tangential shear, and no mass flux for the gas components, so that

$$\langle u \rangle = \frac{\partial \langle v \rangle}{\partial x} = \frac{\partial \langle x_{O_2} \rangle^{(\mathrm{g})}}{\partial x} = \frac{\partial \langle x_{H_2O} \rangle^{(\mathrm{g})}}{\partial x} = 0, \tag{2.27}$$

where $\langle \mathbf{v} \rangle = (\langle u \rangle, \langle v \rangle)$.

2.4.2. Channel/porous backing interface. In addition, matching conditions are required for the fluid-porous interface at $y = 0$, $0 \le x \le L$. The conditions for continuity of normal velocity and normal stress are given, respectively, as

$$v = \langle v \rangle, \tag{2.28}$$

$$p - \mu \frac{\partial v}{\partial y} = \langle p \rangle^{(\mathrm{g})} - \mu_{eff} \frac{\partial \langle v \rangle}{\partial y}, \tag{2.29}$$

where $\mu_{eff}\,(=\mu/\gamma)$ is termed the effective viscosity of the porous medium. The remaining two conditions that are required have been the subject of longstanding debate ever since the work of Beavers and Joseph [4]; a recent contribution is due to Jäger and Mikelić [24]. A summary of possible options for the momentum equation is given by Alazmi and Vafai [1], of which the most relevant for this application, and indeed most consistent in view of our use of the full Navier–Stokes equations for the fluid and a Darcy/Brinkman/Forchheimer formulation for the porous medium, is one due to Ochoa-Tapia and Whitaker [32] when inertial effects are important:

$$u = \langle u \rangle\,, \tag{2.30}$$

$$\frac{\mu}{\gamma}\frac{\partial \langle u \rangle}{\partial y} - \mu\frac{\partial u}{\partial y} = \frac{\beta_1 \mu}{\kappa^{\frac{1}{2}}}u + \beta_2 \rho u^2, \tag{2.31}$$

respectively. Here, β_1 and β_2 are $O\,(1)$ constants which would need to be determined experimentally, although it turns out here that the leading order problem is dictated more by (2.30) than by (2.31) .

Finally, volume-averaging techniques at the interface analogous to those used for heat transfer by [33] are required for the mole fraction transport equations. We do not pursue the details but simply assume the point values for the mole fractions of O_2 and H_2O in the channel to be equal to their intrinsic values in the porous backing, so that

$$\langle x_{O_2} \rangle^{(\mathrm{g})} = x_{O_2}, \quad \langle x_{H_2O} \rangle^{(\mathrm{g})} = x_{H_2O} \text{ at } y = 0, \tag{2.32}$$

and, in addition, that the point values for the mole fraction fluxes of O_2 and H_2O are equal to their superficial values in the porous medium, so that

$$\mathbf{n}_{O_2}.\mathbf{n} = \langle \mathbf{n}_{O_2}.\mathbf{n} \rangle\,, \quad \mathbf{n}_{H_2O}.\mathbf{n} = \langle \mathbf{n}_{H_2O}.\mathbf{n} \rangle\,;$$

using (2.28) and (2.32) , we arrive at

$$\gamma^{\frac{3}{2}}\frac{\partial}{\partial y}\begin{bmatrix} \langle x_{O_2} \rangle^{(\mathrm{g})} \\ \langle x_{H_2O} \rangle^{(\mathrm{g})} \end{bmatrix} = \frac{\partial}{\partial y}\begin{bmatrix} x_{O_2} \\ x_{H_2O} \end{bmatrix}, \tag{2.33}$$

respectively.

2.4.3. Catalyst/porous backing interface. At $y = -h_f$, we would expect $\langle u \rangle\,, \langle v \rangle\,, \langle x_{O_2} \rangle^{(\mathrm{g})}$, and $\langle x_{H_2O} \rangle^{(\mathrm{g})}$ to match to their counterparts in the catalytic layer, although naturally this approach would require us to model the catalyst layer and then by extension the membrane and the corresponding regions on the anode side. This has been done to varying degrees by various authors [5, 6, 14, 17, 18, 19, 20, 23, 31, 35, 36, 37]. An alternative approach, often adopted when the flow field in the porous backing and gas channels rather than the electrochemistry in the catalyst and the membrane is of interest [25, 30, 45, 41], is to prescribe a current density, I, at this interface. Using Faraday's Law, the superficial mass flux of oxygen is given as a function of current density, so that

$$\langle \mathbf{n}_{O_2}.\mathbf{n} \rangle = -\frac{M_{O_2} I}{4F}, \tag{2.34}$$

where F is the Faraday constant. The corresponding expression for water is then taken to be

$$\langle \mathbf{n}_{H_2O}.\mathbf{n} \rangle = \frac{M_{H_2O}(1+2\alpha)I}{2F}, \tag{2.35}$$

where α is a parameter accounting for the water transport by electro-osmosis in the membrane; typical values encountered in the literature are $\alpha = 0.3$ [41] and $0.5 \leq \alpha \leq 1.7$ [30, 44, 45]. Furthermore, since nitrogen does not participate in the reaction at the catalyst layer,

$$\langle \mathbf{n}_{N_2}.\mathbf{n} \rangle = 0. \tag{2.36}$$

This leads to the following boundary conditions for $\langle v \rangle$, $\langle x_{O_2} \rangle^{(\mathrm{g})}$, and $\langle x_{H_2O} \rangle^{(\mathrm{g})}$:

$$\langle \rho \rangle^{(\mathrm{g})} \langle v \rangle = \frac{I}{4F} \left(2(1+2\alpha) M_{H_2O} - M_{O_2} \right) \tag{2.37}$$

and

$$\frac{\langle \rho \rangle^{(\mathrm{g})} \langle v \rangle}{\langle M \rangle^{(\mathrm{g})}} \begin{bmatrix} \langle x_{O_2} \rangle^{(\mathrm{g})} \\ \langle x_{H_2O} \rangle^{(\mathrm{g})} \end{bmatrix} - \frac{\gamma^{\frac{3}{2}} \langle \rho \rangle^{(\mathrm{g})}}{\left(\langle M \rangle^{(\mathrm{g})} \right)^2} \bar{\mathbf{M}} \frac{\partial}{\partial y} \begin{bmatrix} \langle x_{O_2} \rangle^{(\mathrm{g})} \\ \langle x_{H_2O} \rangle^{(\mathrm{g})} \end{bmatrix} = \frac{I}{4F} \begin{bmatrix} -1 \\ 2(1+2\alpha) \end{bmatrix}.$$

3. Analysis.

3.1. Nondimensionalization.

Writing

$$\tilde{x} = \frac{x}{L}, \quad \tilde{y} = \frac{y}{L}, \quad \tilde{\mathbf{v}} = \frac{\mathbf{v}}{U^{in}}, \quad \langle \tilde{\mathbf{v}} \rangle = \frac{\langle \mathbf{v} \rangle}{U^{in}}, \; \tilde{\rho} = \frac{\rho}{[\rho]}, \; \langle \tilde{\rho} \rangle^{(\mathrm{g})} = \frac{\langle \rho \rangle^{(\mathrm{g})}}{[\rho]},$$

$$\tilde{p} = \frac{p - p^{out}}{[\rho] \left(U^{in} \right)^2}, \; \langle \tilde{p} \rangle^{(\mathrm{g})} = \frac{\langle p \rangle^{(\mathrm{g})} - p^{out}}{[\rho] \left(U^{in} \right)^2}, \; \tilde{p}' = \frac{p' - p^{out}}{[\rho] \left(U^{in} \right)^2}, \; \langle \tilde{p}' \rangle^{(\mathrm{g})} = \frac{\langle p' \rangle^{(\mathrm{g})} - p^{out}}{[\rho] \left(U^{in} \right)^2},$$

$$\widetilde{I} = \frac{I}{[I]}, \quad \mathcal{M} = \frac{M}{[M]}, \; \langle \mathcal{M} \rangle^{(\mathrm{g})} = \frac{\langle M \rangle^{(\mathrm{g})}}{[M]}, \quad \tilde{c} = \frac{c}{[\rho] / [M]},$$

$$\mathcal{M}_i = \frac{M_i}{[M]}, \; i = 1, \ldots, 3, \quad \tilde{D}_{ij} = \frac{D_{ij}}{[D]}, \; i, j = 1, \ldots, 3, \quad \tilde{D}_{ij}^{eff} = \frac{D_{ij}^{eff}}{[D]}, \; i, j = 1, \ldots, 3,$$

$$Re = \frac{[\rho] U^{in} L}{\mu}, \quad Sc = \frac{\mu}{[\rho] [D]}, \quad Da = \frac{\kappa}{L^2}, \quad Fr = \frac{U^2}{gL},$$

$$\tilde{\mathbf{M}} = \frac{\mathbf{M}}{[M] [D]}, \; \left\langle \tilde{\mathbf{M}} \right\rangle^{(\mathrm{g})} = \frac{\langle \mathbf{M} \rangle^{(\mathrm{g})}}{[M] [D]},$$

where $[\rho]$ is a density scale, $[D]$ is a diffusion scale, $[I]$ is a current density scale, and $[M]$ is a molecular weight scale (all to be either determined or specified shortly) and Re, Sc, Da, and Fr are the Reynolds, Schmidt, Darcy, and Froude numbers,

respectively, we drop the tildes and arrive at the following nondimensionalized forms. For the channel ($0 \le x \le 1$, $0 \le y \le 1$),

$$\nabla.(\rho \mathbf{v}) = 0, \tag{3.1}$$

$$\nabla.(\rho \mathbf{v} \otimes \mathbf{v}) = -\nabla\left(p + \frac{2\delta^2}{3}\nabla.\mathbf{v}\right) + \delta^2\nabla^2\mathbf{v} - Fr^{-1}\rho\mathbf{j}, \tag{3.2}$$

$$\nabla.\left(\frac{\rho\mathbf{v}}{\mathcal{M}}\begin{bmatrix} x_{O_2} \\ x_{H_2O}\end{bmatrix}\right) = \frac{\delta^2}{Sc}\nabla.\left(\frac{\rho}{\mathcal{M}^2}\mathbf{M}\begin{bmatrix}\nabla x_{O_2} \\ \nabla x_{H_2O}\end{bmatrix}\right), \tag{3.3}$$

where $\delta^2 = Re^{-1}$, and for the porous medium ($0 \le x \le 1$, $-h_p/L \le y \le 0$),

$$\nabla.\left(\langle\rho\rangle^{(g)}\langle\mathbf{v}\rangle\right) = 0, \tag{3.4}$$

$$\frac{\delta^2}{\epsilon^2}\langle\mathbf{v}\rangle = -\nabla\langle p'\rangle^{(g)} + \delta^2\nabla^2\left(\frac{\langle\mathbf{v}\rangle}{\gamma}\right) - Fr^{-1}\langle\rho\rangle^{(g)}\mathbf{j},$$

$$\nabla.\left(\frac{\langle\rho\rangle^{(g)}\langle\mathbf{v}\rangle}{\langle\mathcal{M}\rangle^{(g)}}\begin{bmatrix}\langle x_{O_2}\rangle^{(g)} \\ \langle x_{H_2O}\rangle^{(g)}\end{bmatrix}\right) = \frac{\delta^2}{Sc}\nabla.\left(\gamma^{\frac{3}{2}}\frac{\langle\rho\rangle^{(g)}}{\left(\langle\mathcal{M}\rangle^{(g)}\right)^2}\langle\mathbf{M}\rangle^{(g)}\begin{bmatrix}\nabla\langle x_{O_2}\rangle^{(g)} \\ \nabla\langle x_{H_2O}\rangle^{(g)}\end{bmatrix}\right), \tag{3.5}$$

where $\epsilon^2 = Da$. The boundary conditions are now

$$u = 1, \quad v = 0, \quad x_{O_2} = x_{O_2}^{in}, \quad x_{H_2O} = x_{H_2O}^{in} \quad \text{at} \quad x = 0, \quad 0 \le y \le h_f/L; \tag{3.6}$$

$$u = v = \frac{\partial x_{O_2}}{\partial y} = \frac{\partial x_{H_2O}}{\partial y} = 0 \quad \text{at} \quad 0 \le x \le 1, \quad y = h_f/L; \tag{3.7}$$

$$p = 0, \quad \frac{\partial v}{\partial x} = \frac{\partial x_{O_2}}{\partial x} = \frac{\partial x_{H_2O}}{\partial x} = 0 \quad \text{at} \quad x = 1, \quad 0 \le y \le h_f/L; \tag{3.8}$$

$$\langle u\rangle = \frac{\partial\langle v\rangle}{\partial x} = \frac{\partial\langle x_{O_2}\rangle^{(g)}}{\partial x} = \frac{\partial\langle x_{H_2O}\rangle^{(g)}}{\partial x} = 0 \quad \text{at} \quad x = 0, 1, \quad -h_p/L \le y \le 0. \tag{3.9}$$

The boundary conditions for $0 \le x \le 1$, $y = -h_p/L$ are now

$$\langle u\rangle = 0, \quad \langle\rho\rangle^{(g)}\langle v\rangle = \Lambda\left\{\frac{I}{4}\left(2(1+2\alpha)\mathcal{M}_{H_2O} - \mathcal{M}_{O_2}\right)\right\}, \tag{3.10}$$

$$\frac{\langle\rho\rangle^{(g)}\langle\mathbf{v}\rangle}{\langle\mathcal{M}\rangle^{(g)}}\begin{bmatrix}\langle x_{O_2}\rangle^{(g)} \\ \langle x_{H_2O}\rangle^{(g)}\end{bmatrix} - \frac{\delta^2\gamma^{\frac{3}{2}}\langle\rho\rangle^{(g)}}{Sc\left(\langle\mathcal{M}\rangle^{(g)}\right)^2}\langle\mathbf{M}\rangle^{(g)}\frac{\partial}{\partial y}\begin{bmatrix}\langle x_{O_2}\rangle^{(g)} \\ \langle x_{H_2O}\rangle^{(g)}\end{bmatrix} \tag{3.11}$$

$$= \frac{\Lambda I}{4}\begin{pmatrix}-1 \\ 2(1+2\alpha)\end{pmatrix},$$

TABLE 1
Scales for nondimensionalization.

$[\rho]$	1 kgm^{-3}
$[M]$	$[\rho]\, RT/p^{out}$
$[I]$	10^4 Am^{-2}

where $\Lambda = [I]\,[M]\,/FU^{in}\,[\rho]$. Finally, the boundary conditions along the fluid-porous interface on $y = 0$ reduce to

$$v = \langle v \rangle\,, \tag{3.12}$$

$$p - \delta^2 \frac{\partial v}{\partial y} = \langle p \rangle^{(\mathrm{g})} - \delta^2 \frac{\partial \langle v \rangle}{\partial y}, \tag{3.13}$$

$$u = \langle u \rangle\,, \tag{3.14}$$

$$\frac{1}{\gamma}\frac{\partial \langle u \rangle}{\partial y} - \frac{\partial u}{\partial y} = \left(\frac{\beta_1}{\varepsilon}\right) u + \left(\frac{\beta_2}{\delta^2}\right) \rho u^2 \tag{3.15}$$

and

$$\langle x_{O_2} \rangle^{(\mathrm{g})} = x_{O_2}, \quad \langle x_{H_2O} \rangle^{(\mathrm{g})} = x_{H_2O}, \tag{3.16}$$

$$\gamma^{\frac{3}{2}} \frac{\partial}{\partial y} \begin{bmatrix} \langle x_{O_2} \rangle^{(\mathrm{g})} \\ \langle x_{H_2O} \rangle^{(\mathrm{g})} \end{bmatrix} = \frac{\partial}{\partial y} \begin{bmatrix} x_{O_2} \\ x_{H_2O} \end{bmatrix}. \tag{3.17}$$

3.2. Parameters. Typically, $U^{in} \sim 1$ ms^{-1}, $h_f \sim 10^{-3}$ m, $h_p \sim 3 \times 10^{-4}$ m, $L \geq 10^{-2}$ m, $[I] \sim 10^4$ Am^{-2}, $p^{out} \sim 1$ atm $\sim 10^5$kgm^{-1}s^{-2}, $T \sim 300$–350 K, $0.1 \leq \gamma \leq 0.5$, $0.3 \leq \alpha \leq 1.7$, $\mu \sim O(10^{-5})$ kgm$^{-1}s^{-1}$. In addition, $M_{O_2} = 0.032$ kgmol^{-1}, $M_{H_2O} = 0.018$ kgmol^{-1}, $M_{N_2} = 0.028$ kgmol^{-1}, $F = 96487$ Asmol^{-1}, from which we note that $M_{min} \leq M \leq M_{max}$, where

$$M_{min} = M|_{x_{H_2O}=1, x_{O_2}=0} = 0.018 \text{ kgmol}^{-1},$$
$$M_{max} = M|_{x_{H_2O}=0, x_{O_2}=1} = 0.032 \text{ kgmol}^{-1}.$$

Further, we use the constitutive relation for an ideal gas in order to obtain the density scale $[\rho]$; with $p \sim p^{out}$, we have $\rho \sim 1$ kgm^{-3}, so that $[\rho] \sim 1$ kgm^{-3} seems appropriate. For $[D]$, we take $O\left(10^{-5}\right)$ m^{2}s^{-1} from available literature, e.g., [5, 6]. Note also that the relation (2.17) is satisfied, since the smallest length on the macroscale in the porous backing, h_p, is still much larger than the scale for l_γ, 10^{-5}–10^{-6}m, suggested by the electrochemical literature [16, 40]. The scales used for nondimensionalisation and the physical parameters are summarized in Tables 1 and 2, respectively.

Thence, for the nondimensional parameters Re, Sc, Da, Fr, Λ, we arrive at

$$Re \sim 10^4, \quad Sc \sim 1, \quad Da \leq 10^{-6}, \quad Fr \sim 1, \quad \Lambda \leq 10^{-2},$$

so that $\delta \sim 10^{-2}$ and $\epsilon \leq 10^{-3}$. We note here that some of these parameters have been encountered before in conjunction with the modeling of flow in SOFC [7, 11], in particular, the Reynolds number, Re, which represents the ratio of inertial to viscous forces and the product Schmidt $ReSc$, which is the ratio of gas flow rate to the rate of diffusion. (In fact $ReSc$ in our formulation corresponds to the parameter Q in [7, 11,

TABLE 2
Physical parameters.

M_{O_2}	0.032 kgmol^{-1}
M_{H_2O}	0.018 kgmol^{-1}
M_{N_2}	0.028 kgmol^{-1}
$\mathcal{D}_{O_2,H_2O}$	$3.749[M]/RT$
$\mathcal{D}_{O_2,N_2}$	$2.827[M]/RT$
$\mathcal{D}_{H_2O,N_2}$	$3.923[M]/RT$

26].) Furthermore, the parameter Λ is a measure of the ratio of the electrochemical flux of oxygen to the gas flow rate and thus corresponds to the combination E/Q in [7, 11, 26]. For completeness, we mention that the Froude number, *Fr*, is the ratio of inertial to gravitational forces, whereas the Darcy number, *Da*, is the ratio of the porous medium permeability to the square of the length scale of the entire geometry.

3.3. Narrow-gap approximation. Typically, $h_f/L, h_p/L \ll 1$, which leads us to further rescaling as follows. Writing

$$X = x, \quad Y = \frac{y}{\sigma}, \quad U = u, \; V = \frac{v}{\sigma}, \quad \langle U \rangle = \langle u \rangle, \quad \langle V \rangle = \frac{\langle v \rangle}{\sigma},$$
$$P = p, \; P' = p', \quad \langle P \rangle = \langle p \rangle, \quad \langle P' \rangle = \langle p' \rangle,$$

where $\sigma = h_f/L$, we simplify further by neglecting terms in $O(\sigma)$ or lower, although we retain for the time being terms which contain multiples of σ and the other dimensionless parameters. We introduce the dimensionless parameters Δ, Σ, and Ω, given by

$$\Delta = \delta^2/\sigma^2, \; \Sigma = \sigma^2/\varepsilon, \; \Omega = \Lambda/\sigma,$$

and that an alternative expression for Δ is $\Delta = \left(Re\sigma^2\right)^{-1}$, i.e., the reciprocal of the reduced Reynolds number. We have now, for the channel,

$$\frac{\partial}{\partial X}(\rho U) + \frac{\partial}{\partial Y}(\rho V) = 0, \tag{3.18}$$

$$\rho\left(U\frac{\partial U}{\partial X} + V\frac{\partial U}{\partial Y}\right) = -\frac{\partial P'}{\partial X} + \Delta\frac{\partial^2 U}{\partial Y^2}, \tag{3.19}$$

$$0 = -\frac{\partial P'}{\partial Y}, \tag{3.20}$$

$$\left(\rho U\frac{\partial}{\partial X} + \rho V\frac{\partial}{\partial Y}\right)\begin{bmatrix} 1 & x_{O_2} \\ \mathcal{M} & x_{H_2O} \end{bmatrix} = \frac{\Delta}{Sc}\frac{\partial}{\partial Y}\left(\frac{\rho}{\mathcal{M}^2}\mathbf{M}\frac{\partial}{\partial Y}\begin{bmatrix} x_{O_2} \\ x_{H_2O} \end{bmatrix}\right) \tag{3.21}$$

and for the porous medium,

$$0 = \frac{\partial}{\partial X}\left(\langle\rho\rangle^{(\mathrm{g})}\langle U\rangle\right) + \frac{\partial}{\partial Y}\left(\langle\rho\rangle^{(\mathrm{g})}\langle V\rangle\right), \tag{3.22}$$

$$\langle U\rangle = -\frac{\epsilon}{\Delta\Sigma}\frac{\partial\langle P'\rangle^{(\mathrm{g})}}{\partial X} + \frac{\varepsilon}{\Sigma\gamma}\frac{\partial^2\langle U\rangle}{\partial Y^2}, \tag{3.23}$$

$$\langle V\rangle = -\frac{1}{\Delta\Sigma^2}\frac{\partial\langle P'\rangle^{(\mathrm{g})}}{\partial Y} + \frac{\varepsilon}{\Sigma\gamma}\frac{\partial^2\langle V\rangle}{\partial Y^2}, \tag{3.24}$$

$$\begin{aligned}\left(\langle\rho\rangle^{(\mathrm{g})}\langle U\rangle\frac{\partial}{\partial X} + \langle\rho\rangle^{(\mathrm{g})}\langle V\rangle\frac{\partial}{\partial Y}\right)\left(\frac{1}{\langle\mathcal{M}\rangle^{(\mathrm{g})}}\begin{bmatrix}\langle x_{O_2}\rangle^{(\mathrm{g})}\\ \langle x_{H_2O}\rangle^{(\mathrm{g})}\end{bmatrix}\right)\\ = \frac{\Delta}{Sc}\frac{\partial}{\partial Y}\left(\frac{\gamma^{\frac{3}{2}}\langle\rho\rangle^{(\mathrm{g})}}{\left(\langle\mathcal{M}\rangle^{(\mathrm{g})}\right)^2}\langle\mathbf{M}\rangle^{(\mathrm{g})}\frac{\partial}{\partial Y}\begin{bmatrix}\langle x_{O_2}\rangle^{(\mathrm{g})}\\ \langle x_{H_2O}\rangle^{(\mathrm{g})}\end{bmatrix}\right).\end{aligned} \tag{3.25}$$

Note also that

$$P' = P + O\left(\delta^2\right), \quad \langle P'\rangle^{(\mathrm{g})} = \langle P\rangle^{(\mathrm{g})} + O\left(\delta^2\right),$$

and since $\delta^2 \ll 1$, henceforth, we use the actual pressure rather than the modified pressure. In addition, the gravitational terms in (3.20) and (3.24) are $O\left(Fr^{-1}\sigma\right)$ and have therefore been dropped. The boundary conditions are, for $0 \le X \le 1, Y = 1$,

$$U = V = \frac{\partial x_{O_2}}{\partial Y} = \frac{\partial x_{H_2O}}{\partial Y} = 0; \tag{3.26}$$

for $0 \le X \le 1, Y = 0$,

$$V = \langle V\rangle, \tag{3.27}$$

$$P = \langle P\rangle^{(\mathrm{g})}, \tag{3.28}$$

$$U = \langle U\rangle, \tag{3.29}$$

$$\frac{1}{\gamma}\frac{\partial\langle U\rangle}{\partial Y} - \frac{\partial U}{\partial Y} = \left(\frac{\beta_1\sigma}{\varepsilon}\right)U + \left(\frac{\beta_2\sigma}{\delta^2}\right)\rho U^2, \tag{3.30}$$

$$\langle x_{O_2}\rangle^{(\mathrm{g})} = x_{O_2}, \quad \langle x_{H_2O}\rangle^{(\mathrm{g})} = x_{H_2O}, \tag{3.31}$$

$$\gamma^{\frac{3}{2}}\frac{\partial}{\partial Y}\begin{bmatrix}\langle x_{O_2}\rangle^{(\mathrm{g})}\\ \langle x_{H_2O}\rangle^{(\mathrm{g})}\end{bmatrix} = \frac{\partial}{\partial Y}\begin{bmatrix}x_{O_2}\\ x_{H_2O}\end{bmatrix}; \tag{3.32}$$

for $0 \le X \le 1, Y = -\mathcal{H}\,(= h_p/h_f)$,

$$\langle U\rangle = 0, \quad \langle\rho\rangle^{(\mathrm{g})}\langle V\rangle = \Omega\left\{\frac{I}{4}\left(2(1+2\alpha)\mathcal{M}_{H_2O} - \mathcal{M}_{O_2}\right)\right\}, \tag{3.33}$$

$$\begin{aligned}&\frac{\langle\rho\rangle^{(\mathrm{g})}\langle V\rangle}{\langle\mathcal{M}\rangle^{(\mathrm{g})}}\begin{bmatrix}\langle x_{O_2}\rangle^{(\mathrm{g})}\\ \langle x_{H_2O}\rangle^{(\mathrm{g})}\end{bmatrix} - \frac{\Delta\gamma^{\frac{3}{2}}\langle\rho\rangle^{(\mathrm{g})}}{Sc\left(\langle\mathcal{M}\rangle^{(\mathrm{g})}\right)^2}\langle\mathbf{M}\rangle^{(\mathrm{g})}\frac{\partial}{\partial Y}\begin{bmatrix}\langle x_{O_2}\rangle^{(\mathrm{g})}\\ \langle x_{H_2O}\rangle^{(\mathrm{g})}\end{bmatrix}\\ &= \frac{\Omega I}{4}\begin{bmatrix}-1\\ 2(1+2\alpha)\end{bmatrix}.\end{aligned} \tag{3.34}$$

The neglect of streamwise diffusion terms will of course imply that not all of the original boundary conditions at $X = 0$ and 1 in this reduced formulation can be satisfied, and those terms would need to be reinstated for $X \sim O(\sigma)$ and $1 - X \sim O(\sigma)$. This is beyond the scope of interest here, and for a consistent formulation we simply retain

$$U = 1, \quad x_{O_2} = x_{O_2}^{in}, \quad x_{H_2O} = x_{H_2O}^{in} \quad \text{at} \quad X = 0, \quad 0 \le Y \le 1, \tag{3.35}$$

$$\langle U \rangle = \frac{\partial \langle x_{O_2} \rangle^{(g)}}{\partial X} = \frac{\partial \langle x_{H_2O} \rangle^{(g)}}{\partial X} = 0 \quad \text{at} \quad X = 0, \quad -\mathcal{H} \le Y \le 0. \tag{3.36}$$

For the initial discussion, we proceed under the assumption that $\Sigma, \Delta, \Omega \sim O(1)$; later, we will also require $\Omega \gg 1$. Further simplification is now possible by noting from (3.23) that $\langle U \rangle = 0$ to leading order, which reduces (3.22), (3.24), and (3.25) still further. Turning to the porous region near $Y = 0_-$, there is no reason a priori to suppose that the porous core flow should satisfy (3.27)–(3.30); if it did, we would arrive at $U = \frac{\partial U}{\partial Y} = 0$ at $Y = 0$, and there would be too many boundary conditions for (U, V, P) in the channel. Instead, we require a porous boundary layer for which $Y \sim \varepsilon^{\frac{1}{2}}$, $\langle U \rangle \sim \varepsilon^{\frac{1}{2}}$. Writing

$$Y = \varepsilon^{\frac{1}{2}} \widetilde{Y}, \quad \langle U \rangle = \varepsilon^{\frac{1}{2}} \left\langle \widetilde{U} \right\rangle, \quad \langle P \rangle^{(g)} = \left\langle \widetilde{P} \right\rangle^{(g)}, \quad \langle V \rangle = \left\langle \widetilde{V} \right\rangle,$$

we have, to leading order, in this layer

$$\frac{\partial}{\partial \widetilde{Y}} \left(\langle \rho \rangle^{(g)} \left\langle \widetilde{V} \right\rangle \right) = 0, \tag{3.37}$$

$$\left\langle \widetilde{U} \right\rangle = \frac{\partial^2}{\partial \widetilde{Y}^2} \left(\frac{\left\langle \widetilde{U} \right\rangle}{\gamma} \right), \tag{3.38}$$

$$0 = -\frac{\partial \left\langle \widetilde{P} \right\rangle^{(g)}}{\partial \widetilde{Y}} \tag{3.39}$$

subject to the matching conditions as $\widetilde{Y} \longrightarrow -\infty$

$$\left\langle \widetilde{V} \right\rangle \longrightarrow \langle V \rangle (X, 0), \quad \left\langle \widetilde{U} \right\rangle \longrightarrow 0, \quad \left\langle \widetilde{P} \right\rangle^{(g)} \longrightarrow \langle P \rangle^{(g)} (X, 0),$$

where

$$\langle V \rangle (X, 0) = \lim_{Y \to 0_-} \langle V \rangle, \quad \langle P \rangle^{(g)} (X, 0) = \lim_{Y \to 0_-} \langle P \rangle^{(g)}.$$

At $Y = \widetilde{Y} = 0$, we have

$$V = \left\langle \widetilde{V} \right\rangle, \tag{3.40}$$

$$P = \langle P \rangle^{(g)}, \tag{3.41}$$

$$U = \varepsilon^{\frac{1}{2}} \left\langle \widetilde{U} \right\rangle, \tag{3.42}$$

$$\frac{1}{\gamma} \frac{\partial \left\langle \widetilde{U} \right\rangle}{\partial \widetilde{Y}} - \frac{\partial U}{\partial Y} = \left(\frac{\beta_1 \sigma}{\varepsilon} \right) U + \left(\frac{\beta_2 \sigma}{\delta^2} \right) \rho U^2. \tag{3.43}$$

These equations are then used in the following order. First, the channel flow is determined with boundary conditions, to leading order,

$$U = 0,\ V = \left\langle \widetilde{V} \right\rangle.$$

This gives $P(X)$ which serves a boundary condition for $\langle P \rangle^{(\mathrm{g})}$, and finally $\langle \widetilde{U} \rangle$ can be computed, the boundary condition for this being, at leading order, simply

$$\frac{\partial \left\langle \widetilde{U} \right\rangle}{\partial \widetilde{Y}} = \gamma \left(\frac{\partial U}{\partial Y} \right)_{Y=0} \text{ at } \widetilde{Y} = 0.$$

As for the species equations, no such boundary layer in ε is necessary, with (3.25) being valid all the way up to $Y = 0_-$. In addition, we note that the leading order equations are independent of β_1 and β_2.

3.4. Further simplifications and observations. We invoke the constitutive relation for an ideal gas in dimensionless variables, with

$$\rho = \mathcal{M} + \left(\frac{[\rho] \left(U^{in} \right)^2}{p^{out}} \right) P, \quad \langle \rho \rangle^{(\mathrm{g})} = \langle \mathcal{M} \rangle^{(\mathrm{g})} + \left(\frac{[\rho] \left(U^{in} \right)^2}{p^{out}} \right) \langle P \rangle^{(\mathrm{g})} \tag{3.44}$$

for the channel and porous medium, respectively, which can be reduced to just $\rho = \mathcal{M}$, $\langle \rho \rangle^{(\mathrm{g})} = \langle \mathcal{M} \rangle^{(\mathrm{g})}$, respectively, for the pressures and velocities being considered here. The reduced system of equations is now, for $0 \leq X \leq 1$, $0 \leq Y \leq 1$,

$$\frac{\partial}{\partial X}(\rho U) + \frac{\partial}{\partial Y}(\rho V) = 0, \tag{3.45}$$

$$\rho \left(U \frac{\partial U}{\partial X} + V \frac{\partial U}{\partial Y} \right) = -\frac{dP}{dX} + \Delta \frac{\partial^2 U}{\partial Y^2}, \tag{3.46}$$

$$\frac{\partial}{\partial X} \left(U \begin{bmatrix} x_{O_2} \\ x_{H_2O} \end{bmatrix} \right) + \frac{\partial}{\partial Y} \left(V \begin{bmatrix} x_{O_2} \\ x_{H_2O} \end{bmatrix} \right) = \frac{\Delta}{Sc} \frac{\partial}{\partial Y} \left(\frac{\mathbf{M}}{\mathcal{M}} \frac{\partial}{\partial Y} \begin{bmatrix} x_{O_2} \\ x_{H_2O} \end{bmatrix} \right); \tag{3.47}$$

for $0 \leq X \leq 1$, $-\mathcal{H} \leq Y \leq 0$,

$$\langle \rho \rangle^{(\mathrm{g})} \langle V \rangle = \Omega \left\{ \frac{I}{4} \left(2(1+2\alpha) \mathcal{M}_{H_2O} - \mathcal{M}_{O_2} \right) \right\}, \tag{3.48}$$

$$\langle V \rangle = -\frac{1}{\Delta \Sigma^2} \frac{\partial \langle P \rangle^{(\mathrm{g})}}{\partial Y}, \tag{3.49}$$

$$\langle V \rangle \begin{bmatrix} \langle x_{O_2} \rangle^{(\mathrm{g})} \\ \langle x_{H_2O} \rangle^{(\mathrm{g})} \end{bmatrix} - \frac{\Delta \gamma^{\frac{3}{2}}}{Sc \langle \mathcal{M} \rangle^{(\mathrm{g})}} \langle \mathbf{M} \rangle^{(\mathrm{g})} \frac{\partial}{\partial Y} \begin{bmatrix} \langle x_{O_2} \rangle^{(\mathrm{g})} \\ \langle x_{H_2O} \rangle^{(\mathrm{g})} \end{bmatrix} = \frac{\Omega I}{4} \begin{bmatrix} -1 \\ 2(1+2\alpha) \end{bmatrix}. \tag{3.50}$$

Note here that (3.48), as well a consideration of the physical parameters, now helps to justify neglecting inertia terms between (2.12) and (2.22), as well as dispersion terms in (2.18) and (2.19). First, the Forchheimer correction term (see Whitaker [42]) will be of the order of magnitude of the Reynolds number, Re_γ, based on l_γ, loosely defined by

$$Re_\gamma = \frac{\langle \rho \rangle^{(\mathrm{g})} \langle v \rangle^{(\mathrm{g})} l_\gamma}{\mu}; \tag{3.51}$$

use of $\langle v\rangle^{(\mathrm{g})}$ for the velocity scale is justified since the foregoing analysis indicates that flow in the porous backing will be unidirectional. Consequently, using (2.37),

$$Re_\gamma \sim \frac{I}{4F\mu}\left(2(1+2\alpha)M_{H_2O} - M_{O_2}\right) l_\gamma \ll 1$$

as required. In addition, considerations based on this length scale provide some justification for neglecting dispersion effects in the porous backing, as compared to molecular diffusion. Experimental results for 1D flows (e.g., [2, pp. 606–609]) indicate that dispersion will be negligible if the Peclet number of molecular diffusion, Pe_γ, in the porous medium, defined here by

$$Pe_\gamma = \frac{\langle v\rangle^{(\mathrm{g})}\, l_\gamma}{[D]},$$

is much smaller than one; using the parameters given in section 3.2, this indeed turns out to be the case.

The boundary conditions are, for $0 \le X \le 1, Y = 1$,

$$U = V = \frac{\partial x_{O_2}}{\partial Y} = \frac{\partial x_{H_2O}}{\partial Y} = 0, \tag{3.52}$$

and for $X = 0,\ 0 \le Y \le 1$,

$$U = 1, \quad x_{O_2} = x_{O_2}^{in}, \quad x_{H_2O} = x_{H_2O}^{in} \quad \text{at} \quad X = 0, \quad 0 \le Y \le 1; \tag{3.53}$$

no boundary conditions as such prove to be necessary for $X = 0, -\mathcal{H} \le Y \le 0$ since only ordinary differential equations are solved for $-\mathcal{H} \le Y \le 0$. At $Y = 0$ for $0 \le X \le 1$, porous and fluid quantities are matched through

$$U = 0, \quad V = \langle V\rangle, \quad P = \langle P\rangle, \tag{3.54}$$

$$\langle x_{O_2}\rangle^{(\mathrm{g})} = x_{O_2}, \quad \langle x_{H_2O}\rangle^{(\mathrm{g})} = x_{H_2O}, \tag{3.55}$$

$$\gamma^{\frac{3}{2}} \frac{\partial}{\partial Y}\begin{bmatrix} \langle x_{O_2}\rangle^{(\mathrm{g})} \\ \langle x_{H_2O}\rangle^{(\mathrm{g})} \end{bmatrix} = \frac{\partial}{\partial Y}\begin{bmatrix} x_{O_2} \\ x_{H_2O} \end{bmatrix}. \tag{3.56}$$

In general, I will not be constant; even more generally, it cannot be described a priori but is determined by considering the transport of species in the catalyst, membrane, and the anode side also. However, a common practice in studies which emphasize the investigation of flow in the porous backing and the gas channel is simply to prescribe a current density as a function of mole fraction. For example, if we use the dimensional form of the Tafel law given by He, Yi, and Nguyen [22],

$$I = \frac{a\rho}{M}\exp\left(\frac{\alpha_c F\eta}{RT}\right),$$

where α_c $(= 2)$ is the transfer coefficient of the oxygen reduction reaction (1.2), η is the overpotential for the oxygen reaction, and a $\left(= 10^{-6}\ \mathrm{Am\ mol^{-1}}\right)$ is a constant

TABLE 3
Geometry and operating parameters for the base case.

$x_{O_2}^{in}$	0.21
$x_{H_2O}^{in}$	0
h_f	10^{-3}m
h_p	3×10^{-4}m
L	0.1 m
κ	10^{-12} m^2
γ	0.3
U^{in}	1 ms^{-1}
p^{out}	1 atm
T	353 K
μ	10^{-5}kgm^{-1}s^{-1}

related to the exchange current density and oxygen reference concentration for the oxygen reaction, we obtain the appropriate scale for $[I]$ as

$$[I] = \frac{a\,[\rho]}{[M]} \exp\left(\frac{\alpha_c F \eta}{RT}\right); \tag{3.57}$$

consequently, in dimensionless form,

$$I\left(\langle\rho\rangle^{(\mathrm{g})}, \langle x_{O_2}\rangle^{(\mathrm{g})}, \langle x_{H_2O}\rangle^{(\mathrm{g})}\right) = \frac{\langle\rho\rangle^{(\mathrm{g})}\,\langle x_{O_2}\rangle^{(\mathrm{g})}}{\langle\mathcal{M}\rangle^{(\mathrm{g})}} = \langle x_{O_2}\rangle^{(\mathrm{g})}. \tag{3.58}$$

A dimensional quantity of importance for the determination of polarization curves is the average current density, I_{av}, which is then given by

$$I_{av} = [I]\int_0^1 I dX.$$

This completes the formulation and necessary definitions. As a next step, we consider the possibility of finding an analytical solution in certain parameter ranges; an obvious choice, in view of the geometry, would be the lubrication theory limit $\left(\Delta^{-1} \ll 1\right)$. The data given in Table 3 for the base case physical parameters indicates that $\Delta^{-1} \sim O(1)$. Obviously, taking channels with a smaller aspect ratio or operating the fuel cell at lower inlet gas velocity would reduce Δ^{-1}, motivating us to then consider the lubrication theory limit, since it provides qualitatively useful analytical solutions, as well as a quantitative comparison with our numerical method (see section 6).

4. Asymptotics for $\frac{[\rho]\left(U^{in}\right)^2}{p^{out}} \ll \Delta^{-1} \ll 1$. Assume $\frac{[\rho]\left(U^{in}\right)^2}{p^{out}} \ll \Delta^{-1} \ll 1$, and rescale according to

$$\langle P\rangle^{(\mathrm{g})} = \Delta\,\langle P\rangle^{(\mathrm{g})},\ P = \Delta P;$$

note here that we require a lower restriction on Δ^{-1} for the following development to hold; otherwise the simplifications following (3.44) will not apply and ρ will depend

on P; in practice, the restriction is not unreasonable. Introducing the asymptotic series

$$\begin{aligned}
\chi &= \chi_0 + \Delta^{-1}\chi_1 + O\left(\Delta^{-2}\right), \quad \text{where } \chi = (U, V, P, \rho), \\
\langle\chi\rangle &= \langle\chi_0\rangle + \Delta^{-1}\langle\chi_1\rangle + O\left(\Delta^{-2}\right), \text{ where } \chi = (V, P), \\
\chi &= \chi^{(0)} + \Delta^{-1}\chi^{(1)} + O\left(\Delta^{-2}\right), \text{ where } \chi = (x_{O_2}, x_{H_2O}, \mathbf{M}), \\
\langle\chi\rangle^{(\mathrm{g})} &= \langle\chi_0\rangle^{(g)} + \Delta^{-1}\langle\chi_1\rangle^{(g)} + O\left(\Delta^{-2}\right), \text{ where } \chi = (\rho), \\
\langle\chi\rangle^{(\mathrm{g})} &= \left\langle\chi^{(0)}\right\rangle^{(\mathrm{g})} + \Delta^{-1}\left\langle\chi^{(1)}\right\rangle^{(\mathrm{g})} + O\left(\Delta^{-2}\right), \text{ where } \chi = (x_{O_2}, x_{H_2O}, \mathbf{M}),
\end{aligned}$$

we observe that, at leading order, the governing equations permit a solution of the form

$$\begin{aligned}
x_{O_2}^{(0)} &= \left\langle x_{O_2}^{(0)}\right\rangle^{(\mathrm{g})} = F_{O_2}(X), \\
x_{H_2O}^{(0)} &= \left\langle x_{H_2O}^{(0)}\right\rangle^{(\mathrm{g})} = F_{H_2O}(X),
\end{aligned}$$

with $F_{O_2}(0) = x_{O_2}^{in}$, $F_{H_2O}(0) = x_{H_2O}^{in}$. Then

$$U_0(X, Y) = \frac{1}{2}\frac{dP_0}{dX}\left(Y^2 - Y\right),$$

whereupon, writing $\Phi = (2(1+2\alpha)\mathcal{M}_{H_2O} - \mathcal{M}_{O_2})$ (note that $\Phi > 0$) and using

$$\frac{d}{dX}\left(\int_0^1 \rho_0 U_0 dY\right) = \frac{\Omega\Phi I\left(F_{O_2}(X)\right)}{4},$$

where

$$\rho_0(X) = \mathcal{M}_{N_2} + \left(\mathcal{M}_{O_2} - \mathcal{M}_{N_2}\right)F_{O_2}(X) + \left(\mathcal{M}_{H_2O} - \mathcal{M}_{N_2}\right)F_{H_2O}(X),$$

we have

$$\int_0^1 \rho_0 U_0 dY = \frac{\Omega\Phi}{4}J(X) + \rho_0(0),$$

where

$$J(X) = \int_0^X I\left(F_{O_2}(X')\right)dX'.$$

Hence

$$\frac{dP_0}{dX} = -\frac{12}{\rho_0(X)}\left[\frac{\Omega\Phi}{4}J(X) + \rho_0(0)\right],$$

and so

$$\begin{aligned}
U_0(X, Y) &= \frac{6}{\rho_0(X)}\left[\frac{\Omega\Phi}{4}J(X) + \rho_0(0)\right]\left(Y - Y^2\right), \\
V_0(X, Y) &= -\frac{3\Omega\Phi}{2}\frac{I\left(F_{O_2}(X)\right)}{\rho_0(X)}\left(\frac{Y^2}{2} - \frac{Y^3}{3} - \frac{1}{6}\right), \\
\langle V_0(X)\rangle &= \Omega\Phi\frac{I\left(F_{O_2}(X)\right)}{4\langle\rho_0(X)\rangle}, \\
\langle P_0(X, Y)\rangle &= -\Sigma^2\Omega\left\{\frac{\Phi I\left(F_{O_2}(X)\right)}{4\langle\rho_0(X)\rangle}\right\}Y - 12\left[\frac{\Omega\Phi J(X)}{\langle\rho_0(X)\rangle} + X\right].
\end{aligned}$$

At this stage, $F_{O_2}(X)$ and $F_{H_2O}(X)$ (and hence $U_0, V_0, \langle V_0\rangle, \langle P_0\rangle, \rho_0$) remain undetermined, indicating that the problem at $O(1)$ is degenerate; this appears to be because the boundary conditions for x_{O_2} and x_{H_2O} at $Y = -\mathcal{H}, 1$ at this order are both of Neumann type. This indeterminacy is remedied, however, at $O\left(\Delta^{-1}\right)$ as follows.

At $O\left(\Delta^{-1}\right)$, (3.47) gives

$$\frac{\partial}{\partial X}\left(U_0\begin{bmatrix} F_{O_2}(X) \\ F_{H_2O}(X)\end{bmatrix}\right)+\frac{\partial}{\partial Y}\left(V_0\begin{bmatrix} F_{O_2}(X) \\ F_{H_2O}(X)\end{bmatrix}\right) \tag{4.1}$$

$$=\frac{1}{Sc}\frac{\partial}{\partial Y}\left(\frac{\mathbf{M}^{(0)}}{\rho_0(X)}\frac{\partial}{\partial Y}\begin{bmatrix} x_{O_2}^{(1)} \\ x_{H_2O}^{(1)}\end{bmatrix}\right);$$

for $0 \le X \le 1$, $-\mathcal{H} \le Y \le 0$,

$$\frac{\Omega\Phi I(F_{O_2}(X))}{4\langle\rho_0(X)\rangle}\begin{bmatrix} F_{O_2}(X) \\ F_{H_2O}(X)\end{bmatrix}-\frac{\gamma^{\frac{3}{2}}}{Sc}\frac{\langle\mathbf{M}^{(0)}\rangle^{(\mathrm{g})}}{\langle\rho_0(X)\rangle}\frac{\partial}{\partial Y}\begin{bmatrix} \left\langle x_{O_2}^{(1)}\right\rangle^{(\mathrm{g})} \\ \left\langle x_{H_2O}^{(1)}\right\rangle^{(\mathrm{g})}\end{bmatrix} \tag{4.2}$$

$$=\frac{\Omega I(F_{O_2}(X))}{4}\begin{bmatrix} -1 \\ 2(1+2\alpha)\end{bmatrix}.$$

Equation (4.1) can be rewritten as

$$U_0\begin{bmatrix} F'_{O_2}(X) \\ F'_{H_2O}(X)\end{bmatrix}+\left(\frac{\partial U_0}{\partial X}+\frac{\partial V_0}{\partial Y}\right)\begin{bmatrix} F_{O_2}(X) \\ F_{H_2O}(X)\end{bmatrix}=\frac{1}{Sc}\frac{\mathbf{M}^{(0)}}{\rho_0(X)}\frac{\partial^2}{\partial Y^2}\begin{bmatrix} x_{O_2}^{(1)} \\ x_{H_2O}^{(1)}\end{bmatrix},$$

and then, on using

$$\frac{\partial}{\partial X}(\rho_0 U_0)+\frac{\partial}{\partial Y}(\rho_0 V_0)=0,$$

we have

$$U_0\left\{\begin{bmatrix} F'_{O_2}(X) \\ F'_{H_2O}(X)\end{bmatrix}-\frac{1}{\rho_0(X)}\frac{\partial\rho_0}{\partial X}\begin{bmatrix} F_{O_2}(X) \\ F_{H_2O}(X)\end{bmatrix}\right\}=\frac{1}{Sc}\frac{\mathbf{M}^{(0)}}{\rho_0(X)}\frac{\partial^2}{\partial Y^2}\begin{bmatrix} x_{O_2}^{(1)} \\ x_{H_2O}^{(1)}\end{bmatrix}.$$

Integrating once with respect to Y, we have

$$\lambda_0(X)\left\{\begin{bmatrix} F'_{O_2}(X) \\ F'_{H_2O}(X)\end{bmatrix}-\frac{1}{\rho_0(X)}\frac{\partial\rho_0}{\partial X}\begin{bmatrix} F_{O_2}(X) \\ F_{H_2O}(X)\end{bmatrix}\right\}\left(\frac{Y^2}{2}-\frac{Y^3}{3}-\frac{1}{6}\right) \tag{4.3}$$
$$=\frac{1}{Sc}\frac{\mathbf{M}^{(0)}}{\rho_0(X)}\frac{\partial}{\partial Y}\begin{bmatrix} x_{O_2}^{(1)} \\ x_{H_2O}^{(1)}\end{bmatrix},$$

where we have written $U_0 = \frac{\lambda_0(X)}{\rho_0(X)}\left(Y - Y^2\right)$, with

$$\lambda_0(X) = 6\left[\frac{\Omega\Phi}{4}J(X)+\rho_0(0)\right],$$

and have already implemented (3.52) at $O\left(\Delta^{-1}\right)$. Requiring now, at $Y=0$,

$$\gamma^{\frac{3}{2}}\frac{\partial}{\partial Y}\begin{bmatrix}\left\langle x_{O_2}^{(1)}\right\rangle^{(\mathrm{g})}\\ \left\langle x_{H_2O}^{(1)}\right\rangle^{(\mathrm{g})}\end{bmatrix}=\frac{\partial}{\partial Y}\begin{bmatrix}x_{O_2}^{(1)}\\ x_{H_2O}^{(1)}\end{bmatrix},$$

we combine (4.2) and (4.3) to give

$$-\frac{2\lambda_0(X)}{3\Omega\rho_0(X)I\left(F'_{O_2}(X)\right)}\left\{\begin{bmatrix}F'_{O_2}(X)\\ F'_{H_2O}(X)\end{bmatrix}-\frac{1}{\rho_0(X)}\frac{\partial\rho_0}{\partial X}\begin{bmatrix}F_{O_2}(X)\\ F_{H_2O}(X)\end{bmatrix}\right\}$$
$$=\frac{\Phi}{\rho_0(X)}\begin{bmatrix}F_{O_2}(X)\\ F_{H_2O}(X)\end{bmatrix}-\begin{bmatrix}-1\\ 2(1+2\alpha)\end{bmatrix}.$$

Note that this has led to the elimination of $x_{O_2}^{(1)}$ and $x_{H_2O}^{(1)}$ and has instead led to a pair of nonlinear ordinary differential equations for $F_{O_2}(X)$ and $F_{H_2O}(X)$.

Next, defining

$$\zeta_{O_2}(X)=F_{O_2}(X)/\rho_0(X),\ \zeta_{H_2O}(X)=F_{H_2O}(X)/\rho_0(X),$$

we simplify to

$$\frac{-4}{\Omega I\left(F_{O_2}(X)\right)}\left[\frac{\Omega\Phi}{4}J(X)+\rho_0(0)\right]\frac{\partial}{\partial X}\begin{bmatrix}\zeta_{O_2}(X)\\ \zeta_{H_2O}(X)\end{bmatrix}$$

$$=\Phi\begin{bmatrix}\zeta_{O_2}(X)\\ \zeta_{H_2O}(X)\end{bmatrix}-\begin{bmatrix}-1\\ 2(1+2\alpha)\end{bmatrix},$$

with initial conditions

$$\zeta_{O_2}(0)=\frac{x_{O_2}^{in}}{\mathcal{M}_{N_2}+\left(\mathcal{M}_{O_2}-\mathcal{M}_{N_2}\right)x_{O_2}^{in}+\left(\mathcal{M}_{H_2O}-\mathcal{M}_{N_2}\right)x_{H_2O}^{in}},$$
$$\zeta_{H_2O}(0)=\frac{x_{H_2O}^{in}}{\mathcal{M}_{N_2}+\left(\mathcal{M}_{O_2}-\mathcal{M}_{N_2}\right)x_{O_2}^{in}+\left(\mathcal{M}_{H_2O}-\mathcal{M}_{N_2}\right)x_{H_2O}^{in}}.$$

Replacing the partial derivative, we can simplify to

$$\frac{d\zeta_{O_2}}{d\zeta_{H_2O}}=\frac{\Phi\zeta_{O_2}+1}{\Phi\zeta_{H_2O}-2(1+2\alpha)},$$

whence, on applying the inlet conditions,

$$\left[\Phi\zeta_{O_2}(0)+1\right]\zeta_{H_2O}(X)-\left[\Phi\zeta_{H_2O}(0)-2(1+2\alpha)\right]\zeta_{O_2}(X)$$
$$=2(1+2\alpha)\zeta_{O_2}(0)+\zeta_{H_2O}(0). \tag{4.4}$$

Note, in addition, that this result holds regardless of the expression used for the current density.

Returning now to

$$\frac{-4}{\Omega I\left(F_{O_2}(X)\right)}\left[\frac{\Omega}{4}\Phi J(X)+\rho_0(0)\right]\frac{d\zeta_{O_2}}{dX}=\Phi\zeta_{O_2}+1,$$

this too can be integrated regardless of the form of I. We have, since $J(0) = 0$ for any current density we care to choose,

$$\zeta_{O_2}(X) = \frac{1}{\Phi}\left\{\left(\Phi\zeta_{O_2}(0)+1\right)\left(\frac{4\rho_0(0)}{4\rho_0(0)+\Omega\Phi J(X)}\right)-1\right\}, \tag{4.5}$$

which is effectively an integral equation for $\zeta_{O_2}(X)$. More convenient is a first order ordinary differential equation for $\zeta_{O_2}(X)$, which is obtained after rearranging and differentiating, as

$$I\left(\frac{\mathcal{M}_{N_2}\zeta_{O_2}(X)}{1-\left(\mathcal{M}_{O_2}-\mathcal{M}_{N_2}\right)\zeta_{O_2}(X)-\left(\mathcal{M}_{H_2O}-\mathcal{M}_{N_2}\right)\zeta_{H_2O}(X)}\right)$$

$$= -\frac{4\rho_0(0)}{\Omega}\frac{\left(\Phi\zeta_{O_2}(0)+1\right)}{\left(\Phi\zeta_{O_2}(X)+1\right)^2}\frac{d\zeta_{O_2}}{dX},$$

and then

$$I\left(\frac{\zeta_{O_2}(X)}{\mathcal{A}+\mathcal{B}\zeta_{O_2}(X)}\right) = -\frac{4\rho_0(0)}{\Omega}\frac{\left(\Phi\zeta_{O_2}(0)+1\right)}{\left(\Phi\zeta_{O_2}(X)+1\right)^2}\frac{d\zeta_{O_2}}{dX}, \tag{4.6}$$

where

$$\mathcal{A} = \frac{1}{\mathcal{M}_{N_2}}\left(\left(\mathcal{M}_{H_2O}-\mathcal{M}_{N_2}\right)\left(\frac{2(1+2\alpha)\zeta_{O_2}(0)+\zeta_{H_2O}(0)}{\left[\Phi\zeta_{O_2}(0)+1\right]}\right)-1\right),$$

$$\mathcal{B} = \frac{1}{\mathcal{M}_{N_2}}\left(\left(\mathcal{M}_{O_2}-\mathcal{M}_{N_2}\right)+\left(\mathcal{M}_{H_2O}-\mathcal{M}_{N_2}\right)\frac{\left[\Phi\zeta_{H_2O}(0)-2(1+2\alpha)\right]}{\left[\Phi\zeta_{O_2}(0)+1\right]}\right).$$

Note that (4.6) implies that if oxygen is fully depleted, at which point (say $X = X_0$) $\zeta_{O_2} = I = 0$, then we will necessarily have $\frac{d\zeta_{O_2}}{dX} = 0$ there also.

As an example, and for later use, we note a closed-form solution when $I \equiv \langle x_{O_2}\rangle^{(g)}$; in this case,

$$\frac{\zeta_{O_2}(X)}{\mathcal{A}+\mathcal{B}\zeta_{O_2}(X)} = -\frac{4\rho_0(0)}{\Omega}\frac{\left(\Phi\zeta_{O_2}(0)+1\right)}{\left(\Phi\zeta_{O_2}(X)+1\right)^2}\frac{d\zeta_{O_2}}{dX},$$

which can be integrated exactly to give

$$\mathcal{A}\log\left(\frac{\zeta_{O_2}(X)\left(\Phi\zeta_{O_2}(0)+1\right)}{\zeta_{O_2}(0)\left(\Phi\zeta_{O_2}(X)+1\right)}\right)-(\mathcal{A}\Phi-\mathcal{B})\left(\frac{\zeta_{O_2}(X)-\zeta_{O_2}(0)}{\left(\Phi\zeta_{O_2}(X)+1\right)\left(\Phi\zeta_{O_2}(0)+1\right)}\right)$$

$$= \frac{\Omega X}{4\rho_0(0)\left(\Phi\zeta_{O_2}(0)+1\right)}. \tag{4.7}$$

This formula suggests that for $\Omega \sim O(1)$ there is no possibility for oxygen depletion $\left(\zeta_{O_2} = 0\right)$, since the first term on the left-hand side of (4.7) could not then be balanced by either of the other two terms. In addition, for $\Omega \gg 1$ (and noting that $\mathcal{A} < 0$),

$$\mathcal{A}\log\left(\frac{\zeta_{O_2}(X)\left(\Phi\zeta_{O_2}(0)+1\right)}{\zeta_{O_2}(0)\left(\Phi\zeta_{O_2}(X)+1\right)}\right) \sim \frac{\Omega X}{4\rho_0(0)\left(\Phi\zeta_{O_2}(0)+1\right)},$$

whence

$$\frac{\zeta_{O_2}(X)}{\left(\Phi\zeta_{O_2}(X)+1\right)} \sim \frac{\zeta_{O_2}(0)}{\left(\Phi\zeta_{O_2}(0)+1\right)} \exp\left(\frac{\Omega X}{4\mathcal{A}\rho_0(0)\left(\Phi\zeta_{O_2}(0)+1\right)}\right),$$

and thus

$$\zeta_{O_2}(X) \sim \frac{1}{\left(\Phi+\frac{1}{\zeta_{O_2}(0)}\right)\exp\left(\frac{-\Omega X}{4\mathcal{A}\rho_0(0)\left(\Phi\zeta_{O_2}(0)+1\right)}\right)-\Phi};$$

in this regime, we also have

$$\zeta_{H_2O}(X) \sim \frac{2}{\Phi}(1+2\alpha) + \frac{\zeta_{H_2O}(0)-\frac{2}{\Phi}(1+2\alpha)}{\left(\Phi\zeta_{O_2}(0)+1\right)-\Phi\zeta_{O_2}(0)\exp\left(\frac{\Omega X}{4\mathcal{A}\rho_0(0)\left(\Phi\zeta_{O_2}(0)+1\right)}\right)}.$$

Note also that, thus far, the results are independent of whether or not crossed diffusion is assumed or if nonlinear diffusion coefficients are used or not.

As a corollary, we observe that the solution at $O\left(\Delta^{-1}\right)$, i.e., for $x_{O_2}^{(1)}$, $x_{H_2O}^{(1)}$, $\langle x_{O_2}^{(1)}\rangle^{(\mathrm{g})}$, $\langle x_{H_2O}^{(1)}\rangle^{(\mathrm{g})}$ still remains undetermined, since consideration of the field equations and boundary conditions at $O\left(\Delta^{-1}\right)$ merely leads to the solution being fully determined at $O(1)$. By analogy, to determine the solutions at $O\left(\Delta^{-1}\right)$ completely, we would need to consider the field equations and boundary conditions at $O\left(\Delta^{-2}\right)$; by this stage, however, the algebra becomes lengthy and in the interest of brevity we omit further discussion. In fact, one does not really gain so much by finding these solutions anyway, since there is no compact solution as there is at $O(1)$, and we proceed instead to a numerical solution for the general case when $\Delta^{-1} \sim O(1)$.

5. Numerical method and results. To complement the asymptotics for $\Delta^{-1} \ll 1$, so as to account for regimes when $\Delta^{-1} \sim O(1)$, the simplified parabolized equations were solved numerically using the Keller–Box discretization scheme and Newton iteration (see, for example, Cebeci and Bradshaw [10]). The system of partial differential equations to be solved in the channel is of 8th order, and this is coupled to a 6th order system of ordinary differential equations in the porous region. As is well known, the scheme is second order accurate in both time-like and space-like variables, and we omit any further details here. As an indication of the speed of the computations, we note that a typical run with 500 points across, and 200 points along, the channel required around 100 CPU seconds on a 500 MHz Compaq Alphaserver with 3GB RAM.

Results are presented for the Tafel law given in dimensionless form by (3.58) and used previously for PEFC studies by [22, 25, 30, 45]. Throughout, we keep $\gamma = 0.3, T = 353$ K, and concentrate more on the effect of changes in channel height and length, porous backing thickness and permeability, pressure, inlet speed, and composition. Physically realistic and implementable changes in any of these will result in, at most, an order of magnitude change in the relevant dimensionless parameter. The most sensitive parameter is Ω, which varies over several orders of magnitude as the cell voltage E_{cell} decreases; note here that we revert to using the cell voltage rather than the overpotential, η, with the two being related by

$$E_{cell} = E_0 - \eta,$$

where $E_0 (= 1.1$ V) is termed the open circuit voltage of the fuel cell.

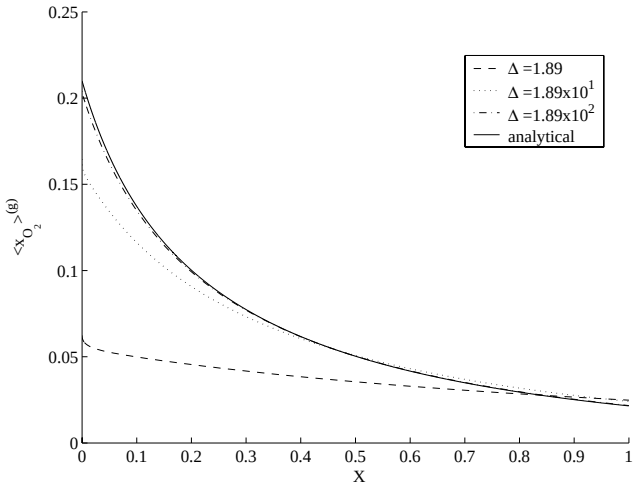


FIG. 3. *Comparison of analytical solution for* $\langle x_{O_2} \rangle^{(g)}$ *at* $Y = -\mathcal{H}$ *with numerical solutions for* $\Delta = 1.89, 1.89 \times 10^1, 1.89 \times 10^2$ ($E_{cell} = 0.75$ V).

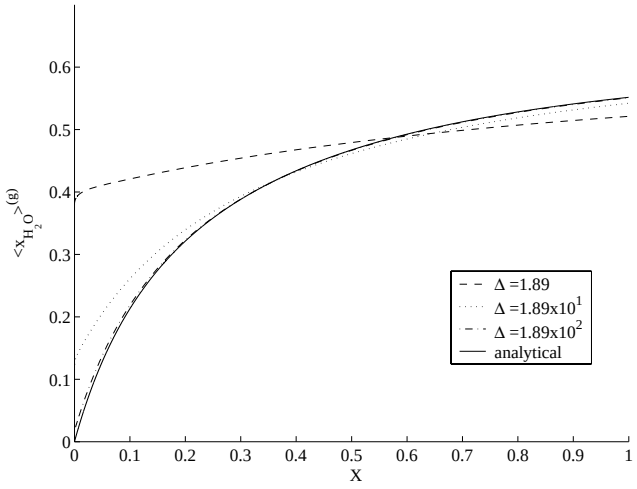


FIG. 4. *Comparison of analytical solution for* $\langle x_{H_2O} \rangle^{(g)}$ *at* $Y = -\mathcal{H}$ *with numerical solutions for* $\Delta = 1.89, 1.89 \times 10^1, 1.89 \times 10^2$ ($E_{cell} = 0.75$ V).

5.1. Effect of Δ and Ω. We show first results for $E_{cell} = 0.75$ V, corresponding to $\Omega = 10.2$, ranging over several orders of magnitude in Δ and compare these with the analytical results in the lubrication theory limit. Figures 3 and 4 are for intrinsic oxygen and water mole fraction at $Y = -\mathcal{H}$, respectively, and demonstrate that the lubrication solution works well for Δ^{-1} as high as $O\left(10^{-2}\right)$.

On the other hand, the base case physical values given in Table 3 correspond to $\Delta = 1.89$. Figure 5 shows the streamwise velocity U at $Y = \frac{1}{2}$ and illustrates the extent of deviation from the classical value $\frac{3}{2}$. An interesting limit occurs as E_{cell} is decreased. In this case, Ω increases although the quantity $\Omega \langle x_{O_2} \rangle^{(g)}$ at $Y = -\mathcal{H}$ remains $O(1)$; this corresponds to the attainment of the limiting current, and the

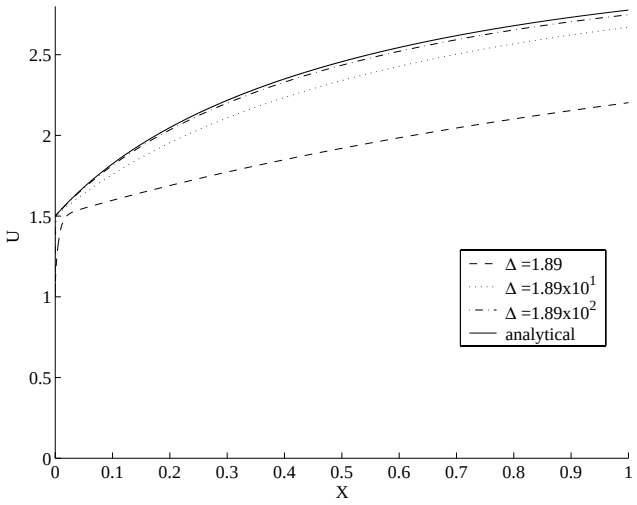


FIG. 5. *Comparison of analytical solution for* U *at* $Y = \frac{1}{2}$ *with numerical solutions for* $\Delta = 1.89, 1.89 \times 10^1, 1.89 \times 10^2$ ($E_{cell} = 0.75$ V).

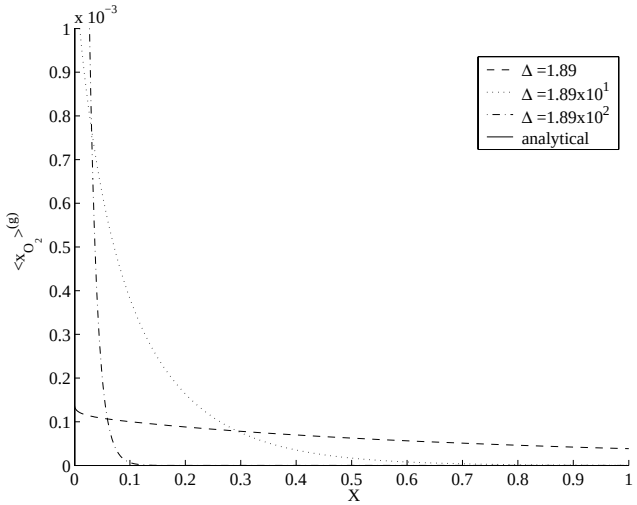


FIG. 6. *Comparison of analytical solution for* $\langle x_{O_2} \rangle^{(g)}$ *at* $Y = -\mathcal{H}$ *with numerical solutions for* $\Delta = 1.89, 1.89 \times 10^1, 1.89 \times 10^2$ ($E_{cell} = 0.65$ V).

corresponding plots are given in Figures 6, 7, and 8; observe that in Figures 6 and 7 the limiting values for intrinsic oxygen and water mole fraction for the analytical solution are reached very rapidly, so that in Figure 6 the curve for $\langle x_{O_2} \rangle^{(g)}$ effectively lies on the X-axis.

With regard to the numerics, it was found that considerably more outer loop iterations for the density were required as Ω was increased. For instance, whereas 4 iterations sufficed for $E_{cell} = 0.75$ V, it was common for 20–30 to be necessary for $E_{cell} = 0.65$ V. In addition, there were difficulties in initiating the marching scheme at $X = 0$ for higher values of Ω; we surmise this to be due to the increased nonlinearity of the equation system. Whilst setting the channel inlet values as an initial guess for

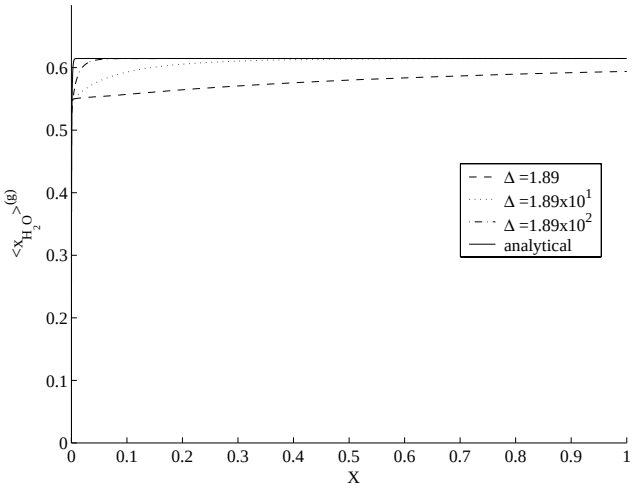


FIG. 7. *Comparison of analytical solution for* $\langle x_{H_2O} \rangle^{(g)}$ *at* $Y = -\mathcal{H}$ *with numerical solutions for* $\Delta = 1.89, 1.89 \times 10^1, 1.89 \times 10^2$ $(E_{cell} = 0.65 \text{ V})$.

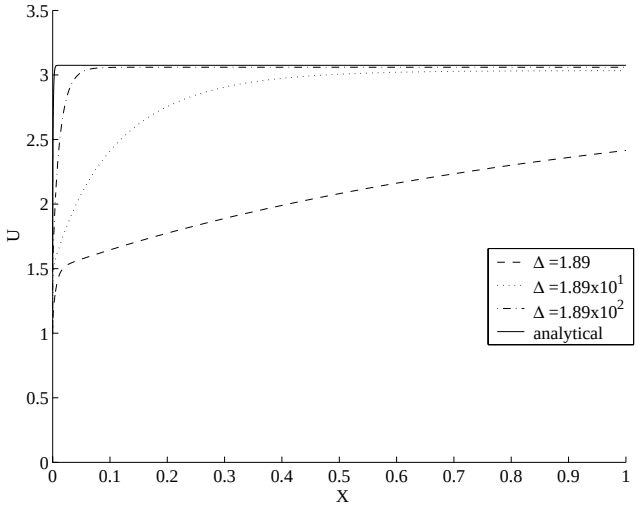


FIG. 8. *Comparison of analytical solution for* U *at* $Y = \frac{1}{2}$ *with numerical solutions for* $\Delta = 1.89, 1.89 \times 10^1, 1.89 \times 10^2$ $(E_{cell} = 0.65 \text{ V})$.

the first step along the channel was adequate for lower values of Ω, this was found to be not sufficient for E_{cell} lower than 0.71 V; for those cases, the first-step solution for $E_{cell} = 0.71$ V had to be used instead, then enabling numerical solutions to be obtained for higher and higher values of Ω until the limiting current was reached.

5.2. "Polarization surfaces". It is customary for fuel cell performance to be given in terms of a polarization curve where the cell potential, E_{cell}, is given as function of the average current density, I_{av}. Generally speaking, if the analysis is done dimensionally, this leads to a vast number of graphs for each alteration made in one of the physical parameters. However, a major benefit of the nondimensional analysis

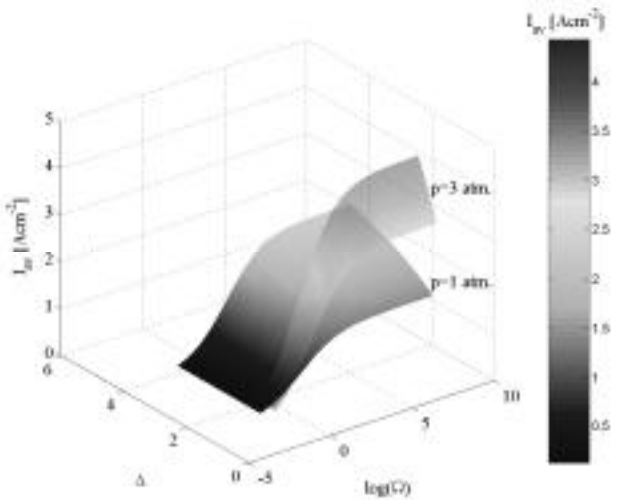


FIG. 9. *Polarization surfaces for* $p^{out} = 1, 3$ *atm* ($\mathcal{H} = 0.3$).

carried out here is that the results can be expressed considerably more compactly by plotting polarization "surfaces"; individual polarization curves will therefore be curves lying on those surfaces. We explain this as follows. From the nondimensionalization given above, the emergent nondimensional parameters were Δ, Σ, Ω, and Sc. In addition, there is γ, which we hold fixed in this study, and $x^{in}_{O_2}$,$x^{in}_{H_2O}$, and $\mathcal{H}$, whose effect on fuel cell performance one would like to explore. First, we observe that, in the parameter range of interest, Σ has no effect on I_{av}, since the dimensionless density is independent of pressure and the pressure in the channel serves as a boundary condition for the pressure in the porous medium. In addition, a change in Sc can be effected only by changes in $[\rho]$, which occurs only if the cathode is run at a different pressure. Consequently, a tidy representation of I_{av} is to plot it as a function of Δ and Ω, for fixed $Sc, x^{in}_{O_2}$,$x^{in}_{O_2}$, and $\mathcal{H}$, the benefit of this being that the effect of four parameters, h_f, L, U^{in}, and E_{cell}, are displayed on one graph; since Ω can vary over several orders of magnitude, it proves more convenient to use $\log(\Omega)$ as a variable. Examples of this are given below.

Figure 9 gives polarization surfaces for $\mathcal{H} = 0.3$, with the pressure at 1 and 3 atmospheres. The limiting current phenomenon is observed as Ω increases, and its value is observed to increase moderately with increasing Δ but strongly with increasing pressure. Figure 10 shows a similar plot, except with computations now for $p^{out} = 1$ atm, for $\mathcal{H} = 0.15$ and $\mathcal{H} = 0.6$. Average current densities are found to be higher for the thinner porous backing, and in both cases a limiting value is evident as Ω is increased. Figure 11 compares the base case for $p^{out} = 1$ atm and $x^{in}_{O_2} = 0.21$ with two other cases at 1 atm which have differing inlet compositions: dry oxygen $\left(x^{in}_{O_2} = 1\right)$ and partially humidified air, for which $x^{in}_{O_2} = 0.13$ and $x^{in}_{H_2O} = 0.36$ (corresponding to 76% relative humidity) [14, 17, 35]. As is evident, increased oxygen content at the inlet raises the average current density; for $x^{in}_{O_2} = 1$, convergence difficulties were experienced for quite low values of Ω, which explains the rather narrow range of values presented for this case, but nonetheless the average current density is much higher than that for the other two cases.

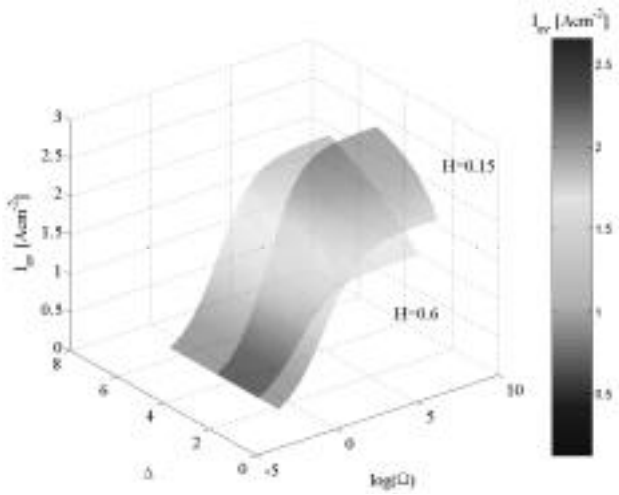


FIG. 10. *Polarization surfaces for* $\mathcal{H} = 0.3, 0.6$ $\left(p^{out} = 1\ atm\right)$.

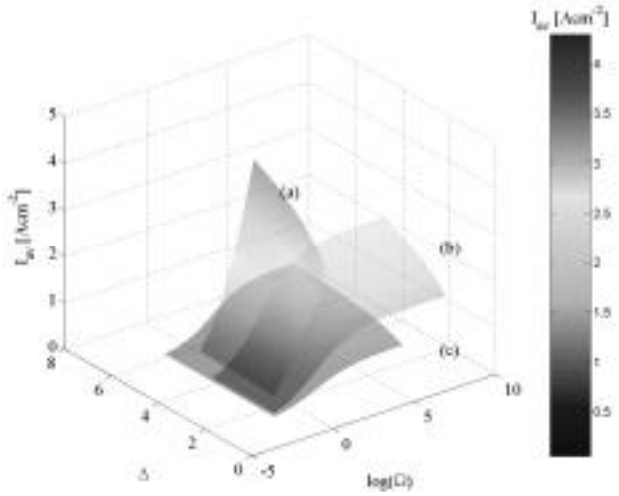


FIG. 11. *Polarization surfaces for* $p^{out} = 1$ *atm with* (a) $x^{in}_{O_2} = 1$, $x^{in}_{H_2O} = 0$; (b) $x^{in}_{O_2} = 0.21$, $x^{in}_{H_2O} = 0$; (c) $x^{in}_{O_2} = 0.13$, $x^{in}_{H_2O} = 0.36$.

6. Conclusions. In this paper, we have considered a 2D model for three-component gaseous flow in the cathode of a PEFC. Assuming a slender geometry, we have derived analytical solutions where possible and complemented these with a numerical study. By choosing to perform the study nondimensionally, we have identified several features that are not evident from earlier work done dimensionally. In summary, we identify four main dimensionless parameters ($\Delta, \Omega, \Sigma, Sc$; see sections 2 and 3 for definitions); other parameters that are present in this model are the porous backing porosity, γ (held fixed at 0.3 in this study), the temperature T (held fixed at 353 K), the ratio of channel and porous backing heights ($\mathcal{H}$), the inlet oxygen and water content ($x^{in}_{O_2}$ and $x^{in}_{H_2O}$, respectively), and the number of water molecules affiliated to each proton that passes across the membrane to the cathode (α). We find that

the flow in the porous backing is essentially unidirectional, although it interacts with a fully 2D flow in the gas channel. Furthermore, Σ is found to play a secondary role, having next to no effect on the gas mole fraction distribution in the porous backing; physically, this implies an insensitivity to porous backing permeability. The fact that the cathode is more or less isobaric gives that the density is a multiple of the molecular weight. In addition, the Schmidt number, Sc, can be affected only by variations in the operating pressure, and we find that a convenient and compact way to understand fuel cell performance is to plot average current density as a function of Δ and Ω for different values of Sc and $\mathcal{H}$; this gives a surface which implicitly contains an infinite family of polarization curves, which is the customary way to assess cell performance. These surfaces have been generated numerically, and in comparatively rapid fashion, using the Keller–Box scheme for systems of parabolic partial differential equations, to provide a rather comprehensive parameter study.

The present work was, needless to say, limited in several respects. To begin with, as has often been stated before, at higher current densities two-phase flow can be expected as water droplets form at the catalytic layer; this will be the starting point for future work. Also, we limited ourselves here to prescribing an often-used Tafel law for the current density relation at the catalytic layer, for one temperature value and one value of porosity; in addition, α was assumed to be constant along the length of the cell. Naturally, the question arises as to whether more sophisticated modeling would lead to a qualitative change in the results. Essentially, such an approach would involve attempting to represent the catalytic layer more faithfully, e.g., as has been attempted for MCFC [34]. A characteristic of this approach would be that a Tafel law is used for reactions across this layer, which would be treated as consisting of the material of which the gas-diffusion electrode as well as polymer electrolyte. Combined with the fact that this layer is much thinner than the gas-diffusion electrode, it is likely that the functional form for the current density would not be much different from what we have used here. On the other hand, if the overpotential and/or temperature are no longer treated as constant, then the exponential term involving these quantities in the expression for the current density could indeed affect the results significantly. Needless to say, this also is the subject of future work.

A final comment concerns the use of nondimensional analysis in fuel cell modeling. The way that the analysis presented here panned out was a function of the magnitudes of the relevant parameters for the PEFC. Subsequent work carried out by us on the anode of a direct methanol fuel cell [9] indicates that the dimensionless parameter Ω there will actually be much smaller than unity, leading to a different treatment of the subsequent equations. This emphasizes the point that the treatment presented here is more far-reaching than just the PEFC.

REFERENCES

[1] B. Alazmi and K. Vafai, *Analysis of fluid flow and heat transfer interfacial conditions between a porous layer and a fluid layer,* Internat. J. Heat Mass Transfer, 44 (2001), pp. 1735–1749.
[2] J. Bear, *Dynamics of Fluids in Porous Media,* Dover, New York, 1988.
[3] J. Bear and J. M. Buchlin, *Modeling and Applications of Transport Phenomena in Porous Media,* Kluwer Academic Publishers, Boston, 1991.
[4] G. S. Beavers and D. D. Joseph, *Boundary conditions at a naturally permeable wall,* J. Fluid Mech., 30 (1967), pp. 197–207.
[5] D. M. Bernardi and M. W. Verbrugge, *Mathematical model of a gas diffusion electrode bonded to a polymer electrolyte*, AIChE J., 37 (1991), pp. 1151–1163.
[6] D. M. Bernardi and M. W. Verbrugge, *A mathematical model of the solid-polymer-*

electrolyte fuel cell, J. Electrochem. Soc., 139 (1992), pp. 2477–2491.
[7] J. Billingham, A. C. King, R. C. Copcutt, and K. Kendall, *Analysis of a model for a loaded, planar, solid oxide fuel cell*, SIAM J. Appl. Math., 60 (2000), pp. 574–601.
[8] R. B. Bird, W. E. Stewart, and E. N. Lightfoot, *Transport Phenomena*, John Wiley, New York, 1960.
[9] E. Birgersson, J. Nordlund, H. Ekström, M. Vynnycky, and G. Lindbergh, *A reduced two-dimensional one-phase model for analysis of the anode of a DMFC,* J. Electrochem. Soc., submitted.
[10] T. Cebeci and P. Bradshaw, *Momentum Transfer in Boundary Layers,* Hemisphere Publishing, Washington, DC, 1977.
[11] R. J. Cooper, J. Billingham, and A. C. King, *Flow and reaction in solid oxide fuel cells,* J. Fluid Mech., 411 (2000), pp. 233–262.
[12] K. Dannenberg, P. Ekdunge, and G. Lindbergh, *Mathematical model of the PEMFC,* J. Appl. Electrochemistry, 30 (2000), pp. 1377–1387.
[13] P. De Vidts and R. E. White, *Governing equations for transport in porous electrodes*, J. Electrochem. Soc., 144 (1997), pp. 1343–1353.
[14] S. Dutta, S. Shimpalee, and J. W. van Zee, *Three-dimensional numerical simulation of straight channel PEM cells*, J. Appl. Electrochemistry, 30 (2000), pp. 135–146.
[15] S. Dutta, S. Shimpalee, and J. W. van Zee, *Numerical prediction of mass exchange between cathode and anode channels in a PEM fuel cell,* Internat. J. Heat Mass Transfer, 44 (2001), pp. 2029–2042.
[16] A. Fischer, J. Jindra, and H. Wendt, *Porosity and catalyst utilization of thin layer cathodes in air operated PEM-fuel cells,* J. Appl. Electrochemistry, 28 (1998), pp. 277–282.
[17] T. F. Fuller and J. Newman, *Water and thermal management in solid-polymer-electrolyte fuel cells*, J. Electrochem. Soc., 140 (1993), pp. 1218–1225.
[18] P. Futerko and I-M. Hsing, *Two-dimensional finite-element study of the resistance of membranes in polymer electrolyte fuel cells*, Electrochimica Acta, 45 (2000), pp. 1741–1751.
[19] V. Gurau, H. Liu, and S. Kakac, *Two-dimensional model for proton exchange membrane fuel cells*, AIChE J., 44 (1998), pp. 2410–2422.
[20] V. Gurau, F. Barbir, and H. Liu, *An analytical solution of a half-cell model for PEM fuel cells,* J. Electrochem. Soc., 147 (2000), pp. 2468–2477.
[21] W. He and Q. Chen, *Three-dimensional simulation of a molten carbonate fuel cell stack using computational fluid dynamics technique*, J. Power Sources, 55 (1995), pp. 25–32.
[22] W. He, J. S. Yi, and T. V. Nguyen, *Two-phase flow model of the cathode of PEM fuel cells using interdigitated flow fields,* AIChE J., 46 (2000), pp. 2053–2064.
[23] I-M. Hsing and P. Futerko, *Two-dimensional simulation of water transport in polymer electrolyte fuel cells*, Chemical Engrg. Sci., 55 (2000), pp. 4209–4218.
[24] W. Jäger and A. Mikelić, *On the interface boundary condition of Beavers, Joseph, and Saffman,* SIAM J. Appl. Math., 60 (2000), pp. 1111–1127.
[25] A. Kazim, H. T. Liu, and P. Forges, *Modelling of performance of PEM fuel cells with conventional and interdigitated flow fields*, J. Appl. Electrochemistry, 29 (1999), pp. 1409–1416.
[26] A. C. King, J. Billingham, and R. J. Cooper, *Performance modelling of solid oxide fuel cells,* Combust. Theory Model., 5 (2001), pp. 639–667.
[27] J. H. Lee, T. R. Lalk, and A. J. Appleby, *Modeling electrochemical performance in large scale proton exchange membrane fuel cell stacks,* J. Power Sources, 70 (1998), pp. 258–268.
[28] J. H. Lee and T. R. Lalk, *Modeling fuel cell stack systems,* J. Power Sources, 73 (1998), pp. 229–241.
[29] G. Maggio, V. Recupero, and C. Mantegaza, *Modelling of temperature distribution in a solid polymer electrolyte fuel cell stack,* J. Power Sources, 62 (1996), pp. 167–174.
[30] T. V. Nguyen, *Modeling two-phase flow in the porous electrodes of proton exchange membrane fuel cells using the interdigitated flow fields,* in Proceedings Volume on Tutorials in Electrochemical Engineering—Mathematical Modeling, 99-14, The Elctrochemical Society, Pennington, NJ, 1999, pp. 222–241.
[31] T. V. Nguyen and R. E. White, *A water and heat management model for proton-exchange-membrane fuel cells*, J. Electrochem. Soc., 140 (1993), pp. 2178–2186.
[32] J. A. Ochoa-Tapia and S. Whitaker, *Momentum jump condition at the boundary between a porous medium and a homogeneous fluid: Inertial effects*, J. Porous Media, 1 (1998), pp. 201–217.
[33] J. A. Ochoa-Tapia and S. Whitaker, *Heat transfer at the boundary between a porous medium and a homogeneous fluid: The one-equation model*, J. Porous Media, 1 (1998), pp. 31–46.
[34] J. A. Prins-Jansen, J. D. Fehribach, K. Hemmes, and J. H. W. DeWit, *A three-phase*

homogeneous model for porous electrodes in molten-carbonate fuel cells, J. Electrochem. Soc., 143 (1996), pp. 1617–1628.

[35] S. Shimpalee and S. Dutta, *Numerical prediction of temperature distribution in PEM fuel cells,* Numerical Heat Transfer, Part A, 38 (2000), pp. 111–128.

[36] D. Singh, D. M. Lu, and N. Djilali, *A two-dimensional analysis of mass transport in proton exchange membrane fuel cells*, Internat. J. Engrg. Sci., 37 (1999), pp. 431–452.

[37] T. E. Springer, T. A. Zawodzinski, and S. Gottesfeld, *Polymer electrolyte fuel cell model*, J. Electrochem. Soc., 138 (1991), pp. 2334–2342.

[38] R. Taylor and R. Krishna, *Multicomponent Mass Transfer,* John Wiley, New York, 1993.

[39] D. Thirumalai and R. E. White, *Mathematical modeling of proton-exchange-membrane fuel-cell stacks,* J. Electrochem. Soc., 144 (1997), pp. 1717–1723.

[40] M. Uchida, Y. Fukuoka, Y. Sugawara, N. Eda, and A. Ohta, *Effects of mictrostructure of carbon support in the catalyst layer on the performance of polymer-electrolyte fuel cells,* J. Electrochem. Soc., 143 (1996), pp. 2245–2252.

[41] Z. H. Wang, C. Y. Wang, and K. S. Chen, *Two-phase flow and transport in the air cathode of proton exchange membrane fuel cells,* J. Power Sources, 94 (2001), pp. 40–50.

[42] S. Whitaker, *The Forchheimer equation: A theoretical development,* Transp. Porous Media, 25 (1996), pp. 27–61.

[43] S. Whitaker, *The Method of Volume Averaging,* Kluwer Academic Publishers, Dordrecht, The Netherlands, 1999.

[44] J. S. Yi and T. V. Nguyen, *An along-the-channel model for proton exchange membrane fuel cells,* J. Electrochem. Soc., 145 (1998), pp. 1149–1159.

[45] J. S. Yi and T. V. Nguyen, *Multicomponent transport in porous electrodes of proton exchange membrane fuel cells using the interdigitated gas distributors*, J. Electrochem. Soc., 146 (1999), pp. 38–45.

SIAM J. APPL. MATH.
Vol. 63, No. 4, pp. 1424–1441

LARGE SOLUTIONS TO THE INITIAL-BOUNDARY VALUE PROBLEM FOR PLANAR MAGNETOHYDRODYNAMICS*

DEHUA WANG†

Abstract. An initial-boundary value problem for nonlinear magnetohydrodynamics (MHD) in one space dimension with general large initial data is investigated. The equations of state have nonlinear dependence on temperature as well as on density. For technical reasons the viscosity coefficients and magnetic diffusivity are assumed to depend only on density. The heat conductivity is a function of both density and temperature, with a certain growth rate on temperature. The existence, uniqueness, and regularity of global solutions are established with large initial data in H^1. It is shown that no shock wave, vacuum, or mass or heat concentration will be developed in a finite time, although the motion of the flow has large oscillations and there is a complex interaction between the hydrodynamic and magnetodynamic effects.

Key words. magnetohydrodynamics (MHD), global solutions, global a priori estimates

AMS subject classifications. 35L65, 35B45, 35B40, 76N10, 76W05, 76X05

PII. S0036139902409284

1. Introduction.

In this paper we are concerned with the initial-boundary value problem and large-time behavior of solutions for plane magnetohydrodynamic flows. Magnetohydrodynamics (MHD) concerns the motion of conducting fluids in an electromagnetic field with a very broad range of applications. The dynamic motion of the fluids and the magnetic field interact strongly with each other. The hydrodynamic and electrodynamic effects are coupled. The equations of three-dimensional magnetohydrodynamic flows have the following form [4, 16, 18]:

$$
\begin{aligned}
&\rho_t + \operatorname{div}(\rho\mathbf{u}) = 0, \qquad (1.1)\\
&(\rho\mathbf{u})_t + \operatorname{div}(\rho\mathbf{u}\otimes\mathbf{u}) + \nabla p = (\nabla\times\mathbf{H})\times\mathbf{H} + \operatorname{div}\mathbf{\Psi},\\
&\mathcal{E}_t + \operatorname{div}(\mathbf{u}(\mathcal{E}+p)) = \operatorname{div}\left((\mathbf{u}\times\mathbf{H})\times\mathbf{H} + \nu\mathbf{H}\times(\nabla\times\mathbf{H}) + \mathbf{u}\mathbf{\Psi} + \kappa\nabla\theta\right),\\
&\mathbf{H}_t - \nabla\times(\mathbf{u}\times\mathbf{H}) = -\nabla\times(\nu\nabla\times\mathbf{H}), \qquad \operatorname{div}\mathbf{H} = 0,
\end{aligned}
$$

where $\mathbf{\Psi} = \lambda'(\operatorname{div}\mathbf{u})\mathbf{I} + \mu(\nabla\mathbf{u} + (\nabla\mathbf{u})^\top)$; ρ denotes the density, $\mathbf{u}\in\mathbb{R}^3$ the velocity, $\mathbf{H}\in\mathbb{R}^3$ the magnetic field, and θ the temperature; $\mathcal{E}$ is the total energy given by

$$
\mathcal{E} = \rho\left(e + \frac{1}{2}|\mathbf{u}|^2\right) + \frac{1}{2}|\mathbf{H}|^2, \qquad (1.2)
$$

with e the internal energy, $\frac{1}{2}|\mathbf{u}|^2$ the kinetic energy, and $\frac{1}{2}|\mathbf{H}|^2$ the magnetic energy; the equations of state $p = p(\rho,\theta)$, $e = e(\rho,\theta)$ relate the pressure p and the internal energy e to the density and temperature of the flow; $\mathbf{I}$ is the 3×3 identity matrix, and $(\nabla\mathbf{u})^\top$ is the transpose of the matrix $\nabla\mathbf{u}$; $\lambda' = \lambda'(\rho,\theta)$ and $\mu = \mu(\rho,\theta)$ are the viscosity coefficients of the flow satisfying $\lambda' + 2\mu > 0$, $\nu = \nu(\rho,\theta)$ is the magnetic diffusivity (see [1]) acting as a magnetic diffusion coefficient of the magnetic field, $\kappa = \kappa(\rho,\theta)$ is the heat conductivity, and all these kinetic coefficients and the magnetic

*Received by the editors June 5, 2002; accepted for publication (in revised form) October 14, 2002; published electronically May 22, 2003. The research was supported in part by the National Science Foundation and the Office of Naval Research.
http://www.siam.org/journals/siap/63-4/40928.html
†Department of Mathematics, University of Pittsburgh, Pittsburgh, PA 15260 (dwang@math.pitt.edu).

diffusivity are independent of the magnitude and direction of the magnetic field (see [18]). The magnetic permeability differs only slightly from unity and therefore is taken to be 1, which does not appear in the equations. The viscosity and heat conduction terms describe the dissipative processes in MHD.

It is well known that the electromagnetic fields are governed by Maxwell's equations. In MHD, the displacement current can be neglected [16, 18]. As a consequence, the last equation in (1.1) is called the induction equation, and the electric field can be written in terms of the magnetic field $\mathbf{H}$ and the velocity $\mathbf{u}$:

$$\mathbf{E} = \nu \nabla \times \mathbf{H} - \mathbf{u} \times \mathbf{H}. \tag{1.3}$$

Although the electric field $\mathbf{E}$ does not appear in (1.1), it is indeed induced according to (1.3) by the moving conductive flow in the magnetic field.

Consider a three-dimensional MHD flow with spatial variables $\mathbf{x} = (x, x_2, x_3)$, which is moving in the x direction and uniform in the transverse direction (x_2, x_3):

$$\begin{aligned} &\rho = \rho(x,t), \qquad \theta = \theta(x,t), \\ &\mathbf{u} = (u, \mathbf{w})(x,t), \quad \mathbf{w} = (u_2, u_3), \\ &\mathbf{H} = (b_1, \mathbf{b})(x,t), \quad \mathbf{b} = (b_2, b_3), \end{aligned} \tag{1.4}$$

where u and b_1 are the longitudinal velocity and longitudinal magnetic field, respectively, and $\mathbf{w}$ and $\mathbf{b}$ are the transverse velocity and transverse magnetic field, respectively. See Figure 1(a) and (b).

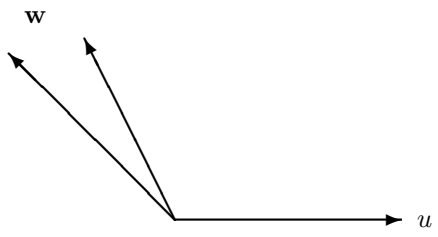


FIG. 1(a): *velocity* $(u, \mathbf{w})$.

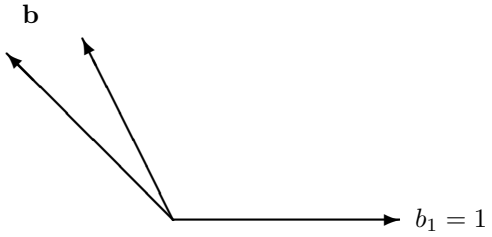


FIG. 1(b): *magnetic field* $(b_1, \mathbf{b})$.

With this special structure (1.4), equations (1.1) are reduced to the following system for the plane magnetohydrodynamic flows with constant longitudinal magnetic field $b_1 = 1$ (without loss of generality) and $\lambda = \lambda' + 2\mu > 0$:

$$\begin{aligned} \rho_t + (\rho u)_x &= 0, \\ (\rho u)_t + \left(\rho u^2 + P\right)_x &= (\lambda u_x)_x, \\ (\rho \mathbf{w})_t + (\rho u \mathbf{w} - \mathbf{b})_x &= (\mu \mathbf{w}_x)_x, \\ \mathbf{b}_t + (u\mathbf{b} - \mathbf{w})_x &= (\nu \mathbf{b}_x)_x, \\ \mathcal{E}_t + \left(u\left(\mathcal{E} + P\right) - \mathbf{w} \cdot \mathbf{b}\right)_x &= (\lambda u u_x + \mu \mathbf{w} \cdot \mathbf{w}_x + \nu \mathbf{b} \cdot \mathbf{b}_x + \kappa \theta_x)_x, \end{aligned} \tag{1.5}$$

where, again as in (1.1), $x \in \mathbb{R}$ is the spatial variable, $t > 0$ is the time variable; ρ denotes the density of the flow, $u \in \mathbb{R}$ the longitudinal velocity, $\mathbf{w} \in \mathbb{R}^2$ the transverse velocity, $\mathbf{b} \in \mathbb{R}^2$ the transverse magnetic field, and θ the temperature; the total energy of the plane magnetohydrodynamic flow is

$$\mathcal{E} = \rho\left(e + \frac{1}{2}(u^2 + |\mathbf{w}|^2)\right) + \frac{1}{2}|\mathbf{b}|^2,$$

with the internal energy e, and

$$P = p + \frac{1}{2}|\mathbf{b}|^2 \tag{1.6}$$

is the full pressure with p the pressure of fluid. The term $\frac{1}{2}|\mathbf{b}|^2$ appears here because of this equality: $(\nabla \times \mathbf{H}) \times \mathbf{H} = (-\frac{1}{2}|\mathbf{b}|^2, \mathbf{b})_x$; both the pressure p and the internal energy e are related to the density and temperature of the flow according to the equations of state:

$$p = p(\rho, \theta), \qquad e = e(\rho, \theta); \tag{1.7}$$

$\lambda = \lambda(\rho, \theta)$ and $\mu = \mu(\rho, \theta)$ are the viscosity coefficients of the flow, $\nu = \nu(\rho, \theta)$ is the magnetic diffusivity, and $\kappa = \kappa(\rho, \theta)$ is the heat conductivity. All these dissipation coefficients depend on both ρ and θ generally.

We consider the initial-boundary value problem of (1.5) in a bounded spatial domain $\Omega = (0, 1)$ (without loss of generality) with the following initial condition and impermeable, thermally insulated boundaries:

$$\begin{aligned} &(\rho, u, \mathbf{w}, \mathbf{b}, \theta)|_{t=0} = (\rho_0, u_0, \mathbf{w}_0, \mathbf{b}_0, \theta_0)(x), \quad x \in \Omega, \\ &(u, \mathbf{w}, \mathbf{b}, \theta_x)|_{\partial\Omega} = 0, \end{aligned} \tag{1.8}$$

where the initial data satisfy certain compatibility conditions as usual. In this paper, we are interested in the well-posedness and regularity of global solutions to this initial-boundary value problem (1.5) and (1.8), as well as the following issue: whether shock waves, vacuum, and mass and heat concentration are developed in the solutions in a finite time, provided that the initial data are bounded, smooth, and do not contain a vacuum. There have been a lot of studies on MHD by physicists and mathematicians because of its physical importance, complexity, rich phenomena, and mathematical challenges; see [1, 3, 4, 6, 8, 12, 13, 16, 18, 19, 20, 23, 24] and the references cited therein. The above initial-boundary value problem is fundamental for the MHD system. We investigate such an important problem for the magnetohydrodynamic fluid flow with the pressure, internal energy, and heat conductivity satisfying certain physical growth conditions on the temperature. These growth conditions are motivated by the physical facts for certain important physical regimes where the temperature is high and experiences rapid change [25]. In particular, the case of perfect gases with $p = R\rho\theta$, $e = c_v\theta$ is included in the class of fluids investigated in this paper, where R is the gas constant, $c_v = R/(\gamma - 1)$ is the heat capacity of the gas at constant volume, and γ is the adiabatic exponent. For perfect gases with small smooth initial data, the existence of global solutions was proved in [12], and the large-time behavior was studied in [20]. For large initial data, these problems have additional difficulties because of the presence of the magnetic field and its interaction with the hydrodynamic motion of the flow of large oscillation. A free boundary value problem on real MHD flow was studied in [5].

The main goal of this paper is to establish the existence and uniqueness of a global solution to the initial-boundary value problem with general large initial data in H^1 and to show that neither shock waves nor vacuum and concentration are developed in a finite time. There have been some similar fundamental results on nonlinear thermoviscoelasticity [7] and on viscous heat-conductive real gases [14]. We consider the real magnetohydrodynamic flows with general pressure and internal energy, and permit the generation of heat by the magnetic field as well as its interaction with

the fluid motion. We remark that the fundamental idea in [15] for the Navier–Stokes equations of perfect compressible viscous flows does not apply to the problem under consideration for the magnetohydrodynamic flows with general equations of state (1.7). Our approach is then based on the methods in [7, 14]. We introduce the Lagrangian variable and transform the initial-boundary value problem (1.5)–(1.8) into a corresponding problem in Lagrangian coordinates. The existence and uniqueness of local solutions can be obtained by using the Banach theorem and the contractivity of the operator defined by the linearization of the problem on a small time interval (see [21]). The existence of global solutions is proved by extending the local solutions globally in time based on the global a priori estimates of solutions. We first obtain an entropy-type energy estimate involving the dissipative effects of viscosity, magnetic diffusion, and heat diffusion, which is essential for deducing the lower and upper bounds of the density. Some new techniques are developed to achieve these bounds. With these bounds, all the required a priori estimates are obtained subsequently by our careful analysis and techniques. In particular, the estimates of the temperature θ are complicated because of the complexity of the system for the general flow and will be achieved by developing some detailed analysis of the energy equation. The boundedness of the temperature will be proved by a maximum principle. We remark that for technical reasons the viscosity coefficients and magnetic diffusion coefficient are assumed to depend on the density only, and the heat conductivity cannot be a constant even in the case of perfect flows. It is an open problem to develop more effective techniques to remove these restrictions.

We reformulate the problem and state the main results in section 2, prove the existence of global solutions with initial data in H^1 by establishing the a priori estimates on the density in section 3, on the velocity and magnetic field in section 4, and on the temperature in section 5.

2. Reformulation of the initial-boundary value problem and main results. Consider the initial-boundary value problem (1.5)–(1.8) with positive lower and upper bounds of the initial density and temperature: $C_0^{-1} \le \rho_0, \theta_0 \le C_0$, for some constant $C_0 > 0$. Without loss of generality, we take $\int_0^1 \rho_0(x)dx = 1$. We first assume that $\rho, \theta > 0$, and then will prove their positive lower bounds later. We introduce the Lagrangian variable:

$$y = y(x,t) = \int_0^x \rho(\xi,t)d\xi. \tag{2.1}$$

We have $0 \le y \le 1$ since y is increasing in x and

$$\int_0^1 \rho(x,t)dx = \int_0^1 \rho_0(x)dx = 1.$$

We translate problem (1.5)–(1.8) in Eulerian coordinates into the following initial-boundary value problem in Lagrangian coordinates (y,t), $y \in \Omega = (0,1)$, a moving coordinate along the particle path:

$$v_t - u_y = 0, \tag{2.2a}$$

$$u_t + P_y = (\lambda\rho u_y)_y, \tag{2.2b}$$

$$\mathbf{w}_t - \mathbf{b}_y = (\mu\rho\mathbf{w}_y)_y, \tag{2.2c}$$

$$(v\mathbf{b})_t - \mathbf{w}_y = (\nu\rho\mathbf{b}_y)_y, \tag{2.2d}$$

$$E_t + (uP - \mathbf{w}\cdot\mathbf{b})_y = (\rho(\lambda u u_y + \mu\mathbf{w}\cdot\mathbf{w}_y + \nu\mathbf{b}\cdot\mathbf{b}_y + \kappa\theta_y))_y, \tag{2.2e}$$

with the initial-boundary conditions

$$(v,u,\mathbf{w},\mathbf{b},\theta)|_{t=0}=(v_0,u_0,\mathbf{w}_0,\mathbf{b}_0,\theta_0)(y),\quad y\in\Omega,$$
$$(u,\mathbf{w},\mathbf{b},\theta_y)|_{\partial\Omega}=0, \tag{2.3}$$

where $v=1/\rho$ is the specific volume, $p=p(v,\theta)$, $e=e(v,\theta)$, and

$$P=p+\frac{1}{2}|\mathbf{b}|^2,\qquad E=e+\frac{1}{2}(u^2+|\mathbf{w}|^2)+\frac{1}{2}v|\mathbf{b}|^2. \tag{2.4}$$

The second law of thermodynamics states the relation between p and e:

$$e_v(v,\theta)+p(v,\theta)=\theta p_\theta(v,\theta). \tag{2.5}$$

Problem (1.5)–(1.8) and problem (2.2)–(2.3) are equivalent for the solutions under consideration. Our main interest is to study the behavior of solutions of this problem with physical equations of state and various physical viscosity coefficients λ, μ, magnetic diffusivity ν, and heat conductivity κ. We assume that p and e are continuously differentiable and κ is twice continuously differentiable in $v>0$ and $\theta\ge 0$. For technical reasons we assume that the viscosity coefficients λ, μ, and ν depend only on v, with continuous first derivatives in $v>0$ such that $\lambda_1\le\lambda\le\lambda_2$, $\mu_1\le\mu\le\mu_2$, $\nu_1\le\nu\le\nu_2$ for $v>0$ and for some positive constants λ_i,μ_i,ν_i $(i=1,2)$. It is an open problem to handle the case where these viscosity coefficients depend also on temperature. We also assume the growth conditions with exponents $r\in[0,1]$ and $q\ge 2+2r$ such that

(1) there exists a constant $e_0>0$ so that, for $v>0$ and $\theta\ge 0$,

$$p_v(v,\theta)\le 0,\quad e(v,\theta)\ge e_0(1+\theta^{1+r}); \tag{2.6}$$

(2) for any given $v_1>0$, there exist positive constants $\kappa_0=\kappa_0(v_1)$, $p_0=p_0(v_1)$, $e_1=e_1(v_1)$ such that, for $v\ge v_1$, $\theta\ge 0$,

$$0\le vp(v,\theta)\le p_0\left(1+\theta^{1+r}\right),\quad \kappa(v,\theta)\ge\kappa_0(1+\theta^q),\quad e_\theta(v,\theta)\ge e_1(1+\theta^r); \tag{2.7}$$

(3) for any given $v_2>v_1>0$, there exist positive constants $p_i=p_i(v_1,v_2)$ $(i=1,2,3)$, $e_j=e_j(v_1,v_2)$ $(j=2,3)$, and $\kappa_1=\kappa_1(v_1,v_2)$ so that, for any $v\in[v_1,v_2]$, $\theta\ge 0$,

$$|vp_\theta(v,\theta)|\le p_1\left(1+\theta^r\right),\qquad -p_3\left(1+\theta^{1+r}\right)\le v^2p_v(v,\theta)\le -p_2\left(1+\theta^{1+r}\right), \tag{2.8}$$
$$|e_v(v,\theta)|\le e_2(1+\theta^{1+r}),\qquad e_\theta(v,\theta)\le e_3(1+\theta^r), \tag{2.9}$$
$$\kappa(v,\theta)+|\kappa_v(v,\theta)|+|\kappa_{vv}(v,\theta)|\le\kappa_1(1+\theta^q). \tag{2.10}$$

These growth conditions are motivated by the physical facts: $e\propto\theta^{1+r}$ with $r\approx 0.5$ and $\kappa\propto\theta^{5/2}$ for important physical regimes where the temperature is high and changes rapidly; see [2, 6, 23, 25]. The perfect gases are included that correspond to the special case $r=0$. We remark that, if $r=0$, we still need the dependence of the heat conductivity on temperature, and new analysis is required to deal with the constant heat conductivity. For the initial-boundary value problem (2.2), (2.3), we will see that, if the initial data is in H^1, then the solution will be at least in H^1, and neither shock waves nor vacuum and concentration are developed in a finite time. Precisely, we have the following results.

MAIN THEOREM. *If for some constant $C_0 > 0$,*

$$C_0^{-1} \le v_0,\, \theta_0 \le C_0, \quad \|(u_0, \mathbf{w}_0, \mathbf{b}_0)\|_{L^4} + \|\theta_0\|_{L^2} \le C_0,$$
$$\|(v_0, u_0, \mathbf{w}_0, \mathbf{b}_0, \theta_0)\|_{H^1} \le C_0,$$

and $v_0 \in W^{1,\infty}(\Omega)$, then the initial-boundary value problem (2.2), (2.3) *has a unique global solution $(v, u, \mathbf{w}, \mathbf{b}, \theta)(y,t)$ such that, for any fixed $T > 0$,*

$$v \in L^\infty(0,T; H^1 \cap W^{1,\infty}(\Omega)), \qquad (u, \mathbf{w}, \mathbf{b}, \theta) \in L^\infty(0,T; H^1(\Omega)),$$

and, for each $(y,t) \in \Omega \times [0,T]$,

$$\begin{aligned} &C^{-1} \le v(y,t),\, \theta(y,t) \le C,\\ &\|(u, \mathbf{w}, \mathbf{b})\|_{L^2(0,T;L^4\cap H^1)} + \|\theta\|_{L^2(0,T;L^2\cap H^1)} \le C,\\ &\|(v, u, \mathbf{w}, \mathbf{b}, \theta)\|^2_{H^1}(t) + \int_0^t \|(v_{yt}, u_{yy}, \mathbf{w}_{yy}, \mathbf{b}_{yy}, \theta_{yy})\|^2_{L^2}(s)ds \le C, \end{aligned} \tag{2.11}$$

where $C > 0$ is some constant.

The results for the initial-boundary value problem (2.2), (2.3) in the main theorem in Lagrangian coordinates can easily be converted to equivalent statements for the corresponding results for the initial-boundary value problem (1.5), (1.8) in Eulerian coordinates. The details are omitted.

The existence and uniqueness of the local solution to the initial-boundary value problem (2.2), (2.3) is known from the standard method based on the Banach theorem and the contractivity of the operator defined by the linearization of the problem on a small time interval [21]. The global existence of a solution can be obtained by combining the local existence and the global a priori estimates (2.11) using a standard argument (see [9], for example). The uniqueness of the global solution follows from the uniqueness of the local solution. Therefore the remaining task is to establish the global a priori estimates (2.11).

3. A priori estimates on density. We will use the following notation: $U = (u, \mathbf{w}, \mathbf{b})$, $\Pi_t = (0,1) \times (0,t)$, and $\vartheta = 1/\theta$. Denote the L^1-norm on $\Omega \subset \mathbb{R}$ by $\|\cdot\| = \int_\Omega |\cdot| dy$, and the L^1-norm on $\Pi_t \subset \mathbb{R}^2$ by $|||\cdot||| = \iint_{\Pi_t} |\cdot|\, dy\, ds$. For $s > 1$, denote the L^s-norm on $\Omega \subset \mathbb{R}$ by $\|\cdot\|_s$, and the L^s-norm on $\Pi_t \subset \mathbb{R}^2$ by $|||\cdot|||_s$. In the rest of this paper, we will establish the a priori estimates of the solutions for $x \in \Omega$ and $t \in (0,T)$ with any T fixed.

In this section we prove the following estimates.

LEMMA 3.1.

$$C^{-1} \le v \le C, \quad \big\|v + \theta + \theta^{1+r} + |U|^2 + e\big\| + \big|\!\big|\!\big| K + |U_y|^2 + \theta_y^2 \big|\!\big|\!\big| \le C,$$

where

$$K = \rho\vartheta\left(\kappa\vartheta\theta_y^2 + \lambda u_y^2 + \mu|\mathbf{w}_y|^2 + \nu|\mathbf{b}_y|^2\right).$$

Proof. We have first from (2.2a), by integration, $\|v\| = \|v_0\| = 1$ without loss of generality. Integrating the energy equation (2.2e) and using (2.6) and the Cauchy–Schwarz inequality yields

$$\|e + u^2 + |\mathbf{w}|^2 + v|\mathbf{b}|^2 + \theta + \theta^{1+r}\| \le C.$$

Set $\psi(v,\theta) \equiv e(v,\theta) - \theta\eta(v,\theta)$, where $\eta(v,\theta)$ is defined by the relations $e_\theta = \theta\eta_\theta$, $\eta_v = p_\theta$. Then $e_\theta = -\theta\psi_{\theta\theta}$ and $\psi_v = -p$. Take $\Phi_1 \equiv 2(\psi(v,\theta) - \psi(v,1) - \psi_\theta(v,\theta)(\theta - 1))$ and $\Phi_2 \equiv 2(\psi(v,1) - \psi(1,1) - \psi_v(1,1)(v-1))$. Then

$$\begin{aligned}&\frac{1}{2}\left(\Phi_1 + \Phi_2 + u^2 + |\mathbf{w}|^2 + v|\mathbf{b}|^2\right)_t + K\\ &\quad = \left(\rho(\lambda u u_y + \mu \mathbf{w}\cdot\mathbf{w}_y + \nu\mathbf{b}\cdot\mathbf{b}_y + \kappa\theta_y) + p(1,1)u - uP + \mathbf{w}\cdot\mathbf{b} - \kappa\rho\vartheta\theta_y\right)_y,\end{aligned}$$

where $p(1,1)$ is the value of $p(v,\theta)$ at $v = \theta = 1$. Integrating the above equation over Π_t yields

$$\frac{1}{2}\int_\Omega (\Phi_1 + \Phi_2)dy + \frac{1}{2}\left\|u^2 + |\mathbf{w}|^2 + v|\mathbf{b}|^2\right\| + \|\!|K|\!\| \le C.$$

By (2.6),

$$\Phi_1 \ge 2(\theta-1)^2 \int_0^1 \frac{e_\theta(v, 1 + s(\theta-1))}{1 + s(\theta-1)}\, s\, ds \ge 0,$$

and $\Phi_2 = -p_v(\xi,1)(v-1)^2 \ge 0$ for some ξ using Taylor's expansion; therefore

$$\|\!|K|\!\| \le C. \tag{3.1}$$

Set

$$w(y,t) \equiv \int_0^t (\lambda\rho u_y - P)(y,s)ds + \int_0^y u_0(\xi)d\xi;$$

then $w_y = u$ from (2.2b), and thus

$$w(y,t) = w(a(t),t) + \int_{a(t)}^y u(\xi,t)d\xi,$$

where $a(t) \in [0,1]$ will be determined later. On the other hand,

$$w(y,t) = \int_0^t (V_t - P)ds + \int_0^y u_0(\xi)d\xi,$$

where $V(v) \equiv \int_1^v \lambda(\xi)\xi^{-1}d\xi$ is increasing in v and $\lambda_1 \ln v \le V(v) \le \lambda_2 \ln v$; then

$$\begin{aligned}V(y,t) &= V(y,0) + w(y,t) + \int_0^t P ds - \int_0^y u_0(\xi)d\xi\\ &= V(y,0) + w(a(t),t) + \int_0^t P ds + \int_{a(t)}^y u(\xi,t)d\xi - \int_0^y u_0(\xi)d\xi.\end{aligned} \tag{3.2}$$

From the definition of w, we have $w_t = \lambda\rho u_y - P$; then

$$(vw)_t = (uw)_y - u^2 - vP + \lambda u_y,$$

and by integration,

$$\int_\Omega vw dy = \int_\Omega v_0 w_0 dy - \iint_{\Pi_t} \left(u^2 + vP\right) dy\, ds + \iint_{\Pi_t} \lambda u_y\, dy\, ds.$$

There exists a function $a(t) \in [0,1]$ for any $t > 0$ such that

$$\int_\Omega vwdy = w(a(t),t)\int_\Omega vdy = w(a(t),t).$$

Therefore

$$\begin{aligned} V(y,t) = V(y,0) &+ \int_\Omega v_0(y)\int_0^y u_0(\xi)d\xi dy + \int_0^t P(y,s)ds - \iint_{\Pi_t}\left(u^2+vP\right)\,dy\,ds \\ &+ \int_{a(t)}^y u(\xi,t)d\xi - \int_0^y u_0(\xi)d\xi + \iint_{\Pi_t}\lambda u_y\,dy\,ds. \end{aligned} \tag{3.3}$$

Notice that, from (3.1),

$$\begin{aligned} \|\lambda u_y\| &\le \left\|\lambda\rho\vartheta u_y^2\right\| + C\int_0^t \max_y\theta\,\|v\|\,ds \\ &\le C + C\,\|\theta+\theta_y\| \le C + C\int_0^t \left\|\kappa\rho\vartheta^2\theta_y^2\right\|\,\|v\|\,ds \le C + C\,\|K\| \le C, \end{aligned} \tag{3.4}$$

and from (2.7),

$$\|u^2 + vP\| \le C\|u^2 + 1 + \theta^{1+r} + |\mathbf{b}|^2\| \le C.$$

Then $V(y,t) \ge -C$ since $P \ge 0$. Thus $v \ge C^{-1}$ and $\rho \le C$.

From (2.2), we have

$$e_t + pu_y = (\kappa\rho\theta_y)_y + \rho\left(\lambda u_y^2 + \mu|\mathbf{w}_y|^2 + \nu|\mathbf{b}_y|^2\right). \tag{3.5}$$

Integrating (3.5) and using (2.7), we obtain

$$\begin{aligned} &\iint_{\Pi_t}\rho\left(\lambda u_y^2 + \mu|\mathbf{w}_y|^2 + \nu|\mathbf{b}_y|^2\right)\,dy\,ds \\ &\quad\le C + \frac{1}{2}\iint_{\Pi_t}\lambda\rho u_y^2\,dy\,ds + C\iint_{\Pi_t}\rho\left(1+\theta^{2+2r}\right)\,dy\,ds, \end{aligned}$$

and then, from (2.7) and (3.1),

$$\begin{aligned} \frac{1}{2}\left\|\rho\left(\lambda u_y^2 + \mu|\mathbf{w}_y|^2 + \nu|\mathbf{b}_y|^2\right)\right\| &\le C + C\iint_{\Pi_t}\left(\|\theta^{1+r}\| + \|\theta^r\theta_y\|\right)^2\,dy\,ds \\ &\le C + C\int_0^t \|\kappa\rho\vartheta^2\theta_y^2\|\,\|v\|\,ds \le C + C\,\|K\| \le C. \end{aligned}$$

Thus, from the boundary condition (2.3) on $\mathbf{b}$ and the estimates above,

$$\int_0^t \frac{1}{2}|\mathbf{b}|^2 ds = \int_0^t\int_0^y \mathbf{b}\cdot\mathbf{b}_y ds \le \|v|\mathbf{b}|^2\| + C\|\nu\rho|\mathbf{b}_y|^2\| \le C.$$

Then

$$\begin{aligned} V(y,t) &\le C + C\int_0^t \theta^{1+r}ds + \int_0^t \frac{1}{2}|\mathbf{b}|^2 ds \le C + C\|\theta^{1+r} + \theta^r\theta_y\| \\ &\le C + C\|\kappa\rho\vartheta^2\theta_y^2\| + C\,\|v\| \le C, \end{aligned}$$

that is, $v \le C$ and $\rho \ge C^{-1}$. This completes the proof of Lemma 3.1. ∎

4. A priori estimates on velocity and the magnetic field. We now establish the following estimates on $U = (u, \mathbf{w}, \mathbf{b})$.

LEMMA 4.1.

$$\|(v_y, U_y)\|_2^2 + \|\!|U_{yy}|\!\|_2^2 + \|\!|U_y|\!\|_4^4 + |U|^2 \le C.$$

Proof. Rewrite (2.2b) as $(V_y - u)_t = P_y$, multiply it by $V_y - u$, and integrate it to obtain

$$\frac{1}{2}\|V_y - u\|_2^2 \le C + \iint_{\Pi_t} (p_v v_y + p_\theta \theta_y + \mathbf{b}\cdot\mathbf{b}_y)(V_y - u)\, dy\, ds.$$

From (2.7) we observe that

$$\begin{aligned} &\int_0^t \max_{y\in\Omega}\left(1+\theta^{1+r}\right) ds + \int_0^t \max_{y\in\Omega}\left(1+\theta^{2+2r}\right) ds + \int_0^t \max_{y\in\Omega}\left(1+\theta^{1+r}\right)^2 ds \\ &\le C\int_0^t \max_{y\in\Omega}\left(1+\theta^{1+r}\right)^2 ds \le C\int_0^t \left(1+\|\theta^{1+r}\| + \|\theta^r\theta_y\|\right)^2 ds \\ &\le C + C\|\!|\kappa\vartheta^2\theta_y^2|\!\| \le C; \end{aligned} \tag{4.1}$$

then, using (2.8) and $V_y = \lambda\rho v_y$, we obtain that there exists some constant $C_1 > 0$ such that

$$\begin{aligned} \frac{1}{2}\|V_y - u\|_2^2 &\le C - C_1\|\!|\left(1+\theta^{1+r}\right)v_y^2|\!\| + C\|\!|\left(1+\theta^{1+r}\right)v_y u|\!\| \\ &\quad + C\|\!|(1+\theta^r)\,\theta_y\,(V_y+u)|\!\| + \iint_{\Pi_t}\mathbf{b}\cdot\mathbf{b}_y\,(V_y-u)\,dy\,ds \\ &\le C - \frac{C_1}{2}\|\!|\left(1+\theta^{1+r}\right)v_y^2|\!\| + C\int_0^t \max_{y\in\Omega}\left(1+\theta^{1+r}\right)\|u^2\| ds \\ &\quad + C\|\!|\kappa\vartheta^2\theta_y^2|\!\| + C\,\|\!|(\mathbf{b}\cdot\mathbf{b}_y, u)|\!\|_2^2 \\ &\le C - \frac{C_1}{2}\|\!|\left(1+\theta^{1+r}\right)v_y^2|\!\| + C\,\|\!|\mathbf{b}\cdot\mathbf{b}_y|\!\|_2^2. \end{aligned}$$

Then

$$\|V_y - u\|_2^2 + \|\!|\left(1+\theta^{1+r}\right)v_y^2|\!\| \le C + C\,\|\!|\mathbf{b}\cdot\mathbf{b}_y|\!\|_2^2$$

and

$$\begin{aligned} \|v_y\|_2^2 &\le C\,\|V_y\|_2^2 \le C\,\|V_y - u\|_2^2 + C\,\|u\|_2^2 \\ &\le C + C\,\|\!|\mathbf{b}\cdot\mathbf{b}_y|\!\|_2^2. \end{aligned} \tag{4.2}$$

Multiplying (2.2e) by $K_1 E$ and (2.2b) by $K_2 u^3$, taking the inner product of (2.2c) with $K_3|\mathbf{w}|^2\mathbf{w}$ and (2.2d) with $K_4|v\mathbf{b}|^2 v\mathbf{b}$, multiplying (4.2) by K_5, respectively, for proper positive constants $K_j, 1 \le j \le 5$, integrating them over $[0,1]\times[0,t]$, adding them all together, and using Gronwall's inequality, we have $\|v_y\|_2^2 + \|\!|\mathbf{b}\cdot\mathbf{b}_y|\!\|_2^2 \le C$ by tedious calculations. Multiply (2.2b) by u_{yy} and integrate on $[0,1]\times[0,t]$ to obtain, using the interpolation inequality on u_y,

$$|u_y|^2 \le C(1+\delta^{-1})\,\|u_y\|_2^2 + \delta\|\lambda\rho u_{yy}^2\| \tag{4.3}$$

for any $\delta > 0$, and from (4.1),

$$\begin{aligned}
&\frac{1}{2}\|u_y\|_2^2 \le \frac{1}{2}\|u_{0y}\|_2^2 - \iint_{\Pi_t} \left(\lambda\rho u_{yy} + (\lambda\rho)_v v_y u_y - p_v v_y - p_\theta \theta_y - \mathbf{b}\cdot\mathbf{b}_y\right) u_{yy}\, dy\, ds \\
&\le C - \frac{1}{2}\left\|\!\left|\lambda\rho u_{yy}^2\right|\!\right\| + C\int_0^t \max_y \left(u_y^2 + \theta^{2+2r}\right)\left\|v_y^2\right\| ds + C\left\|\!\left|\kappa\vartheta^2\theta_y^2 + |\mathbf{b}\cdot\mathbf{b}_y|^2\right|\!\right\| \\
&\le C - \frac{1}{4}\left\|\!\left|\lambda\rho u_{yy}^2\right|\!\right\| + C\left\|u_y\right\|_2^2 \le C - \frac{1}{4}\left\|\!\left|\lambda\rho u_{yy}^2\right|\!\right\|;
\end{aligned}$$

then

$$\|u_y\|_2^2 + \left\|u_{yy}\right\|_2^2 \le C$$

and

$$\left\|u_y\right\|_4^4 = \int_0^t \max_y u_y^2 \|u_y\|_2^2\, ds \le C\left\|(u_y, u_{yy})\right\|_2^2 \le C.$$

Multiplying (2.2c) by $\mathbf{w}_{yy}$ and then integrating, and using the interpolation inequality on $\mathbf{w}_y$ similar to (4.3), we have

$$\begin{aligned}
\frac{1}{2}\|\mathbf{w}_y\|_2^2 &= \frac{1}{2}\|\mathbf{w}_y(y,0)\|_2^2 - \iint_{\Pi_t}\left(\mathbf{b} + \frac{\mu\mathbf{w}_y}{v}\right)_y \cdot \mathbf{w}_{yy}\, dy\, ds \\
&\le C - C_2\left\|\mathbf{w}_{yy}\right\|_2^2 + C\left\|(|\mathbf{b}_y| + |v_y||\mathbf{w}_y|)\,\mathbf{w}_{yy}\right\| \\
&\le C - \frac{3C_2}{4}\left\|\mathbf{w}_{yy}\right\|_2^2 + C\left\|\mathbf{b}_y\right\|_2^2 + C\int_0^t \max_{y\in\Omega}|\mathbf{w}_y|^2\,\|v_y\|_2^2\, ds \\
&\le C - \frac{C_2}{2}\left\|\mathbf{w}_{yy}\right\|_2^2 + C\left\|\mathbf{w}_y\right\|_2^2 \le C - \frac{C_2}{2}\left\|\mathbf{w}_{yy}\right\|_2^2,
\end{aligned}$$

and then

$$\|\mathbf{w}_y\|_2^2 + \left\|\mathbf{w}_{yy}\right\|_2^2 \le C.$$

As a consequence,

$$\left\|\mathbf{w}_y\right\|_4^4 \le \int_0^t \max_{y\in\Omega}|\mathbf{w}_y|^2\,\|\mathbf{w}_y\|_2^2\, ds \le C\left\|(\mathbf{w}_y, \mathbf{w}_{yy})\right\|_2^2 \le C.$$

Using $v_t = u_y$, we rewrite (2.2d) as

$$\mathbf{b}_t = \frac{u_y}{v}\mathbf{b} + \frac{1}{v}\left(\mathbf{w} + \frac{\nu\mathbf{b}_y}{v}\right)_y. \tag{4.4}$$

Multiplying the above equation by $\mathbf{b}_{yy}$, integrating it, and using Lemma 3.1 and the similar interpolation inequalities on $\mathbf{b}$ and $\mathbf{b}_y$ to (4.3) yields

$$\begin{aligned}
\frac{1}{2}\|\mathbf{b}_y\|_2^2 &\le C - C_3\left\|\mathbf{b}_{yy}\right\|_2^2 + C\left\|(|u_y||\mathbf{b}| + |\mathbf{w}_y| + |v_y||\mathbf{b}_y|)\,\mathbf{b}_{yy}\right\| \\
&\le C - \frac{3C_3}{4}\left\|\mathbf{b}_{yy}\right\|_2^2 + C\max_{y,t}|\mathbf{b}|^2\left\|u_y\right\|_2^2 + C\int_0^t \max_{y\in\Omega}|\mathbf{b}_y|^2\,\|v_y\|_2^2\, ds \\
&\le C - \frac{C_3}{2}\left\|\mathbf{b}_{yy}\right\|_2^2 + C\|\mathbf{b}\|_2^2 + \frac{1}{4}\|\mathbf{b}_y\|_2^2 + C\left\|\mathbf{b}_y\right\|_2^2 \\
&\le C - \frac{C_3}{2}\left\|\mathbf{b}_{yy}\right\|_2^2 + \frac{1}{4}\|\mathbf{b}_y\|_2^2.
\end{aligned}$$

Then

$$\|\mathbf{b}_y\|_2^2 + \|\mathbf{b}_{yy}\|_2^2 \le C.$$

Thus

$$\|\mathbf{b}_y\|_4^4 \le \int_0^t \max_{y\in\Omega} |\mathbf{b}_y|^2 \, \|\mathbf{b}_y\|_2^2 \, ds \le C \, \|(\mathbf{b}_y, \mathbf{b}_{yy})\|_2^2 \le C.$$

Using Lemma 3.1 and the above estimates, one obtains the following:

$$|U|^2 \le C \, \|U\|_2^2 + C \, \|U_y\|_2^2 \le C.$$

The proof of Lemma 4.1 is complete. □

5. A priori estimates on temperature. We now make estimates on the temperature θ. Due to the complicated structures of the pressure, internal energy, and heat conductivity, as well as the strong coupling of the equations, the estimates on the temperature θ are quite complex.

LEMMA 5.1.

$$\|(\theta_y, U_t, U_{yy})\|_2^2 + \|(\theta_t, U_t)\|_2^2 + |(v_y, U_y)|^2 \le C, \quad C^{-1} \le \theta \le C. \tag{5.1}$$

Proof. To make the estimates on θ, we first define the following:

$$\Theta = \max_{\Pi_T} \theta(y,t),$$
$$X = \left\| \left(1+\theta^{q+r}\right)\theta_t^2 \right\|, \quad Y = \max_{t\in[0,T]} \left\| \left(1+\theta^{2q}\right)\theta_y^2 \right\|, \quad Z = \max_{t\in[0,T]} \|u_{yy}\|_2^2.$$

We will show that Θ, X, and Y can be controlled by Z, and then we will derive an inequality on Z which yields the upper bound of Z and thus the upper bounds of Θ, X, and Y.

From Lemma 4.1, one has

$$u_y^2 \le \|u_y\|_2^2 + 2 \, \|u_y\|_2 \, \|u_{yy}\|_2 \le C + CZ^{1/2},$$

which implies

$$|u_y| \le C + CZ^{1/4}. \tag{5.2}$$

Define $y(t) \in [0,1]$ such that $\theta(y(t),t) = \int_0^1 \theta(y,t)dy \le C$. From Lemma 3.1, we have

$$\begin{aligned}\theta^{(2q+3+r)/2}(y,t) &= \|\theta\|^{(2q+3+r)/2} + \frac{2q+3+r}{2} \int_{y(t)}^{y} \theta^{(2q+1+r)/2}\theta_y d\xi \\ &\le C + C\left(\int_0^1 \theta^{2q}\theta_y^2 dy\right)^{1/2}\left(\int_0^1 \theta^{1+r} dy\right)^{1/2} \le C + CY^{1/2},\end{aligned}$$

and thus

$$\Theta \le C + CY^{\beta_1}, \tag{5.3}$$

where $\beta_1 = 1/(2q+3+r)$. The rest of the proof can be divided into four steps.

Step 1. We now show that X and Y can be controlled by Z. Use (2.5) and rewrite (3.5) as

$$e_\theta\theta_t+\theta p_\theta u_y=\left(\frac{\kappa\theta_y}{v}\right)_y+\frac{\lambda u_y^2}{v}+\frac{\mu|\mathbf{w}_y|^2}{v}+\frac{\nu|\mathbf{b}_y|^2}{v}. \tag{5.4}$$

Set $H(v,\theta)=v^{-1}\int_0^\theta \kappa(v,\xi)\,d\xi$. Multiplying (5.4) by H_t, using integration by parts and the boundary condition (2.3), we have

$$\iint_{\Pi_t}\left(e_\theta\theta_t+\theta p_\theta u_y-\frac{\lambda}{v}u_y^2-\frac{\mu}{v}|\mathbf{w}_y|^2-\frac{\nu}{v}|\mathbf{b}_y|^2\right)H_t\,dy\,ds+\iint_{\Pi_t}\frac{\kappa}{v}\theta_y H_{ty}\,dy\,ds=0, \tag{5.5}$$

where

$$H_t=H_v u_y+\frac{\kappa}{v}\theta_t,\qquad H_{ty}=H_v u_{yy}+H_{vv}u_y v_y+\left(\frac{\kappa}{v}\right)_v\theta_t v_y+\left(\frac{\kappa}{v}\theta_y\right)_t.$$

We now estimate all the terms in the above equality. First we have, from (2.6) and (2.7),

$$\iint_{\Pi_t} e_\theta\theta_t\frac{\kappa}{v}\theta_t\,dy\,ds\ge C_4X,\quad \int_0^T\int_0^1\frac{\kappa}{v}\theta_y\left(\frac{\kappa}{v}\theta_y\right)_t\,dy\,ds\ge C_5Y-C,$$

for some positive constants C_4 and C_5. From (2.10),

$$|H_v|+|H_{vv}|\le C\int_0^\theta\left(|\kappa|+|\kappa_v|+|\kappa_{vv}|\right)d\xi\le C\left(1+\theta^{q+1}\right).$$

Using (2.8), (2.9), (5.2), (5.3), Lemmas 3.1 and 4.1, and Young's inequality, we have

$$\begin{aligned}&\iint_{\Pi_t} e_\theta\theta_t H_v u_y\,dy\,ds\le C\left\|\left(1+\theta^{q+r+1}\right)\theta_t u_y\right\|\\&\le\frac{C_4}{8}X+C\left(1+\Theta^{q+2+r}\right)\left\|u_y^2\right\|\le\frac{C_4}{8}X+CY^{(q+2+r)\beta_1}+C\\&\le\frac{C_4}{8}X+\frac{C_5}{8}Y+C\end{aligned}$$

and

$$\begin{aligned}&\iint_{\Pi_t}\left(\theta p_\theta u_y-\frac{\lambda}{v}u_y^2-\frac{\mu}{v}|\mathbf{w}_y|^2-\frac{\nu}{v}|\mathbf{b}_y|^2\right)H_v u_y\,dy\,ds\\&\le C\left(1+\Theta^{q+1}\right)\max_{y,t}|u_y|\left\|\left(1+\theta^{1+r}\right)|u_y|+u_y^2+|\mathbf{w}_y|^2+|\mathbf{b}_y|^2\right\|\\&\le C\left(1+\Theta^{q+1}\right)\left(1+Z^{1/4}\right)\left\|\left(1+\theta^{1+r}\right)^2+u_y^2+|\mathbf{w}_y|^2+|\mathbf{b}_y|^2\right\|\\&\le C+\frac{C_5}{8}Y+CZ^{\beta_2},\end{aligned}$$

with $\beta_2=(2q+3+r)/(4(q+1+r))$ and

$$\begin{aligned}&\iint_{\Pi_t}\left(\theta p_\theta u_y-\frac{\lambda}{v}u_y^2-\frac{\mu}{v}|\mathbf{w}_y|^2-\frac{\nu}{v}|\mathbf{b}_y|^2\right)\frac{\kappa}{v}\theta_t\,dy\,ds\\&\le C\left\|\left(\left(1+\theta^{1+r}\right)|u_y|+u_y^2+|\mathbf{w}_y|^2+|\mathbf{b}_y|^2\right)\left(1+\theta^q\right)\theta_t\right\|\\&\le\frac{C_4}{8}X+C\left(1+\Theta^{q+2+r}\right)\|u_y\|_2^2+C\left(1+\Theta^{q-r}\right)\|(u_y,\mathbf{w}_y,\mathbf{b}_y)\|_4^4\\&\le\frac{C_4}{8}X+C\Theta^{q+2+r}+C\le\frac{C_4}{8}X+\frac{C_5}{8}Y+C.\end{aligned}$$

Since $\theta^q(y,t) = \|\theta\|^q + \int_{y(t)}^{y} q\theta^{q-1}\theta_y d\xi$, we use Lemma 3.1 and Young's inequality to obtain

$$\begin{aligned}
&\int_0^t \max_{y\in\Omega}\theta^q ds \le C + C\left\|\!\left|\theta^{q-1}\theta_y\right|\!\right\| \le C + \int_0^t \max_{y\in\Omega}\theta^{q-1}\,\|\theta\|\,ds + C\left\|\!\left|\kappa\vartheta^2\theta_y^2\right|\!\right\| \\
&\le C + \int_0^t \max_{y\in\Omega}\theta^{q-1} ds \le C + \frac{1}{2}\int_0^t \max_{y\in\Omega}\theta^q ds,
\end{aligned}$$

and then

$$\int_0^t \max_{y\in\Omega}\theta^q ds \le C. \tag{5.6}$$

From (2.6)–(2.10), (5.2), (5.3), (5.6), Lemmas 3.1 and 4.1, and Young's inequality, we have

$$\begin{aligned}
&\iint_{\Pi_t} \frac{\kappa}{v}\theta_y H_{vv} u_y v_y\, dy\, ds \le C\left(1+\Theta^{q+1}\right)\max_{y,t}|u_y|\,\|\!|\kappa\theta_y v_y|\!\| \\
&\le C\left(1+\Theta^{q+2}\right)\max_{y,t}|u_y|\left(\int_0^t \max_y\left(1+\theta^q\right)\left\|v_y^2\right\| ds\right)^{1/2}\left\|\!\left|\kappa\vartheta^2\theta_y^2\right|\!\right\|^{1/2} \\
&\le C(1+Z^{1/4})(1+Y^{(q+2)\beta_1}) \le \frac{C_6}{8}Y + CZ^{\beta_2} + C.
\end{aligned}$$

Using (2.10), (5.3), Lemma 4.1, and Young's inequality yields

$$\begin{aligned}
&\iint_{\Pi_t} \frac{\kappa}{v}\theta_y\left(\frac{\kappa}{v}\right)_v \theta_t v_y\, dy\, ds \le C\,\|\!|\left(1+\theta^q\right)\theta_t\left(\kappa\rho\theta_y\right)v_y|\!\| \\
&\le \frac{C_5}{16}X + C\left(1+\Theta^{q-r}\right)\int_0^t \max_y\left|\kappa\rho\theta_y\right|^2\|v_y\|_2^2\, ds \\
&\le \frac{C_5}{16}X + C\left(1+\Theta^{q-r}\right)\int_0^t\left(\|\kappa\rho\theta_y\|_2^2 + 2\|\kappa\rho\theta_y\left(\kappa\rho\theta_y\right)_y\|\right) ds \\
&\le \frac{C_5}{16}X + M\left(1+\Theta^{2q+2-r}\right)\left\|\!\left|\kappa\vartheta^2\theta_y^2\right|\!\right\| + C\left(1+\Theta^{q-r}\right)\|\!|\kappa\rho\theta_y\left(\kappa\rho\theta_y\right)_y|\!\| \\
&\le \frac{C_5}{16}X + C + CY^{(2q+2-r)\beta_1} + C\left(1+\Theta^{q+1-r}\right)\|\!|\kappa\left(\kappa\rho\theta_y\right)_y^2|\!\|^{1/2}\left\|\!\left|\kappa\vartheta^2\theta_y^2\right|\!\right\|^{1/2}.
\end{aligned}$$

Using (5.4), we continue to have

$$\begin{aligned}
&\iint_{\Pi_t} \frac{\kappa}{v}\theta_y\left(\frac{\kappa}{v}\right)_v \theta_t v_y\, dy\, ds \\
&\le \frac{C_5}{16}X + \frac{C_6}{16}Y + C + C\left(1+\Theta^{q+1-r}\right)\left\|\!\left|\kappa\left(e_\theta^2\theta_t^2 + \theta^2 p_\theta^2 u_y^2 + |U_y|^4\right)\right|\!\right\|^{1/2} \\
&\le \frac{C_5}{16}X + \frac{C_6}{16}Y + C + C(1+\Theta^{(2q+2-r)/2})X^{1/2} + C(1+\Theta^{(3q+4)/2})\,\|\!|u_y|\!\|_2 \\
&\quad + M(1+\Theta^{(3q+2-2r)/2})\,\|\!|U_y|\!\|_4^2 \\
&\le \frac{C_5}{16}X + \frac{C_6}{16}Y + C + C(1+\Theta^{(2q+2-r)/2})X^{1/2} + C\Theta^{(3q+4)/2} + C \\
&\le \frac{C_5}{8}X + \frac{C_6}{16}Y + C\Theta^{2q+2-r} + C \le \frac{C_5}{8}X + \frac{C_6}{8}Y + C.
\end{aligned}$$

From (2.6), (2.10), (5.2), (5.3), (5.6), Lemma 3.1, and Young's inequality, we have

$$\begin{aligned}
&\iint_{\Pi_t} \frac{\kappa}{v}\theta_y H_v u_{yy}\,dy\,ds \le C\left\|\left(1+\theta^{q+1}\right)\kappa\theta_y u_{yy}\right\| \\
&\le C(1+\Theta^{(4q+5)/4})\left\|\kappa\vartheta^2\theta_y^2\right\|^{1/2}\left(Z\int_0^t\left(1+\max_y\theta^{(2q+3)/2}\right)ds\right)^{1/2} \\
&\le C(1+Y^{\beta_1(4q+5)/4})Z^{1/2} \le \frac{C_6}{8}Y + CZ^{\beta_3} + C,
\end{aligned}$$

where $\beta_3 = 2(2q+3+r)/(4q+7+4r) \in (0,1)$ and the following estimate is employed:

$$\int_0^t \max_y \theta^{(2q+3)/2} ds \le C, \tag{5.7}$$

which will be shown in the next step. Assuming the estimate (5.7), we conclude, from (5.5), that

$$X + Y \le C + CZ^{\beta_4}, \tag{5.8}$$

where $0 < \beta_4 = \max\{\beta_2, \beta_3\} < 1$.

Step 2. The proof of (5.7) is as follows. For some small δ, from Lemma 3.1,

$$\begin{aligned}
\int_0^t \max_y \theta^{(2q+3)/2} ds &\le \int_0^t \left(\|\theta\|^{(2q+3)/2} + \|\theta^{(2q+1)/2}\theta_y\|\right) ds \\
&\le C + \delta\int_0^t \max_y \theta^{(2q+3)/2}\|\theta^{1+r}\| ds + C\|\theta^{(2q-3-2r)/2}\theta_y^2\| \\
&\le C + \frac{1}{2}\int_0^t \max_y \theta^{(2q+3)/2} ds + C\|\theta^{(2q-3-2r)/2}\theta_y^2\|,
\end{aligned}$$

and then

$$\int_0^t \max_y \theta^{(2q+3)/2} ds \le C + C\|\theta^{(2q-3-2r)/2}\theta_y^2\|.$$

For $r = 1$, using Lemma 3.1, we obtain (5.7) since

$$\int_0^t \max_y \theta^{(2q+3)/2} ds \le C + C\|\theta^{(2q-5)/2}\theta_y^2\| \le C + C\left\|\kappa\vartheta^2\theta_y^2\right\| \le C.$$

For $0 < r < 1$, define $G(v,\theta) = \int_0^\theta \xi^{-r} e_\theta(v,\xi)d\xi$. Then, from (2.6)–(2.10), Lemma 3.1, and (3.5),

$$\|G\| \le C\|\theta + \theta^{1-r}\| \le C\,\|\theta + 1\| \le C$$

and

$$G_t + (1-r)u_y\int_0^\theta \xi^{-r} p_\theta(v,\xi)d\xi = \frac{\lambda u_y^2 + \mu|\mathbf{w}_y|^2 + \nu|\mathbf{b}_y|^2}{v\theta^r} + \frac{r\kappa\theta_y^2}{v\theta^{r+1}} + \left(\frac{\kappa\theta_y}{v\theta^r}\right)_y.$$

Integrating the above equation and using (2.6)–(2.10) and Lemma 3.1, we have

$$\begin{aligned}
&\left\|\rho\vartheta^r\left(\lambda u_y^2 + \mu|\mathbf{w}_y|^2 + \nu|\mathbf{b}_y|^2\right) + r\kappa\rho\theta_y^2\vartheta^{r+1}\right\| \\
&\quad\le C + C\left\|u_y^2 + (1+\theta)^2\right\| \le C + C\left\|\kappa\vartheta^2\theta_y^2\right\| \le C.
\end{aligned}$$

Then, from the above estimate, we have

$$\int_0^t \max_y \theta^{(2q+3)/2} ds \le C + C\|\theta^{(2q-3-2r)/2}\theta_y^2\| \le C + C\left\|\kappa\theta_y^2\vartheta^{1+r}\right\| \le C.$$

If $r = 0$, replace the above definition of G by $G(v,\theta) = \int_0^\theta \xi^{-1/2} e_\theta(v,\xi)d\xi$ and follow the same procedure to obtain

$$\|2\rho\vartheta^{1/2}\left(\lambda u_y^2 + \mu|\mathbf{w}_y|^2 + \nu|\mathbf{b}_y|^2\right) + \kappa\rho\theta_y^2\vartheta^{3/2}\| \le C.$$

Then

$$\int_0^t \max_y \theta^{(2q+3)/2} ds \le C + C\|\theta^{(2q-3)/2}\theta_y^2\| \le C + C\|\kappa\theta_y^2\vartheta^{3/2}\| \le C.$$

Step 3. Now we estimate Z. Differentiate (2.2b) with respect to t, multiply it by u_t, and then integrate to obtain

$$\frac{1}{2}\left\|u_t^2\right\| + \iint_{\Pi_t}\left(\frac{\lambda}{v}u_{yt} + \left(\frac{\lambda}{v}\right)_v u_y^2 - p_t - \mathbf{b}\cdot\mathbf{b}_t\right)u_{yt}\,dy\,ds \le C,$$

from integration by parts and the initial-boundary condition (2.3). Using Lemmas 3.1 and 4.1 and (5.3), one has the following estimates:

$$\iint_{\Pi_t}\frac{\lambda}{v}u_{yt}^2\,dy\,ds \ge C_7\left\|u_{yt}\right\|_2^2,$$

$$\iint_{\Pi_t}\left(\frac{\lambda}{v}\right)_v u_y^2 u_{yt}\,dy\,ds \le \frac{C_7}{8}\left\|u_{yt}\right\|_2^2 + C\left\|u_y\right\|_4^4 \le \frac{C_7}{8}\left\|u_{yt}\right\|_2^2 + C,$$

and

$$\begin{aligned}
&\iint_{\Pi_t} p_t u_{yt}\,dy\,ds \le \left\|\left(p_v u_y + p_\theta\theta_t\right)u_{yt}\right\| \\
&\le \frac{C_7}{8}\left\|u_{yt}\right\|_2^2 + C\left(1+\Theta^{2+2r}\right)\left\|u_y\right\|_2^2 + C\left\|\left(1+\theta^{2r}\right)\theta_t^2\right\| \\
&\le \frac{C_7}{8}\left\|u_{yt}\right\|_2^2 + C(X+Y+1) \le \frac{C_7}{8}\left\|u_{yt}\right\|_2^2 + C + CZ^{\beta_4}.
\end{aligned}$$

From (4.4),

$$|\mathbf{b}_t|^2 \le M\left(|\mathbf{b}|^2 u_y^2 + |\mathbf{w}_y|^2 + |\mathbf{b}_{yy}|^2 + |\mathbf{b}_y|^2 v_y^2\right),$$

and then, by Lemmas 3.1 and 4.1,

$$\begin{aligned}
\left\|\mathbf{b}_t\right\|_2^2 &\le C\left\||\mathbf{b}|^2 u_y^2 + |\mathbf{w}_y|^2 + |\mathbf{b}_{yy}|^2 + |\mathbf{b}_y|^2 v_y^2\right\| \\
&\le C\max_t\left\|\mathbf{b}_y\right\|_2^2\left\|u_y\right\|_2^2 + C\left\|(\mathbf{w}_y,\mathbf{b}_{yy})\right\|_2^2 + C\int_0^t \max_y|\mathbf{b}_y|^2\left\|v_y\right\|_2^2 ds \qquad (5.9)\\
&\le C + C\left\|(\mathbf{b}_y,\mathbf{b}_{yy})\right\|_2^2 \le C.
\end{aligned}$$

Thus, we have

$$\iint_{\Pi_t} u_{yt}\mathbf{b}\cdot\mathbf{b}_t\,dy\,ds \le \frac{C_7}{8}\left\|u_{yt}\right\|_2^2 + C\left\|\mathbf{b}_t\right\|_2^2 \le \frac{C_7}{8}\left\|u_{yt}\right\|_2^2 + C,$$

and then

$$\|u_t\|_2^2 + \|\!|u_{yt}|\!\|_2^2 \le C + CZ^{\beta_4}.$$

From (2.2b),

$$u_{yy} = \frac{v}{\lambda}\left(u_t - p_y - \mathbf{b}\cdot\mathbf{b}_y - \left(\frac{\lambda}{v}\right)_v v_y u_y\right),$$

and then, by Lemma 4.1 and (5.2),

$$\begin{aligned}\|u_{yy}\|_2^2 &\le \left\|u_t^2 + p_v^2 v_y^2 + p_\theta^2\theta_y^2 + |\mathbf{b}|^2 + |\mathbf{b}_y|^2 + v_y^2 u_y^2\right\| \\ &\le C + CZ^{\beta_4} + C\Theta^{2+2r}\|v_y\|_2^2 + CY + C\max_{y,t} u_y^2 \|v_y\|_2^2 \le C + CZ^{1/2} + CZ^{\beta_4},\end{aligned}$$

that is, $Z \le C + CZ^{1/2} + CZ^{\beta_4}$. Therefore $Z \le C$ since the exponents of Z on the right-hand side of the above inequality are both less than one. Then (5.8) implies $X + Y \le C$.

Step 4. We now prove the positive lower bound of the temperature with the aid of a maximum principle. It follows from (5.2) that $|u_y| \le C$. Differentiate (2.2c) with respect to t and then multiply by $\mathbf{w}_t$ and integrate it to obtain

$$\begin{aligned}\frac{1}{2}\|\mathbf{w}_t\|_2^2 + \|\!|\mu\rho|\mathbf{w}_{yt}|^2|\!\| &\le C - \iint_{\Pi_t}\left(\frac{\mu}{v}\right)_v v_t \mathbf{w}_y\cdot\mathbf{w}_{yt}\,dy\,ds - \iint_{\Pi_t}\mathbf{b}_t\cdot\mathbf{w}_{yt}\,dy\,ds \\ &\le \frac{1}{2}\|\!|\mu\rho|\mathbf{w}_{yt}|^2|\!\| + C\|\!|u_y^4 + |\mathbf{w}_y|^4 + |\mathbf{b}_t|^2|\!\| \le \frac{1}{2}\|\!|\mu\rho|\mathbf{w}_{yt}|^2|\!\| + C,\end{aligned}$$

from integration by parts, the initial-boundary condition (2.3), Lemma 4.1, and (5.9). Then

$$\|\mathbf{w}_t\|_2^2 + \|\!|\mu\rho|\mathbf{w}_{yt}|^2|\!\| \le C.$$

From (2.2c),

$$\mathbf{w}_{yy} = \frac{v}{\mu}\left(\mathbf{w}_t - \mathbf{b}_y - \left(\frac{\mu}{v}\right)_v v_y \mathbf{w}_y\right),$$

and then

$$\begin{aligned}\|\mathbf{w}_{yy}\|_2^2 &\le C\left\||\mathbf{w}_t|^2 + |\mathbf{b}_y|^2 + v_y^2|\mathbf{w}_y|^2\right\| \le C + C\max_y|\mathbf{w}_y|^2\|v_y\|_2^2 \\ &\le C + C\|\mathbf{w}_y\|_2^2 + \frac{1}{2}\|\mathbf{w}_{yy}\|_2^2,\end{aligned}$$

which implies that

$$\|\mathbf{w}_{yy}\|_2^2 \le C + C\|\mathbf{w}_y\|_2^2 \le C$$

and

$$|\mathbf{w}_y|^2 \le C\|\mathbf{w}_y\|_2^2 + C\|\mathbf{w}_{yy}\|_2^2 \le C.$$

Rewrite (5.4) as

$$e_\theta\theta_t + \theta p_\theta u_y - \frac{\lambda}{v}u_y^2 - \frac{\mu}{v}|\mathbf{w}_y|^2 - \frac{\nu}{v}|\mathbf{b}_y|^2 = \left(\frac{\kappa}{v}\right)_v v_y\theta_y + \frac{\kappa_\theta}{v}\theta_y^2 + \frac{\kappa}{v}\theta_{yy}.$$

Then, by Lemma 4.1, the above estimates, and an interpolation inequality on θ_y^2,

$$\begin{aligned}\|\kappa\rho\theta_{yy}\|_2^2 &\le C\left\|\theta_t^2+u_y^2+u_y^4+|\mathbf{w}_y|^4+|\mathbf{b}_y|^4+v_y^2\theta_y^2+\theta_y^4\right\| \\ &\le C+C\int_0^t \max_y \theta_y^2\left\|v_y^2+\theta_y^2\right\| ds \le C+C\|\theta_y\|_2^2+\frac{1}{2}\|\kappa\rho\theta_{yy}\|_2^2,\end{aligned}$$

which implies

$$\|\kappa\rho\theta_{yy}\|_2^2 \le C+C\|\theta_y\|_2^2 \le C. \tag{5.10}$$

From (3.3),

$$V(v(y,t))_y = V(v(y,0))_y + u(y,t) - u_0(y) + \int_0^t (p_y + \mathbf{b}\cdot\mathbf{b}_y)dy.$$

Then, using the interpolation inequalities on $|\mathbf{b}_y|^2$ and θ_y^2, Lemmas 3.1 and 4.1, and (5.10), we have

$$\begin{aligned}v_y^2 &\le C(V(v)_y)^2 \le C+C\int_0^t \left(|\mathbf{b}_y|^2+p_v^2v_y^2+p_\theta^2\theta_y^2\right) ds \\ &\le C+C\|(\mathbf{b}_y,\mathbf{b}_{yy},\theta_y,\theta_{yy})\|_2^2+C\int_0^t v_y^2 ds \le C+C\int_0^t v_y^2 ds,\end{aligned}$$

which yields, from Gronwall's inequality, $v_y^2 \le M$. Differentiate (2.2d) with respect to t and then multiply by $(v\mathbf{b})_t$ and integrate to obtain

$$\begin{aligned}\frac{1}{2}\|(v\mathbf{b})_t\|_2^2 &\le C-\frac{1}{2}\left\|\nu\mathbf{b}_{yt}^2\right\|+C\left\||\mathbf{w}_t|^2+u_y^2|\mathbf{b}_y|^2+u_{yy}^2+v_y^2|\mathbf{b}_t|^2\right\| \\ &\le C-\frac{1}{2}\left\|\nu\mathbf{b}_{yt}^2\right\|+C\int_0^t \max_y|\mathbf{b}_y|^2\|u_y\|_2^2\,ds+C\|\mathbf{b}_t\|_2^2 \\ &\le C-\frac{1}{2}\left\|\nu\mathbf{b}_{yt}^2\right\|+C\|(\mathbf{b}_y,\mathbf{b}_{yy})\|_2^2 \le C-\frac{1}{2}\left\|\nu\mathbf{b}_{yt}^2\right\|,\end{aligned}$$

which yields

$$\|(v\mathbf{b})_t\|_2^2+\left\|\nu\mathbf{b}_{yt}^2\right\| \le C.$$

From (2.2d),

$$\||\mathbf{b}_{yy}|\|_2^2 \le C\left\|(v\mathbf{b})_t^2+|\mathbf{w}_y|^2+v_y^2|\mathbf{b}_y|^2\right\| \le C+C\|\mathbf{b}_y\|_2^2 \le C,$$

and then

$$|\mathbf{b}_y|^2 \le C\|(\mathbf{b}_y,\mathbf{b}_{yy})\|_2^2 \le C.$$

As a consequence, we have $C^{-1} \le \theta \le C$ from the maximum principle (see [22]) applied to (5.4) and the boundedness of $(v_y, u_y, \mathbf{w}_y, \mathbf{b}_y)$, $0<\theta\le C$ (from (5.3)), and the positive lower bound of θ_0.

This completes the proof of the main theorem. ∎

Acknowledgments. The author would like to thank Reza Malek-Madani for many stimulating conversations and discussions. The author also thanks the referees for their comments and suggestions.

REFERENCES

[1] R. Balescu, *Transport Processes in Plasmas* I: *Classical Transport Theory*, North–Holland, New York, 1988.
[2] E. Becker, *Gasdynamik*, Teubner, Stuttgart, 1966.
[3] M. Brio and C. C. Wu, *An upwind differencing scheme for the equations of magnetohydrodynamics*, J. Comput. Phys., 75 (1988), pp. 400–422.
[4] H. Cabannes, *Theoretical Magnetofluiddynamics*, Academic Press, New York, 1970.
[5] G.-Q. Chen and D. Wang, *Global solution of nonlinear magnetohydrodynamics with large initial data*, J. Differential Equations, 182 (2002), pp. 344–376.
[6] P. C. Clemmow and J. P. Dougherty, *Electrodynamics of Particles and Plasmas*, Addison-Wesley, New York, 1990.
[7] C. M. Dafermos, *Global smooth solutions to the initial-boundary value problem for the equations of one-dimensional nonlinear thermoviscoelasticity*, SIAM J. Math. Anal., 13 (1982), pp. 397–408.
[8] H. Freistühler and P. Szmolyan, *Existence and bifurcation of viscous profiles for all intermediate magnetohydrodynamic shock waves*, SIAM J. Math. Anal., 26 (1995), pp. 112–128.
[9] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer, Berlin, 1994.
[10] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Springer-Verlag, New York, 1981.
[11] Y. I. Kanel, *On a model system of equations of one-dimensional gas motion*, Differential Equations, 4 (1968), pp. 374–380.
[12] S. Kawashima and M. Okada, *Smooth global solutions for the one-dimensional equations in magnetohydrodynamics*, Proc. Japan Acad. Ser. A Math. Sci., 58 (1982), pp. 384–387.
[13] S. Kawashima and Y. Shizuta, *Magnetohydrodynamic approximation of the complete equations for an electromagnetic fluid*, Tsukuba J. Math., 10 (1986), pp. 131–149.
[14] B. Kawohl, *Global existence of large solutions to initial boundary value problems for a viscous, heat-conducting, one-dimensional real gas*, J. Differential Equations, 58 (1985), pp. 76–103.
[15] V. Kazhikhov and V. V. Shelukhin, *Unique global solution with respect to time of initial-boundary-value problems for one-dimensional equations of a viscous gas*, J. Appl. Math. Mech., 41 (1977), pp. 273–282.
[16] A. G. Kulikovskiy and G. A. Lyubimov, *Magnetohydrodynamics*, Addison-Wesley, Reading, MA, 1965.
[17] O. A. Ladyženskaja, V. A. Solonnikov, and N. N. Ural'ceva, *Linear and Quasilinear Equations of Parabolic Type*, Transl. Math. Monogr. 23, AMS, Providence, RI, 1988.
[18] L. D. Laudau and E. M. Lifshitz, *Electrodynamics of Continuous Media*, 2nd ed., Pergamon, New York, 1984.
[19] R. J. LeVeque, D. Mihalas, E. Dorfi, and E. Müller, *Computational Methods in Astrophysical Fluid Flow*, Twenty-seventh Saas-Fee Advances Course 27, Springer-Verlag, New York, 1998.
[20] T.-P. Liu and Y. Zeng, *Large time behavior of solutions for general quasilinear hyperbolic-parabolic systems of conservation laws*, Mem. Amer. Math. Soc. 125 (1997), no. 599.
[21] J. Nash, *Le problème de Cauchy pour les équations différentielles d'un fluide général*, Bull. Soc. Math. France, 90 (1962), pp. 487–497.
[22] M. H. Protter and H. F. Weinberger, *Maximum Principles in Differential Equations*, Prentice–Hall, Englewood Cliffs, NJ, 1967.
[23] L. C. Woods, *Principles of Magnetoplasma Dynamics*, Oxford University Press, New York, 1987.
[24] C. C. Wu, *Formation, structure, and stability of MHD intermediate shocks*, J. Geophys. Res., 95 (1990), pp. 8149–8175.
[25] Y. B. Zel'dovich and Y. P. Raizer, *Physics of Shock Waves and High-Temperature Hydrodynamic Phenomena*, Vol. 2, Academic Press, New York, 1967.

SIAM J. APPL. MATH.
Vol. 63, No. 4, pp. 1442–1460

THE WIENER–HOPF TECHNIQUE FOR IMPENETRABLE WEDGES HAVING ARBITRARY APERTURE ANGLE*

V. G. DANIELE†

Abstract. Diffraction by impenetrable wedges having arbitrary aperture angle is studied by means of the Wiener–Hopf (W-H) technique. A system of functional equations called generalized Wiener–Hopf equations (GWHE) is obtained. Only for certain values of the aperture angle are these equations recognizable as standard or classical Wiener–Hopf equations (CWHE). However, in all cases a mapping is found that reduces the GWHE to CWHE. It means that the diffraction by an impenetrable wedge always reduces to a standard W-H factorization. The solution for the diffraction by a wedge with given face impedances (the Malyuzhinets problem) is obtained in closed form by an explicit factorization of the kernel.

Key words. Wiener–Hopf technique, diffraction, scattering, wedge, half-plane

AMS subject classifications. 78A45, 47A68, 35J25, 45E10, 47B35

PII. S0036139901400239

1. Introduction.

The first rigorous studies of waves in the presence of geometric discontinuities are due to Poincaré [1] and Sommerfeld [2], who considered the presence of a half-plane in free space. The generalization of the half-plane problem led to the wedge problem, which constituted an important and challenging subject of applied mathematics in the last century. There are a vast number of papers that address the wedge problem. They concern many disciplines such as electromagnetism, acoustics, hydrodynamics, fracture mechanics, and so on. The impenetrable wedges arise from the introduction of approximate boundary conditions on the surfaces of the wedges. These conditions considerably simplify the study of wedge problems. In fact, the external problem (i.e., the evaluation of the fields outside the wedge) is decoupled from the internal problem (i.e., the evaluation of the fields inside the wedge). This article deals only with the impenetrable case.

There are many analytical methods for studying fields and waves in angular regions. Among them, the Malyuzhinets method [3] is particularly well known. This method is based on the use of the Sommerfeld integral. The Wiener–Hopf (W-H) technique is the most powerful method for solving field problems in the presence of geometrical discontinuities. However, concerning wedge problems there is the belief that this technique can be applied only for certain values of the aperture angle. We have never been satisfied with this limited use of the W-H technique for the wedge problems. The aim of this work is to show that the W-H technique can handle wedge problems also in the presence of arbitrary aperture angle. This proposed task is not easy. For instance, the W-H formulation of wedge problems yields functional equations (generalized W-H equations) that substantially differ from the well-known classical W-H equations studied in the literature [4]. However, in this paper it will be shown that, in a suitable complex plane, the generalized W-H equations (GWHE) always

*Received by the editors December 28, 2001; accepted for publication (in revised form) January 6, 2003; published electronically May 29, 2003. Part of this paper appeared in preliminary form in the Proceedings of the 2001 International Conference on Electromagnetics in Advanced Applications (ICEAA01), 2001, Torino, Italy, pp. 385–393. This work was supported by the Italian Ministry of University and Scientific Research (MIUR) under grant MM093227718.

http://www.siam.org/journals/siap/63-4/40023.html

†Dipartimento di Elettronica, Politecnico di Torino, C.so Duca degli Abruzzi 24, 10129 Torino, Italy (daniele@polito.it).

reduce to classical W-H equations (CWHE). That means that the diffraction by an impenetrable wedge always reduces to a standard W-H factorization. For instance, the solution for the diffraction by a wedge with given face impedances (Malyuzhinets problem) is obtained in closed form in this paper by an explicit factorization of the matrix kernel. This yields a new solution of this important problem in a form completely different from, although equivalent to, the Malyuzhinets solution.

At present, important problems involving angular regions are concerned with either penetrable wedges or impenetrable wedges surrounded by media in which more types of waves propagate. In general, for these problems closed form solutions are not available. However, some recent progress has resulted in efficient approximate solutions of important cases [5, 6, 7, 8, 9, 10]. The techniques used in these works are substantially based on the regularization method for singular integral equations. The regularization method for singular integral equations is well known and used by many authors to obtain approximate solutions of singular integral equations when closed form solutions are not obtainable by other methods. The regularization method consists of reducing the singular integral equation to a Fredholm equation of the second kind. Such a technique is very convenient since Fredholm equations are amenable to efficient numerical integration schemes. Every year several papers are produced to reduce particular wedge problems to the solution of Fredholm equations. The most important of them differs only in the technique used to obtain the regularization of the kernels. For example, some works by Budaev [5] and Budaev and Bogy [6, 8] make use of the Sommerfeld–Malyuzhinets method. Conversely, in the work of Gautesen [10], the reduction of the singular equation to a Fredholm equation is obtained in part by using the W-H technique through a decomposition of functions. Even though the involved method is, of course, approximate, Gautesen's paper is very interesting since the kind of regularization performed produces very nice results, though on only one well-defined wedge problem.

It is worth noticing that the extension of the W-H technique reported in this paper to deal with wedge problems is not something that has been done merely for academic reasons, but could have important applications to obtaining the solution of yet-unsolved wedge problems by introducing approximate factorizations of the kernels [11], in contrast to methods based on the regularization of the operators.

This paper is organized as follows: Section 2 reports some important functional equations deduced in Appendix A. After indicating the geometry of the problem, the boundary conditions are considered in section 3. Their imposition yields the GWHE for impenetrable wedge problems. Section 4 describes some particular cases in which the GWHE are immediately recognizable as CWHE. Section 5 deals with the general case for which a fundamental mapping to a new complex plane is provided. This mapping reduces the GWHE to the CWHE. Section 6 concerns the diffraction by a perfectly electrical conducting (PEC) wedge; this problem involves scalar W-H equations. Section 7 describes the diffraction by a wedge with given face impedances (the Malyuzhinets problem). This problem involves vector W-H equations. Section 8 presents some physical aspects of the W-H solution.

2. Some functional equations occurring in angular regions. We consider only time-harmonic electromagnetic fields with a time dependence specified by the factor $e^{j\omega t}$, which is omitted. These electromagnetic fields will be studied in the angular region indicated in Figure 1 that is defined by the aperture angle γ_1, $(0 \leq \varphi \leq \gamma_1)$. This region is filled by an isotropic homogeneous medium having complex permittivity ε and complex permeability μ. Without loss of generality, we assume

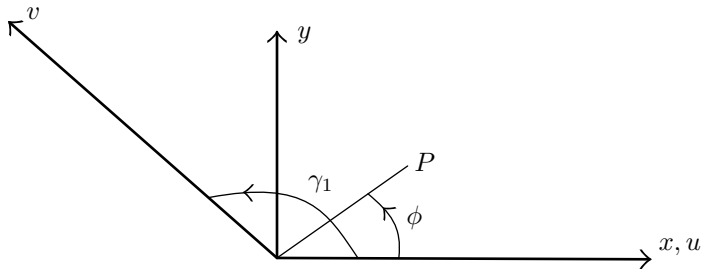


FIG. 1. *Angular region* $0 \le \phi \le \gamma_1$.

the z dependence of the electromagnetic field $\mathbf{E}$ and $\mathbf{H}$ specified by the factor $e^{-j\alpha_o z}$, which is omitted. According to the circumstances, Cartesian coordinates $\{x, y, z\}$, polar coordinates $\{\rho, \varphi, z\}$, or oblique coordinates $\{u, v, z\}$ will be used.

The W-H technique [4] for wedge problems is based on the introduction of the following Laplace transforms:

$$V_{z+}(\eta,\varphi) = \int_0^\infty E_z(\rho,\varphi)e^{j\eta\rho}d\rho, \qquad I_{z+}(\eta,\varphi) = \int_0^\infty H_z(\rho,\varphi)e^{j\eta\rho}d\rho, \tag{1}$$

$$V_{\rho+}(\eta,\varphi) = \int_0^\infty E_\rho(\rho,\varphi)e^{j\eta\rho}d\rho, \qquad I_{\rho+}(\eta,\varphi) = \int_0^\infty H_\rho(\rho,\varphi)e^{j\eta\rho}d\rho, \tag{2}$$

where the subscript + indicates plus functions, i.e., functions having regular half-planes of convergence that are upper half-planes in the η-plane.

Notice that even if E_ρ and H_ρ may be singular for vanishing values of ρ, there are no problems of existence of the Laplace transforms in physical problems. To avoid singular points on the real axis of the η-plane, we assume small losses also in the presence of no dissipative media. Consequently, the propagation constant (or wave number) defined by $k = \omega\sqrt{\mu\varepsilon}$ always has negative imaginary part $\mathrm{Im}[k] < 0$.

As will be shown in Appendix A, the following equations hold for $0 \le \gamma_1 \le \pi$:

$$\begin{aligned} &\xi V_{z+}(\eta,0) - \frac{\tau_o^2}{\omega\varepsilon} I_{\rho+}(\eta,0) - \frac{\alpha_o\eta}{\omega\varepsilon} I_{z+}(\eta,0) \\ &\qquad = -n_1 V_{z+}(-m_1,\gamma_1) - \frac{\tau_o^2}{\omega\varepsilon} I_{\rho+}(-m_1,\gamma_1) + \frac{\alpha_o m_1}{\omega\varepsilon} I_{z+}(-m_1,\gamma_1), \end{aligned} \tag{3}$$

$$\begin{aligned} &\xi I_{z+}(\eta,0) + \frac{\tau_o^2}{\omega\mu} V_{\rho+}(-\eta,0) + \frac{\alpha_o\eta}{\omega\mu} V_{z+}(\eta,0) \\ &\qquad = -n_1 I_{z+}(-m_1,\gamma_1) + \frac{\tau_o^2}{\omega\mu} V_{\rho+}(-m_1,\gamma_1) - \frac{\alpha_o m_1}{\omega\mu} V_{z+}(-m_1,\gamma_1), \end{aligned} \tag{4}$$

with $\tau_o = \sqrt{k^2 - \alpha_o^2}$, $\mathrm{Im}[\tau_o] \le 0$, $\xi = \xi(\eta) = \sqrt{\tau_o^2 - \eta^2}$ with the branch $\xi(0) = \tau_o$, and with m_1 and n_1 the functions $m = m(\eta)$ and $n = n(\eta)$ defined by $m = m(\eta) = -\eta\cos\gamma + \xi\sin\gamma$ and $n = n(\eta) = -\xi\cos\gamma - \eta\sin\gamma$ evaluated for $\gamma = \gamma_1$. These functional equations relate the Laplace transforms of the tangential components of $\mathbf{E}$ and $\mathbf{H}$ on the two boundaries $\varphi = 0$ and $\varphi = \gamma_1$ of the considered angular region. They are fundamental to deducing the W-H equations for wedge-shaped regions in the next section.

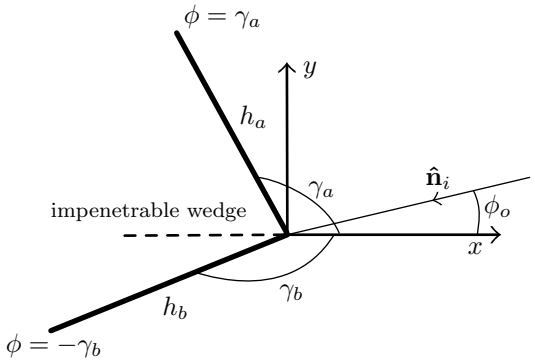


FIG. 2. *Scattering by an impenetrable wedge.*

When the angular region is defined for $-\gamma_2 \leq \varphi \leq 0$ ($0 \leq \gamma_2 \leq \pi$), a slight modification of the deduction reported in Appendix A yields similar equations:

$$\begin{aligned}(5)\quad & \xi V_{z+}(\eta,0) + \frac{\tau_o^2}{\omega\varepsilon} I_{\rho+}(\eta,0) + \frac{\alpha_o \eta}{\omega\varepsilon} I_{z+}(\eta,0) \\ & \quad = -n_2 V_{z+}(-m_2,-\gamma_2) + \frac{\tau_o^2}{\omega\varepsilon} I_{\rho+}(-m_2,-\gamma_2) - \frac{\alpha_o m_2}{\omega\varepsilon} I_{z+}(-m_2,-\gamma_2),\end{aligned}$$

$$\begin{aligned}(6)\quad & \xi I_{z+}(\eta,0) - \frac{\tau_o^2}{\omega\mu} V_{\rho+}(-\eta,0) - \frac{\alpha_o \eta}{\omega\mu} V_{z+}(\eta,0) \\ & \quad = -n_2 I_{z+}(-m_2,-\gamma_2) - \frac{\tau_o^2}{\omega\mu} V_{\rho+}(-m_2,-\gamma_2) + \frac{\alpha_o m_2}{\omega\mu} V_{z+}(-m_2,-\gamma_2),\end{aligned}$$

where m_2 and n_2 are the functions $m = m(\eta)$ and $n = n(\eta)$ defined by $m = m(\eta) = -\eta\cos\gamma + \xi\sin\gamma$ and $n = n(\eta) = -\xi\cos\gamma - \eta\sin\gamma$ evaluated for $\gamma = \gamma_2$. It is worth observing that (3)–(6) also hold in the presence of plane waves.

3. Generalized W-H equations for wedge-shaped regions. Figure 2 shows the problem that we will study. A plane wave with skew incidence $\hat{\mathbf{n}}_i$ excites an impenetrable wedge where linear conditions are defined on the boundaries $\varphi = \gamma_a$ and $\varphi = -\gamma_b$. The longitudinal field components of the plane wave are given by

$$(7)\qquad E_z^i = E_o e^{j\tau_o \rho\cos(\varphi-\varphi_o)} e^{-j\alpha_o z}, \qquad H_z^i = H_o e^{j\tau_o \rho\cos(\varphi-\varphi_o)} e^{-j\alpha_o z},$$

where E_o and H_o are known quantities, ϑ_o is the angle between the $\hat{\mathbf{n}}_i$ and $\hat{\mathbf{z}}$, $\alpha_o = k\cos\vartheta_o$, and $\tau_o = k\sin\vartheta_o$.

The Leontovich conditions on the boundaries of the wedge $\varphi = \gamma_a$ and $\varphi = -\gamma_b$ are expressed by

$$(8)\qquad \begin{bmatrix} E_z(\rho,\gamma_a) \\ H_z(\rho,\gamma_a) \end{bmatrix} = h_a \begin{bmatrix} H_\rho(\rho,\gamma_a) \\ -E_\rho(\rho,\gamma_a) \end{bmatrix}, \qquad \begin{bmatrix} E_z(\rho,-\gamma_b) \\ H_z(\rho,-\gamma_b) \end{bmatrix} = -h_b \begin{bmatrix} H_\rho(\rho,-\gamma_b) \\ -E_\rho(\rho,-\gamma_b) \end{bmatrix},$$

with the matrices

$$h_{a,b} = \begin{bmatrix} Z_e^{a,b} & T_e^{a,b} \\ T_h^{a,b} & Y_h^{a,b} \end{bmatrix}$$

depending on the wedge material.

The Maxwell equations provide the following relations between the radial components E_ρ and H_ρ and the longitudinal ones E_z and H_z:

$$\rho E_\rho = \frac{1}{j\tau_o^2}\left[\alpha_o\rho\frac{\partial E_z}{\partial\rho} + \omega\mu\frac{\partial H_z}{\partial\varphi}\right], \qquad \rho H_\rho = \frac{1}{j\tau_o^2}\left[\alpha_0\rho\frac{\partial H_z}{\partial\rho} - \omega\varepsilon\frac{\partial E_z}{\partial\varphi}\right]. \tag{9}$$

The mathematical problem to solve consists of finding the solutions E_z and H_z of the wave equations

$$\frac{\partial^2}{\partial x^2}E_z + \frac{\partial^2}{\partial y^2}E_z + \tau_o^2 E_z = 0, \qquad \frac{\partial^2}{\partial x^2}H_z + \frac{\partial^2}{\partial y^2}H_z + \tau_o^2 H_z = 0 \tag{10}$$

such that (9) and (8) hold on the boundaries $\varphi = \gamma_a$ and $\varphi = -\gamma_b$ of the wedge, and $E_z - E_z^i$ and $H_z - H_z^i$ satisfy the Sommerfeld radiation conditions in the angular region $-\gamma_b \le \varphi \le \gamma_a$. For physical problems the materials that constitute the wedge yield parameters of $h_{a,b}$ that ensure the existence and the uniqueness of the solution of this problem.

Equations (3)–(6) and the Leontovich conditions (8) written in the Laplace domain yield the following four equations:

$$\begin{aligned}\xi V_{z+}(\eta,0) - \frac{\tau_o^2}{\omega\varepsilon}I_{\rho+}(\eta,0) - \frac{\alpha_o\eta}{\omega\varepsilon}I_{z+}(\eta,0)\\ = (a_1\eta + b_1\xi + e_1)I_{\rho+}(-m_a,\gamma_a) + (c_1\eta + d_1\xi + f_1)V_{\rho+}(-m_a,\gamma_a),\end{aligned} \tag{11}$$

$$\begin{aligned}\xi V_{z+}(\eta,0) + \frac{\tau_o^2}{\omega\varepsilon}I_{\rho+}(\eta,0) + \frac{\alpha_o\eta}{\omega\varepsilon}I_{z+}(\eta,0)\\ = (a_2\eta + b_2\xi + e_2)I_{\rho+}(-m_b,-\gamma_b) + (c_2\eta + d_2\xi + f_2)V_{\rho+}(-m_b,-\gamma_b),\end{aligned} \tag{12}$$

$$\begin{aligned}\xi I_{z+}(\eta,0) + \frac{\tau_o^2}{\omega\mu}V_{\rho+}(-\eta,0) + \frac{\alpha_o\eta}{\omega\mu}V_{z+}(\eta,0)\\ = (a_3\eta + b_3\xi + e_3)I_{\rho+}(-m_a,\gamma_a) + (c_3\eta + d_3\xi + f_3)V_{\rho+}(-m_a,\gamma_a),\end{aligned} \tag{13}$$

$$\begin{aligned}\xi I_{z+}(\eta,0) - \frac{\tau_o^2}{\omega\mu}V_{\rho+}(-\eta,0) - \frac{\alpha_o\eta}{\omega\mu}V_{z+}(\eta,0)\\ = (a_4\eta + b_4\xi + e_4)I_{\rho+}(-m_b,-\gamma_b) + (c_4\eta + d_4\xi + f_4)V_{\rho+}(-m_b,-\gamma_b),\end{aligned} \tag{14}$$

with the constants a_i, b_i, c_i, d_i, e_i, f_i $(i = 1,2,3,4)$ depending on the geometrical and electromagnetic parameters involved in the problem being considered. They are not specified here.

Notice that, besides the four functions $V_{z+}(\eta,0)$, $I_{z+}(\eta,0)$, $V_{\rho+}(\eta,0)$, $I_{\rho+}(\eta,0)$ that are plus functions in the η-plane, there are four other functions which are regular in m_a-lower half-planes and in m_b-lower half-planes. Equations (11)–(14) will be called GWHE.

4. The cases involving classical W-H equations. In some important cases, (11)–(14) simplify considerably and constitute a vector system of classical W-H equations. These cases are indicated in Figure 3, and they will be discussed in some detail.

(a) *The problem of the half-plane.* The half-plane problem (Figure 3(a)) has been studied extensively in the past [12]. The most recent solution by means of the W-H technique [13] concerns the incidence of a skew plane wave on a wedge where the matrices h_a and h_b assume the form

$$h_{a,b} = \begin{bmatrix} Z^{a,b} & 0 \\ 0 & \dfrac{1}{Z^{a,b}} \end{bmatrix}. \tag{15}$$

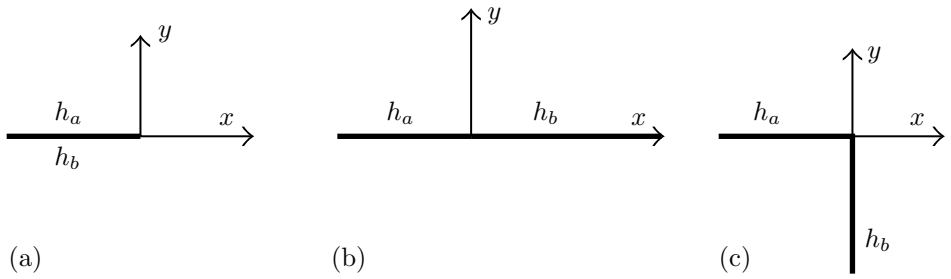


FIG. 3. (a) *Half-plane:* $\gamma_a = \gamma_b = \pi$; (b) *half-plane junction:* $\gamma_a = \pi$, $\gamma_b = 0$; (c) *right wedge:* $\gamma_a = \pi$, $\gamma_b = \pi/2$.

The problem of diffraction by an anisotropic impedance half-plane has been addressed by Hurd and Lüneburg [14] and by Senior and Legault [15].

In the case of the half-plane $\gamma_a = \gamma_b = \pi$, the functions m_a and m_b simplify considerably:

$$m_{a,b} = m_{a,b}(\eta) = -\eta \cos \gamma_{a,b} + \xi \sin \gamma_{a,b} = \eta. \tag{16}$$

After algebraic manipulation, the GWHEs (11)–(14) reduce to the classical W-H system

$$G_\pi(\eta) X_+(\eta) = Y_+(-\eta), \tag{17}$$

where

$$X_+(\eta) = \begin{bmatrix} V_{z+}(\eta, 0) \\ I_{\rho+}(\eta, 0) \\ I_{z+}(\eta, 0) \\ V_{\rho+}(\eta, 0) \end{bmatrix}, \qquad Y_+(-\eta) = \begin{bmatrix} I_{\rho+}(-\eta, \pi) \\ I_{\rho+}(-\eta, -\pi) \\ V_{\rho+}(-\eta, \pi) \\ V_{\rho+}(-\eta, -\pi) \end{bmatrix},$$

$$G_\pi(\eta) = R_0(\eta) + \xi R_1(\eta). \tag{18}$$

The function $Y_+(-\eta)$ is a minus function. These functions are regular in a lower half-plane in the η-plane. They will be denoted with the subscript $-$ as in $Y_+(-\eta) = X_-(\eta)$. Equation (18) defines the kernel matrix $G_\pi(\eta)$ through the two matrices $R_0(\eta)$ and $R_1(\eta)$. It should be observed that $R_0(\eta)$ and $R_1(\eta)$ are rational matrices of η. In general, they have complicated expressions that are not reported here.

The reason for the possibility of closed form solutions for the half-problem follows from the structure of the kernel $G_\pi(\eta) = G(\eta)$, which allows explicit factorizations. For instance, since we can explicitly factorize the rational matrix $R_0(\eta) = R_{0-}(\eta) R_{0+}(\eta)$, it is convenient to put $G(\eta)$ in the form

$$G(\eta) = R_{0-}(\eta) \Big[1 + \xi R_{0-}(\eta)^{-1} R_1(\eta) R_{0+}(\eta)^{-1} \Big] R_{0+}(\eta). \tag{19}$$

Taking into account that

$$R_{0-}(\eta)^{-1} R_1(\eta) R_{0+}(\eta)^{-1} = \frac{P(\eta)}{d(\eta)}, \tag{20}$$

where $P(\eta)$ and $d(\eta)$ are, respectively, matrix and scalar polynomial functions, we reduce the factorization of the matrix $G(\eta)$ to that of the matrix $1 + \frac{\sqrt{k^2-\eta^2}}{d(\eta)} P(\eta)$. This matrix does commute with the polynomial matrix $P(\eta)$ and can be explicitly factorized by using the procedure indicated in [16, 17, 20]. However, it must be observed that the more general cases are very cumbersome to deal with, and they still require considerable mathematical skill. This topic is beyond the scope of this work and will not be further pursued here.

(b) *The problem of the two half-plane junction.* The problem of the two half-plane junction (Figure 3(b)) has been studied extensively in the past [18]. To this author's knowledge, the more recent W-H solution is reported in [19]. For this problem, $\gamma_a = \pi \Rightarrow m = \eta$, and we can ignore (5) and (6). By imposing the boundary equation (8) on the two faces $\varphi = 0$ and $\varphi = \pi$ (Figure 3(b)), (3) and (4) yield a classical W-H equation of order two, where again the kernel has the form given in (18). Consequently, the same considerations as in the half-plane problem apply. It should be observed that, in presence of a skew incident plane wave, when h_a and h_b assume the form (15), the factorization of the kernel of order two can be obtained by slightly modifying the method indicated in [13].

(c) *The problem of the right wedge.* Even though it is well known that the right wedge (Figure 3(c)) involves classical W-H equations, these problems are usually approached with the Malyuzhinets method. The W-H formulation is based again on (11)–(14), where $\gamma_a = \pi$, $\gamma_b = \pi/2$. It yields $m_a = \eta$ and $m_b = \xi = \sqrt{k^2 - \eta^2}$. Changing η with $-\eta$ in (11)–(14) adds four independent equations that introduce the six new functions $V_{z+}(-\eta, 0)$, $I_{z+}(-\eta, 0)$, $V_{\rho+}(-\eta, 0)$, $I_{\rho+}(-\eta, 0)$, $I_{\rho+}(-\eta, \pi)$, $V_{\rho+}(-\eta, \pi)$. We now have eight equations that involve the fourteen unknowns $V_{z+}(\eta, 0)$, $I_{z+}(\eta, 0)$, $V_{\rho+}(\eta, 0)$, $I_{\rho+}(\eta, 0)$, $I_{\rho+}(\eta, \pi)$, $V_{\rho+}(\eta, \pi)$, $I_{\rho+}(-\xi, -\frac{\pi}{2})$, $V_{\rho+}(-\xi, -\frac{\pi}{2})$, $V_{z+}(-\eta, 0)$, $I_{z+}(-\eta, 0)$, $V_{\rho+}(-\eta, 0)$, $I_{\rho+}(-\eta, 0)$, $I_{\rho+}(-\eta, \pi)$, and $V_{\rho+}(-\eta, \pi)$. Two equations allow us to eliminate the two functions $I_{\rho+}(-\xi, -\frac{\pi}{2})$ and $V_{\rho+}(-\xi, -\frac{\pi}{2})$. Rearranging the other six yields a system of classical W-H equations that involve the six plus functions $V_{z+}(\eta, 0)$, $I_{z+}(\eta, 0)$, $V_{\rho+}(\eta, 0)$, $I_{\rho+}(\eta, 0)$, $I_{\rho+}(\eta, \pi)$, $V_{\rho+}(\eta, \pi)$ and the six minus functions $V_{z+}(-\eta, 0)$, $I_{z+}(-\eta, 0)$, $V_{\rho+}(-\eta, 0)$, $I_{\rho+}(-\eta, 0)$, $I_{\rho+}(-\eta, \pi)$, $V_{\rho+}(-\eta, \pi)$. Again the kernel matrix $G(\eta)$ of this system has the form (18). Consequently, the same considerations apply as in the half-plane problem.

5. The standard W-H equations of the wedge problem having arbitrary aperture angle. When the wedge has an arbitrary aperture angle, the GWHEs (11)–(14) constitute a closed mathematical problem that has been considered in [21]. By using the concept of generalized decomposition, these equations can be reduced to Fredholm systems. However, it is very remarkable that we can remain in the framework of CWHE by introducing a suitable mapping $\eta = \eta(\overline{\eta})$ that, by eliminating the m-plane, allows one to deal only with minus and plus functions of the $\overline{\eta}$-plane (see Appendix B). As will be seen later, in the $\overline{\eta}$-plane explicit W-H solutions will be provided for the problems solved by the Sommerfeld–Malyuzhinets method. In fact, for these problems the scalar decomposition can be accomplished with the classical Cauchy decomposition formula, and the matrices to factorize are of the Daniele–Khrapkov type.

With reference to Figure 4(a), equations (11)–(14) can be rearranged in the form

$$G_\Phi(\eta) X_+(\eta) = Y_+(-m), \tag{21}$$

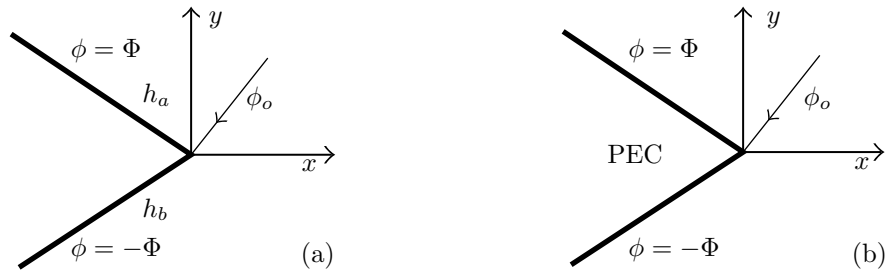


FIG. 4. (a) *An impenetrable wedge with arbitrary aperture angle* $\gamma_a = \gamma_b = \Phi$. (b) *PEC wedge with arbitrary aperture angle* $\gamma_a = \gamma_b = \Phi$.

where

$$X_+(\eta) = \begin{vmatrix} V_{z+}(\eta,0) \\ I_{z+}(\eta,0) \\ -\dfrac{\tau_o^2}{\omega\varepsilon} I_{\rho+}(\eta,0) - \dfrac{\alpha_o \eta}{\omega\varepsilon} I_{z+}(\eta,0) \\ \dfrac{\tau_o^2}{\omega\mu} V_{\rho+}(-\eta,0) + \dfrac{\alpha_o \eta}{\omega\mu} V_{z+}(\eta,0) \end{vmatrix},$$

$$Y_+(-m) = \begin{vmatrix} -\dfrac{\tau_o^2}{\omega\varepsilon} I_{\rho+}(-m,\Phi) + \dfrac{\alpha_o m}{\omega\varepsilon} I_{z+}(-m,\Phi) \\ \dfrac{\tau_o^2}{\omega\mu} V_{\rho+}(-m,\Phi) - \dfrac{\alpha_o m}{\omega\mu} V_{z+}(-m,\Phi) \\ \dfrac{\tau_o^2}{\omega\varepsilon} I_{\rho+}(-m,-\Phi) - \dfrac{\alpha_o m}{\omega\varepsilon} I_{z+}(-m,-\Phi) \\ -\dfrac{\tau_o^2}{\omega\mu} V_{\rho+}(-m,-\Phi) + \dfrac{\alpha_o m}{\omega\mu} V_{z+}(-m,-\Phi) \end{vmatrix}, \tag{22}$$

$$m = m(\eta) = -\eta\cos\Phi + \xi\sin\Phi, \qquad n = n(\eta) = -\xi\cos\Phi - \eta\sin\Phi.$$

In the general case, the matrix $G_\Phi(\eta)$ has a complicated expression that is not reported here.

The mapping to be used to obtain classical W-H equations in impenetrable wedge problems is (see Appendix B)

$$\eta = \eta(\overline{\eta}) = -\tau_o \cos\left[\frac{\Phi}{\pi}\arccos\left(-\frac{\overline{\eta}}{\tau_o}\right)\right]. \tag{23}$$

With this mapping, a generic plus function $F_+(\eta)$ in the η-plane is also a plus function $\overline{F}_+(\overline{\eta}) = F_+(\eta)$ in the $\overline{\eta}$-plane. Besides, since this mapping implies that $m = -\eta(-\overline{\eta})$, it follows that the minus function $Y_+(-m)$ in the m-plane becomes a minus function $\overline{Y}_+(-\overline{\eta}) = Y_+(-m)$ in the $\overline{\eta}$-plane. Consequently, the mapping reduces the equations (21) to the following W-H system in the $\overline{\eta}$-plane:

$$\overline{G}_\Phi(\overline{\eta})\overline{X}_+(\overline{\eta}) = \overline{Y}_+(-\overline{\eta}), \tag{24}$$

where $\overline{G}_\Phi(\overline{\eta}) = G_\Phi(\eta)$, $\overline{X}_+(\overline{\eta}) = X_+(\eta)$, $\overline{Y}_+(-\overline{\eta}) = Y_+(-m)$.

The solution of the classical W-H equation (24) requires the introduction of a source term. To this end we decompose the functions $Y_+(-m)$ in terms of the diffracted field (superscript d) and geometrical optical field (superscript g) as follows:

$$Y_+(-m) = Y_+^d(-m) + Y_+^g(-m). \tag{25}$$

The geometrical optical field is known. It depends on the direction of the plane wave φ_o and implies that the function $Y_+^g(-m)$ has a pole in $m_o = k\cos(\Phi - \varphi_o)$ with residue A_o. Taking into account that $m = -\eta(-\overline{\eta})$, it follows that $\overline{Y}_+^g(-\overline{\eta}) = Y_+^g(\eta(-\overline{\eta}))$ has a pole in $\overline{\eta}_o = -\tau_o \cos\frac{\pi}{\Phi}\varphi_o$, with residue R_o given by

$$R_o = A_o \left.\frac{d\overline{\eta}}{dm}\right|_{\overline{\eta}=\overline{\eta}_o} = A_o \frac{\pi}{\Phi}\frac{\sin\frac{\pi}{\Phi}\varphi_o}{\sin(\Phi - \varphi_o)}. \tag{26}$$

The standard factorization of $\overline{G}_\Phi(\overline{\eta}) = \overline{G}_{\Phi-}(\overline{\eta})\overline{G}_{\Phi+}(\overline{\eta})$ yields

$$\begin{aligned} &\overline{G}_{\Phi+}(\overline{\eta})\overline{X}_+(\overline{\eta}) - \overline{G}_{\Phi-}^{-1}(\overline{\eta}_o)\frac{R_o}{\overline{\eta} - \overline{\eta}_o} \\ &\qquad = \overline{G}_{\Phi-}^{-1}(\overline{\eta})[\overline{Y}_-^d(-\overline{\eta}) + \overline{Y}_+^g(-\overline{\eta})] - \overline{G}_{\Phi-}^{-1}(\overline{\eta}_o)\frac{R_o}{\overline{\eta} - \overline{\eta}_o} = w(\overline{\eta}). \end{aligned} \tag{27}$$

The member on the right-hand side of the first equality does not have the pole $\overline{\eta}_o = -\tau_o \cos\frac{\pi}{\Phi}\varphi_o$ and involves only minus functions that are regular for $\mathrm{Im}[\overline{\eta}] < \mathrm{Im}[-\tau_o]$; since the first member involves only plus functions, it follows that the function $w(\overline{\eta})$ is entire. Taking into account that $\eta \approx \overline{\eta}^{\Phi/\pi}$ as $\overline{\eta} \to \infty$, the presence of Laplace transforms, and the fact that standard factorized matrices have algebraic behavior as $\overline{\eta} \to \infty$ leads to the conclusion that $w(\overline{\eta})$ is vanishing. This provides the following solution of the W-H system (24):

$$\overline{X}_+(\overline{\eta}) = \overline{G}_{\Phi+}^{-1}(\overline{\eta})\overline{G}_{\Phi-}^{-1}(\overline{\eta}_o)\frac{R_o}{\overline{\eta} - \overline{\eta}_o}. \tag{28}$$

The W-H technique requires the decomposition in the $\overline{\eta}$-plane of arbitrary functions $\overline{X}(\overline{\eta})$:

$$X(\eta) = \overline{X}(\overline{\eta}) = X_+(\eta) + X_-(m) = \overline{X}_+(\overline{\eta}) + \overline{X}_-(\overline{\eta}). \tag{29}$$

An analytical manipulation of the Cauchy decomposition formula [4, p. 17] yields the following expression of the decomposed plus $\overline{X}_+(\overline{\eta}) = X_+(\eta)$:

$$\overline{X}_+(\overline{\eta}) = X_+(\eta) = -\frac{1}{\pi j}\int_{-\infty}^{\infty}[\hat{X}(-\pi + ju) - \hat{X}(-\pi - ju)]\frac{\sinh u}{\cosh u - \overline{\eta}/\tau_o}du - \sum_i \frac{R_i}{\overline{\eta} - \overline{\eta}_i}, \tag{30}$$

where $\hat{X}(\overline{w}) = \overline{X}(-\tau_o \cos\overline{w})$ and R_i represents the residue of $\overline{X}(\overline{\eta})$ in the poles $\overline{\eta}_i$ located in the half-plane $\mathrm{Im}[\overline{\eta}] < 0$.

To ascertain the validity of the solution (28), when possible we must verify that the results obtained with the W-H technique are equivalent to those obtained with other methods. We encountered some difficulties when, after considering specific problems, we compared the expressions obtained in this paper with the ones obtained by the Malyuzhinets method. In fact, the two approaches use two different spectral representations. These representations are the Laplace transform in the W-H technique and the Sommerfeld functions in the Malyuzhinets method. The difficulties were overcome only after we used expressions that relate the Sommerfeld function to the Laplace transform [3, 5, 22].

6. Solution of the CWHE in the scalar case. Let us consider the case of a PEC wedge excited by an E-polarized plane wave ($\partial/\partial z = 0, \tau_o = k$) (see Figure 4(b)). Even though simple, this case is significant since it shows all the difficulties involved in obtaining explicit W-H solutions in wedge problems having arbitrary aperture angles. For this problem, (4) and (6) may be ignored since the unknowns present in them are vanishing. Equations (3) and (5) simplify, becoming

$$\xi V_{z+}(\eta,0) - \omega\mu I_{\rho+}(\eta,0) = -\omega\mu I_{\rho+}(-m,\Phi), \tag{31}$$

$$\xi V_{z+}(\eta,0) + \omega\mu I_{\rho+}(\eta,0) = \omega\mu I_{\rho+}(-m,-\Phi), \tag{32}$$

where $\xi = \xi(\eta) = \sqrt{k^2-\eta^2}$ and $m = m(\eta) = -\eta\cos\Phi + \xi\sin\Phi$. Summing (31) and (32) yields the scalar generalized W-H equation

$$\xi X_+(\eta) = Y_+(-m), \tag{33}$$

with $X_+(\eta) = 2V_{z+}(\eta,0)$ and $Y_+(-m) = \omega\mu[I_{\rho+}(m,-\Phi) - I_{\rho+}(m,\Phi)]$. The mapping $\eta = \eta(\overline{\eta}) = -k\cos[\frac{\Phi}{\pi}\arccos(-\frac{\overline{\eta}}{k})]$ yields the following scalar CWHE equation in the $\overline{\eta}$-plane:

$$\overline{\xi}(\overline{\eta})\overline{X}_+(\overline{\eta}) = \overline{Y}_+(-\overline{\eta}) = \overline{X}_-(\overline{\eta}). \tag{34}$$

The key point for solving the CWHE is the factorization of the kernel $\overline{\xi}(\overline{\eta}) = \overline{\xi}_-(\overline{\eta})\overline{\xi}_+(\overline{\eta})$. Since we are dealing with scalar functions, this factorization can be accomplished by using well-known procedures. We obtain (see Appendix C)

$$\overline{\xi}_-(\overline{\eta}) = \sqrt{\frac{\tau_o+\overline{\eta}}{2}}, \qquad \overline{\xi}_+(\overline{\eta}) = \frac{\xi(\eta(\overline{\eta}))}{\overline{\xi}_-(\overline{\eta})}. \tag{35}$$

Taking the source into account, we consider, for illustrative purposes, the case (see Figure 4) in which the face $\varphi = -\Phi$ is in the shadow region: $I^g_{\rho+}(-m,-\Phi) = 0$. It yields

$$Y^g_+(-m) = \omega\mu[I^g_{\rho+}(-m,-\Phi) - I^g_{\rho+}(-m,\Phi)] = \frac{A_o}{[m - k\cos(\Phi-\varphi_o)]}, \tag{36}$$

with $A_o = 2jk\sin(\Phi-\varphi_o)E_o$. The residue R_o of $\overline{Y}^g_+(-\overline{\eta}) = Y^g_+(-m)$ in the simple pole $\overline{\eta}_o = -k\cos\frac{\pi}{\Phi}\varphi_o$ is given by (26) as follows:

$$R_o = A_o\left.\frac{d\overline{\eta}}{dm}\right|_{\overline{\eta}_o} = A_o\frac{\pi}{\Phi}\frac{\sin 2\frac{\varphi_o}{n}}{\sin(\Phi-\varphi_o)} = 4jkE_o\frac{\sin 2\frac{\varphi_o}{n_\Phi}}{n_\Phi},$$

with $n_\Phi = \frac{2\Phi}{\pi}$. Equation (28) provides the solution:

$$\overline{X}_+(\overline{\eta}) = \overline{\xi}^{-1}_+(\overline{\eta})\overline{\xi}^{-1}_-(\overline{\eta}_o)\frac{R_o}{\overline{\eta}-\overline{\eta}_o}. \tag{37}$$

To compare 37 to known solutions, we introduce the w-plane defined by $\eta = -k\cos w$ (see Appendix B). Algebraic manipulation on $V_{z+}(\eta,0) = \widetilde{X}_+(\overline{\eta})/2$ yields the well-known result

$$\begin{aligned} &V_{z+}(-k\cos w,0)\\ &\quad= \frac{jE_o\cos(\varphi_o/n_\Phi)}{n_\Phi k\sin w}\left(\frac{1}{\sin w/n_\Phi - \sin\varphi_o/n_\Phi} + \frac{1}{\sin w/n_\Phi + \sin\varphi_o/n_\Phi}\right). \end{aligned} \tag{38}$$

7. Solution of the CWHE in the vector case. In the wedge with two different face impedances excited by an E-polarized plane wave ($\partial/\partial z = 0, \tau_o = k$) (Malyuzhinets problem), (21) simplifies, and one obtains a vector generalized W-H equation of order two:

$$G_\Phi(\eta)F_+(\eta) = F_-(m), \tag{39}$$

where

$$F_+(\eta) = \begin{vmatrix} V_{z+}(\eta,0) \\ I_{\rho+}(\eta,0) \end{vmatrix}, \qquad F_-(m) = \begin{vmatrix} I_{\rho+}(-m,\Phi) \\ I_{\rho+}(-m,-\Phi) \end{vmatrix},$$

and the kernel $\overline{G}_\Phi(\overline{\eta})$ is the matrix given by

$$G_\Phi(\eta) = \begin{vmatrix} -\dfrac{\xi}{Z_a(n+n_a)} & \dfrac{\xi}{Z_a(n+n_a)}\dfrac{\omega\mu}{\xi} \\ \dfrac{\xi}{Z_b(n+n_b)} & \dfrac{\xi}{Z_b(n+n_b)}\dfrac{\omega\mu}{\xi} \end{vmatrix}, \tag{40}$$

where $n = -\xi\cos\Phi - \eta\sin\Phi$ and $n_{a,b} = \frac{\omega\mu}{Z_{a,b}}$.

In the $\overline{\eta}$-plane we will obtain an explicit factorization by reducing $\overline{G}_\Phi(\overline{\eta})$ to a Daniele–Khrapkov matrix. For this task an algebraic manipulation allows us to rewrite $\overline{G}_\Phi(\overline{\eta})$ in the form

$$\begin{aligned} &\overline{G}_\Phi(\overline{\eta}) \\ &= \frac{(Z_a - Z_b)\xi n}{2Z_aZ_b(n+n_a)(n+n_b)} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \dfrac{\sqrt{\omega\mu}}{\overline{\xi}_-(\overline{\eta})} \end{bmatrix} \begin{bmatrix} 1 & a \\ b & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \dfrac{\sqrt{\omega\mu}}{\overline{\xi}_+(\overline{\eta})} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \end{aligned} \tag{41}$$

where

$$a = f(\eta)\frac{\sqrt{\omega\mu}}{\overline{\xi}_-(\overline{\eta})}, \quad b = f(\eta)\frac{\overline{\xi}_-(\overline{\eta})}{\sqrt{\omega\mu}}, \quad f(\eta) = 1 - \frac{2(n+n_a)Z_a}{n(Z_a - Z_b)}.$$

Equation (41) reduces the matrix factorization of $\overline{G}_\Phi(\overline{\eta})$ to the factorization of the scalar $\xi n/(n+n_a)(n+n_b)$ and the matrix factorization of the central matrix $\left[\begin{smallmatrix} 1 & a \\ b & 1 \end{smallmatrix}\right]$. The factorization of the scalar needs the factorization of ξ, n, and $\frac{n}{n+n_{a,b}}$ in the $\overline{\eta}$-plane. The factorization of ξ is reported in (35). Using the same procedure, we obtain the factorization of $\overline{n}(\overline{\eta}) = \overline{n}_-(\overline{\eta})\overline{n}_+(\overline{\eta}) : \overline{n}_+(\overline{\eta}) = \sqrt{\frac{k-\overline{\eta}}{2}}$, $\overline{n}_-(\overline{\eta}) = \frac{n(\eta(\overline{\eta}))}{\overline{n}_+(\overline{\eta})}$. For what concerns the factorization of $\frac{n}{n+n_{a,b}} = [\frac{n}{n+n_{a,b}}]_-[\frac{n}{n+n_{a,b}}]_+$, we can use the logarithmic decomposition. In the presence of inductive impedances $Z_{a,b}$, (30) yields

$$\left[\frac{n+n_{a,b}}{n}\right]_+ = d_\Phi(\overline{\eta}) = \exp\left[\frac{1}{\pi}\int_0^\infty \arctan\left[\frac{\sin\vartheta_{a,b}}{\sinh\left[\frac{\Phi}{\pi}u\right]}\right]\frac{\sinh u}{\cosh u - \overline{\eta}/k}du\right]. \tag{42}$$

We are left with the more difficult problem consisting of factorizing the central matrix. Since $\overline{\xi}^2_-(\overline{\eta}) = \frac{1}{2}(k+\overline{\eta})$, the ratio $\frac{a}{b} = \frac{\omega\mu}{\overline{\xi}^2_-(\overline{\eta})}$ is a rational function of $\overline{\eta}$. It follows that the matrix $\left[\begin{smallmatrix} 1 & a \\ b & 1 \end{smallmatrix}\right]$ is a Daniele–Khrapkov matrix. For these matrices there is a well-known method [20]. It yields $\left[\begin{smallmatrix} 1 & a \\ b & 1 \end{smallmatrix}\right] = \left[\begin{smallmatrix} 1 & a \\ b & 1 \end{smallmatrix}\right]_- \bullet \left[\begin{smallmatrix} 1 & a \\ b & 1 \end{smallmatrix}\right]_+$, where the minus and plus

factorized matrices are given by

$$\begin{bmatrix} 1 & a \\ b & 1 \end{bmatrix}_- = \sqrt{\overline{g}_-(\overline{\eta})} \begin{vmatrix} \cosh\left[\frac{1}{2}\log\left[\frac{-Z_b}{Z_a}\right]\right] & \frac{\sqrt{\omega\mu}}{\overline{\xi}_-(\overline{\eta})}\sinh\left[\frac{1}{2}\log\left[\frac{-Z_b}{Z_a}\right]\right] \\ \frac{\overline{\xi}_-(\overline{\eta})}{\sqrt{\omega\mu}}\sinh\left[\frac{1}{2}\log\left[\frac{-Z_b}{Z_a}\right]\right] & \cosh\left[\frac{1}{2}\log\left[\frac{-Z_b}{Z_a}\right]\right] \end{vmatrix}$$

$$\bullet \begin{vmatrix} \cosh\left[\frac{\sqrt{\omega\mu}}{2}\overline{t}_-(\overline{\eta})\overline{\xi}_-(\overline{\eta})\right] & \frac{\sqrt{\omega\mu}}{\overline{\xi}_-(\overline{\eta})}\sinh\left[\frac{\sqrt{\omega\mu}}{2}\overline{t}_-(\overline{\eta})\overline{\xi}_-(\overline{\eta})\right] \\ \frac{\overline{\xi}_-(\overline{\eta})}{\sqrt{\omega\mu}}\sinh\left[\frac{\sqrt{\omega\mu}}{2}\overline{t}_-(\overline{\eta})\overline{\xi}_-(\overline{\eta})\right] & \cosh\left[\frac{\sqrt{\omega\mu}}{2}\overline{t}_-(\overline{\eta})\overline{\xi}_-(\overline{\eta})\right] \end{vmatrix}, \tag{43}$$

$$\begin{bmatrix} 1 & a \\ b & 1 \end{bmatrix}_+ = \sqrt{\overline{g}_+(\overline{\eta})} \begin{vmatrix} \cosh\left[\frac{\sqrt{\omega\mu}}{2}\overline{t}_+(\overline{\eta})\overline{\xi}_-(\overline{\eta})\right] & \frac{\sqrt{\omega\mu}}{\overline{\xi}_-(\overline{\eta})}\sinh\left[\frac{\sqrt{\omega\mu}}{2}\overline{t}_+(\overline{\eta})\overline{\xi}_-(\overline{\eta})\right] \\ \frac{\overline{\xi}_-(\overline{\eta})}{\sqrt{\omega\mu}}\sinh\left[\frac{\sqrt{\omega\mu}}{2}\overline{t}_+(\overline{\eta})\overline{\xi}_-(\overline{\eta})\right] & \cosh\left[\frac{\sqrt{\omega\mu}}{2}\overline{t}_+(\overline{\eta})\overline{\xi}_-(\overline{\eta})\right] \end{vmatrix}.$$

The factorization of the scalar

$$g(\eta) = 1 - f^2(\eta) = -\frac{4Z_aZ_b}{(Z_a - Z_b)^2}\frac{n+n_a}{n}\frac{n+n_b}{n} = \overline{g}_-(\overline{\eta})\overline{g}_+(\overline{\eta})$$

and the decomposition of the scalar

$$\begin{aligned} t(\eta) &= \frac{1}{\sqrt{\omega\mu}\overline{\xi}(\overline{\eta})}\log\left[\frac{1+f(\eta)}{1-f(\eta)}\right] - \frac{1}{\sqrt{\omega\mu}\overline{\xi}(\overline{\eta})}\log\left[-\frac{Z_b}{Z_a}\right] \\ &= \frac{1}{\sqrt{\omega\mu}\overline{\xi}_-(\overline{\eta})}\log\left[\frac{n+n_b}{n+n_a}\right] = \overline{t}_-(\overline{\eta}) + \overline{t}_+(\overline{\eta}) \end{aligned}$$

can be accomplished without any difficulty through (30). For instance, in the presence of inductive impedances $Z_{a,b}$, we have

$$\begin{aligned} \overline{t}_+(\overline{\eta}) &= \left\{\frac{1}{\overline{\xi}_-(\overline{\eta})}\log\left[\frac{n+n_b}{n+n_a}\right]\right\}_+ \\ &= -\frac{2}{\pi}\int_0^\infty \left[\arctan\left[\frac{\sin\vartheta_b}{\sinh\left[\frac{\Phi}{\pi}u\right]}\right] - \arctan\left[\frac{\sin\vartheta_a}{\sinh\left[\frac{\Phi}{\pi}u\right]}\right]\right]\frac{\sinh\frac{u}{2}}{\cosh u - \frac{\overline{\eta}}{k}}du, \end{aligned} \tag{44}$$

with $n_{a,b} = k\sin\vartheta_{a,b}$. Notice that $\overline{g}_-(\overline{\eta})$ and $\overline{g}_+(\overline{\eta})$ behave as constants as $\overline{\eta}\to\infty$. In addition, taking into account that $\eta \approx \overline{\eta}^{\Phi/\pi}$, it follows that $\overline{t}_\pm(\overline{\eta}) = O[\frac{1}{\overline{\eta}^{1/2}\overline{\eta}^{\Phi/\pi}}]$. Since the argument of the hyperbolic functions in (43) is constant or vanishing as $\overline{\eta}\to\infty$, it yields an algebraic behavior of both the factorized matrices.

To complete the solution process, one needs to know the value of $\overline{\eta}_o$ and R_o in order to use (28); once again, one has $\overline{\eta}_o = -k\cos\frac{\pi}{\Phi}\varphi_o$. Furthermore, by taking into account (26) and the reflection coefficient on the face $\varphi = \Phi$, one obtains

$$A_o = \begin{vmatrix} -\frac{2j\sin(\Phi-\varphi_o)}{Z_o + Z_a\sin(\Phi-\varphi_o)}E_o \\ 0 \end{vmatrix}, \qquad R_o = \begin{vmatrix} -\frac{\pi}{\Phi}\frac{2j\sin\frac{\pi}{\Phi}\varphi_o}{Z_o + Z_a\sin(\Phi-\varphi_o)}E_o \\ 0 \end{vmatrix}.$$

Even though the integrals (42) and (44) may be evaluated numerically without any difficulty, it is worth observing that they involve the Malyuzhinets function $\Psi_\Phi(w)$. In fact, a cumbersome mathematical deduction not reported here shows that

(45)
$$\left[\frac{n+n_{a,b}}{\sqrt{k}}\right]_+\Bigg|_{\eta=-k\cos w} = d_\Phi(-k)\frac{\Psi_\Phi(w+\Phi+\widetilde{\vartheta}_{a,b})\Psi_\Phi(w+\Phi-\widetilde{\vartheta}_{a,b})\Psi_\Phi(w-\Phi+\widetilde{\vartheta}_{a,b})\Psi_\Phi(w-\Phi-\widetilde{\vartheta}_{a,b})}{\Psi_\Phi(\Phi+\widetilde{\vartheta}_{a,b})\Psi_\Phi(\Phi-\widetilde{\vartheta}_{a,b})\Psi_\Phi(-\Phi+\widetilde{\vartheta}_{a,b})\Psi_\Phi(-\Phi-\widetilde{\vartheta}_{a,b})},$$

(46)
$$\left\{\frac{\sqrt{k}}{\overline{\xi}_-(\overline{\eta})}\log\left[\frac{n+n_b}{n+n_a}\right]\right\}_+\Bigg|_{\eta=-k\cos w} = \frac{1}{\sin\left[\frac{\pi w}{2\Phi}\right]}\log\left[\frac{\Psi_\Phi(w-\Phi-\widetilde{\vartheta}_b)\Psi_\Phi(w-\Phi+\widetilde{\vartheta}_b)\Psi_\Phi(w+\Phi-\widetilde{\vartheta}_a)\Psi_\Phi(w+\Phi+\widetilde{\vartheta}_a)}{\Psi_\Phi(w+\Phi-\widetilde{\vartheta}_b)\Psi_\Phi(w+\Phi+\widetilde{\vartheta}_b)\Psi_\Phi(w-\Phi-\widetilde{\vartheta}_a)\Psi_\Phi(w-\Phi+\widetilde{\vartheta}_a)}\right],$$

with $\eta=-k\cos w$, $w=\frac{\Phi}{\pi}\overline{w}$, $\widetilde{\vartheta}_{a,b}=\frac{\pi}{2}-\vartheta_{a,b}$. These alternative expressions are very important since they allow one to verify that the W-H solution (28) agrees completely with the Malyuzhinets solution [3].

8. Physical considerations on the W-H technique applied to wedge problems. The W-H solution $X_+(\eta)=\overline{X}_+(\overline{\eta})$ of (28) provides the Laplace transforms of the electromagnetic field only in the direction $\varphi=0$. In order to obtain $E_z(\rho,\varphi)$ and $H_z(\rho,\varphi)$ for every value of φ, it can be shown [23] that the two functions $s_E(w)$ and $s_H(w)$ defined by

(47)
$$\begin{aligned} s_E(w) &= \frac{j}{2}\left[-\tau_o\sin w V_{z+}(-\tau_o\cos w,0)\right.\\ &\qquad \left.+\frac{\tau_o^2}{\omega\varepsilon}I_{\rho+}(-\tau_o\cos w,0)-\frac{\alpha_o\tau_o\cos w}{\omega\varepsilon}I_{z+}(-\tau_o\cos w,0)\right],\\ s_H(w) &= \frac{j}{2}\left[-\tau_o\sin w I_{z+}(-\tau_o\cos w,0)\right.\\ &\qquad \left.-\frac{\tau_o^2}{\omega\mu}V_{\rho+}(-\tau_o\cos w,0)+\frac{\alpha_o\tau_o\cos w}{\omega\mu}V_{z+}(-\tau_o\cos w,0)\right] \end{aligned}$$

are the Sommerfeld functions of the problem. It yields the following representation of the longitudinal components valid for every value of φ:

(48)
$$\begin{aligned} E_z(\rho,\varphi) &= \frac{1}{2\pi}\left[\int_{c_1+c_2}s_E[w+\varphi]e^{+j\tau_o\cos[w]\rho}dw\right],\\ H_z(\rho,\varphi) &= \frac{1}{2\pi}\left[\int_{c_1+c_2}s_H[w+\varphi]e^{+j\tau_o\cos[w]\rho}dw\right], \end{aligned}$$

where c_1+c_2 is the Sommerfeld contour. By using well-known techniques such as the saddle point integration method [24] and the Watson lemma, (47) and (48) permit us to gain a complete physical insight into the solution $X_+(\eta)=\overline{X}_+(\overline{\eta})$, as obtained from (28). For instance, these solutions provide the diffraction coefficients of the

geometrical theory of diffraction (GTD) [18, 24] as well as the field behaviors near the edge $\rho = 0$ (see [22]). From the engineering standpoint, applications of (28) also include well-known refinements such as the uniform theory of diffraction (UTD) and the uniform asymptotic theory (UAT) [18].

When studying diffraction problems where no exact solutions are possible, of course we must introduce approximate solutions. The basic technique in electromagnetic engineering is the physical optics approximation. The W-H solution immediately provides the physical optics solution by observing that the presence of the denominator $\overline{\eta} - \overline{\eta}_o$ in the solution (28) indicates that the dominant spectrum is for $\overline{\eta} \approx \overline{\eta}_o$. It suggests that setting $\overline{G}_{\Phi+}^{-1}(\overline{\eta}) \approx \overline{G}_{\Phi+}^{-1}(\overline{\eta}_o)$ yields

$$\overline{X}_+(\overline{\eta}) \approx \overline{G}_{\Phi+}^{-1}(\overline{\eta}_o)\overline{G}_{\Phi-}^{-1}(\overline{\eta}_o)\frac{R_o}{\overline{\eta} - \overline{\eta}_o} = \overline{G}_{\Phi}^{-1}(\overline{\eta}_o)\frac{R_o}{\overline{\eta} - \overline{\eta}_o}. \tag{49}$$

This approximation represents the physical optical solution. To improve this approximation we can introduce rational approximants for the kernels. They yield rational matrices that in general can be factorized in closed form. It has been ascertained that this procedure can be very efficient in obtaining accurate solutions of W-H equations [11, 25, 26].

9. Conclusions. This paper shows that the wedge problems solved with the Sommerfeld–Malyuzhinets method can be successfully solved also by the classical W-H technique. There are many differences between the two methods. For instance, the Sommerfeld–Malyuzhinets approach requires the solution of difference equations, whereas the W-H technique involves decomposition-factorization problems. Even though it seems that both of the methods have the same capability to solve wedge problems in closed form, it should be observed that, with the extension also to arbitrary angular regions, there is no doubt that the W-H technique constitutes the most general analytical method for solving field problems involving geometrical discontinuities.

Appendix A. Derivation of the functional equation (3) of section 2. For illustrative purposes, we consider only the reasoning behind (3) in the free source angular region indicated in Figure 1. In this region the field E_z must satisfy the wave equation

$$\frac{\partial^2}{\partial x^2}E_z + \frac{\partial^2}{\partial y^2}E_z + \tau_o^2 E_z = 0. \tag{A1}$$

Let us introduce the oblique Cartesian coordinates u and v to study this region ($\gamma = \gamma_1$):

$$u = x - y\cot\gamma, \quad v = \frac{y}{\sin\gamma} \quad \text{or} \quad x = u + v\cos\gamma, \quad y = v\sin\gamma. \tag{A2}$$

With these coordinates (see Figure 1), (1) assumes the form

$$\frac{\partial^2 E_z}{\partial u^2} + \frac{\partial^2 E_z}{\partial v^2} - 2\cos\gamma\frac{\partial^2 E_z}{\partial u \partial v} + \tau_0^2 \sin^2\gamma E_z = 0. \tag{A3}$$

Equation (A3) will be solved in the Laplace domain by introducing the Laplace transform

$$\widetilde{E}_z(\eta, v) = \mathrm{L}\left[E_z(u, v)\right] = \int_0^\infty E_z(u, v)e^{j\eta u}du. \tag{A4}$$

It follows that

$$\frac{d^2\widetilde{E}_z}{dv^2} + 2j\eta\cos\gamma\frac{d\widetilde{E}_z}{dv} + (\tau_0^2\cos^2\lambda - \eta^2)\widetilde{E}_z = f_\eta(v), \tag{A5}$$

where

$$f_\eta(v) = -2\cos\gamma\frac{dE_z(0_+,v)}{dv} - j\eta E_z(0_+,v) + \left.\frac{\partial E_z(u,v)}{\partial u}\right|_{u=0_+}. \tag{A6}$$

A standard procedure [24, p. 274] yields the solution of (A5) as follows:

$$\widetilde{E}_z(\eta,v) = A(\eta)e^{-j\vec{m}(\eta)v} - \int_0^\infty \frac{e^{j\eta(v-v_1)\cos\gamma}e^{-j\xi|v-v_1|\sin\gamma}}{2j\sin\gamma\xi} f_\eta(v_1)dv_1, \tag{A7}$$

where $A(\eta)$ is an unknown function and $\vec{m}(\eta)$ and $\overleftarrow{m}(\eta)$ are the two solutions of the characteristic equation $-m^2 + 2\eta\cos\gamma m + \tau_o^2\sin^2\gamma - \eta^2 = 0$ ($\frac{d}{dv} \Rightarrow -jm$):

$$\vec{m}(\eta) = \cos\gamma\eta + \sin\gamma\xi, \qquad \overleftarrow{m}(\eta) = \cos\gamma\eta - \sin\gamma\xi, \tag{A8}$$

with $\xi = \xi(\eta) = \sqrt{\tau_o^2 - \eta^2}$, $\xi(0) = \tau_o$ for $\eta = 0$. On the semiaxis $v = 0$, (A7) provides the equation

$$V_+(\eta,0) = \widetilde{E}_z(0,\eta) = A(\eta) - \int_0^\infty \frac{e^{-jm(\eta)v_1}}{2j\sin\gamma\xi} f_\eta(v_1)dv_1, \tag{A9}$$

where

$$m(\eta) = -\eta\cos\gamma + \sqrt{\tau_o^2 - \eta^2}\sin\gamma. \tag{A10}$$

The magnetic field H_x is obtained by the Maxwell equations

$$H_x = -\frac{j\omega\varepsilon}{\tau_o^2}\left(-\frac{\partial E_z}{\partial v}\frac{1}{\sin\gamma} + \frac{\partial E_z}{\partial u}\cot\gamma\right) + \frac{j\alpha_o}{\tau_o^2}\frac{\partial H_z}{\partial x}. \tag{A11}$$

Applying the Laplace transform $\widetilde{H}_x(\eta,v) = \mathrm{L}[H_y(u,v)] = \int_0^\infty H_x(u,v)e^{j\eta u}du$ and taking into account (A7) yields

$$\begin{aligned} I_{\rho+}(\eta,0) &= \widetilde{H}_x(0,\eta) \\ &= \frac{j\omega\varepsilon}{\tau_o^2}\left(\frac{1}{\sin\gamma}\left[j\vec{m}A(\eta) + \frac{jm}{2j\sin\gamma\xi}\int_0^\infty e^{-jm(\eta)v}f_\eta(v)dv\right.\right. \\ &\qquad \left.\left. - j\eta\cos\gamma\widetilde{E}_z(0,\eta) - E_z(0,0)\cos\gamma\right]\right) \\ &\quad + \frac{j\alpha_o}{\tau_0^2}[-j\eta I_{z+}(\eta,0) - H_z(0,0)]. \end{aligned} \tag{A12}$$

Algebraic manipulations on the Maxwell equations written in the coordinates u and v yield

$$\frac{\partial E_z}{\partial u} = -\frac{\sin\gamma}{\omega\varepsilon}\left(-\alpha_o\frac{\partial H_z}{\partial u} - j\tau_o^2 H_v - \omega\varepsilon\cot\gamma\frac{\partial E_z}{\partial v}\right). \tag{A13}$$

Substituting (A13) into (A6) leads to the following result:

$$\int_0^\infty e^{-jm(\eta)v} f_\eta(v)dv = -j\cos\gamma(mV_{\rho+}(-m,\gamma) - E_z(0,0)) - j\eta V_{\rho+}(-m,\gamma) + \frac{\sin\gamma}{\omega\varepsilon}(\alpha_o jmI_{z+}(-m,0) - \alpha_o H_z(0,0) - j\tau_o^2 I_{\rho+}(-m,\gamma)). \tag{A14}$$

Eliminating $A(\eta)$ in (A7) and (A12) and taking into account (A14) yields the final equation (3) of section 2.

Appendix B. The mapping $\eta = \overline{\eta}(\eta)$. In this appendix we will find a mapping $\eta = \eta(\overline{\eta})$ that will reduce the GWHE (21) in the η-plane to a system of classical W-H equations (24) in the $\overline{\eta}$-plane. For facilitating function-theoretic manipulations that may be obscure in the η-plane, we introduce also the angular complex variable w defined by

$$\eta = k\cos(w+\pi) = -k\cos w. \tag{B1}$$

We also will use the notation

$$f_+(\eta) = f_+(-k\cos w) = \widehat{f}_+(w). \tag{B2}$$

The key point is the observation that the plus functions $f_+(\eta)$ involved in wedge problems yield functions $f_+(-k\cos w) = \widehat{f}_+(w)$, whose analytical continuations are even functions of w in the w-plane, i.e., $\widehat{f}_+(-w) = \widehat{f}_+(w)$. In addition, the functions $\widehat{f}_+(w)$ are also regular in the point $w = 0$. In fact, from both mathematical and physical considerations, it is possible to ascertain that the function $f_+(\eta)$ has only one branch point that is located in $\eta = k$. It follows that $\widehat{f}_+(w)$ are meromorphic functions of w, i.e., $\widehat{f}_+(w) = \frac{n(w)}{d(w)}$ with $n(w)$ and $d(w)$ being entire functions. From (11)–(14) the functions $f_+(\eta)$ appear as functions of $\sqrt{\tau_o - \eta} = \sqrt{2\tau_o}\cos(\frac{w}{2})$ and $\eta = -\tau_o\cos w = \tau_o(1 - 2\cos^2\frac{w}{2})$. It means that near $\eta = \tau_o$, i.e., $w = -\pi$, we have

$$n(w) = a_0 + a_2w^2 + a_4w^4 + \cdots, \qquad d(w) = b_0 + b_2w^2 + b_4w^4 + \cdots \quad (w \approx -\pi). \tag{B3}$$

Since we are dealing with entire functions, (B3) holds for every w. It implies the evenness of $f_+(\eta) = n(w)/d(w)$ everywhere. Moreover, since the point $w = 0$ corresponds to the regular point $\eta = -\tau_o$, $f_+(-\tau_o\cos w)$ is regular in $w = 0$.

Conversely, the meromorphic functions $f(-\tau_o\cos w)$ that are even in the w-plane and regular in $w = 0$ do not involve the branch point $\eta = -\tau_o$ in the η-plane. In fact, a Taylor expansion near $w = 0$ involves only even powers of w:

$$f(-k\cos w) = c_0 + c_2w^2 + c_4w^4 + \cdots \quad (w \approx 0). \tag{B4}$$

Since $w = \sqrt{\tau_o^2 - \eta^2}s_+(\eta)$, where the plus function $s_+(\eta)$ is given by

$$s_+(\eta) = \frac{1}{j\pi\sqrt{\tau_o^2 - \eta^2}}\ln\frac{j\sqrt{\tau_o^2 - \eta^2} - \tau_o - \eta}{j\sqrt{\tau_o^2 - \eta^2} + \tau_o + \eta}, \tag{B5}$$

we can write

$$f(\eta) = c_0 + c_2(\tau_o^2 - \eta^2)(s_+(\eta))^2 + c_4(\tau_o^2 - \eta^2)^2(s_+(\eta))^4 + \cdots \quad (\eta \approx -\tau_o). \tag{B6}$$

It follows that $f(\eta)$ does not have a branch point in $\eta = -\tau_o$. We can also extend our discussion to minus functions $f_-(\eta)$ by relating them to plus function $g_+(\eta)$ through

$$f_-(\eta) = g_+(-\eta). \tag{B7}$$

In the proper sheet it must be

$$\xi = \sqrt{\tau_o^2 - \eta^2} = \sqrt{\tau_o^2 - (-\eta)^2} = -\tau_o \sin w. \tag{B8}$$

This equality holds only if $w(-\eta) = -w(\eta) - \pi$. Using the notation (B2), it follows that $g_+(-\eta) = \hat{g}_+(-w - \pi)$. This result also shows that a minus function does not change value when in the w-plane w is substituted by $-2\pi - w$.

In the w-plane we have $m = \tau_o \cos(w + \Phi)$, and the GWHE (21) assumes the form

$$\hat{G}_\Phi(w)\hat{X}_+(w) = \hat{Y}_+(w + \Phi). \tag{B9}$$

Now the mapping

$$w = \frac{\Phi}{\pi}\overline{w} \tag{B10}$$

yields

$$\widetilde{G}_\Phi(\overline{w})\widetilde{X}_+(\overline{w}) = \widetilde{Y}_+(\overline{w} + \pi), \tag{B11}$$

where the following notation has been used:

$$\widetilde{G}_\Phi(\overline{w}) = \hat{G}_\Phi(w) = \hat{G}_\Phi\left(\frac{\Phi}{\pi}\overline{w}\right). \tag{B12}$$

The mapping $w = \frac{\Phi}{\pi}\overline{w}$ ensures that $\widetilde{X}_+(\overline{w})$ and $\widetilde{Y}_+(\overline{w})$ are also even functions of $\overline{w}$. Consequently, in the plane $\overline{\eta}$ defined by $\overline{\eta} = -\tau_o \cos\overline{w} = -\tau_o \cos\frac{\pi}{\Phi}w$ or

$$\overline{\eta} = -\tau_o \cos\left[\frac{\pi}{\Phi}\arccos\left(-\frac{\eta}{\tau_o}\right)\right], \qquad \eta = \eta(\overline{\eta}) = -\tau_o \cos\left[\frac{\Phi}{\pi}\arccos\left(-\frac{\overline{\eta}}{\tau_o}\right)\right], \tag{B13}$$

these functions do not involve the branch point $\overline{\eta} = -\tau_o$. In addition, we have

$$m = -\eta(-\overline{\eta}), \tag{B14}$$

$$Y_+(-m) = \hat{Y}_+(-w - \Phi) = \widetilde{Y}_+(-\overline{w} - \pi) = \widetilde{Y}_+(\overline{w}(-\overline{\eta})) = \hat{Y}_+(w(-\overline{\eta})) = \overline{Y}_+(-\overline{\eta}). \tag{B15}$$

It means that $\overline{Y}_{\Phi+}(-\overline{\eta})$ do not involve the branch point $\overline{\eta} = \tau_o$ in the $\overline{\eta}$-plane.

The singularities involved in the $\overline{\eta}$-plane derive from the ones present in the η-plane. They consist of (a) a finite number of poles (structural surface waves), (b) branch points $\overline{\eta} = \pm\tau_o$, and (c) a limited number of poles that arise from the geometrical optic field. The limited number of the poles involved allows one to conclude that $\overline{X}_+(\overline{\eta})$ has a regular upper half-plane and that $\overline{X}_-(\overline{\eta}) = \overline{Y}_+(-\overline{\eta})$ has a regular lower half-plane. Consequently, (24) constitutes a classical system of W-H equations.

To facilitate the applications of the mapping $\eta = \overline{\eta}(\eta)$ defined in (B13), it is useful to remember that the variables w and $\overline{w}$ are defined by

$$\eta = -\tau_o \cos w, \quad \overline{\eta} = -\tau_o \cos \overline{w}, \quad w = \frac{\Phi}{\pi}\overline{w}. \tag{B16}$$

With these variables, we have $\xi = -\tau_o \sin w$, $m = \tau_o \cos(w+\Phi)$, and $n = \tau_o \sin(w+\Phi)$. As an example, the pole $m_o = \tau_o \cos(\Phi - \varphi_o) = \tau_o \cos(w_o + \Phi)$ implies $w_o = -\varphi_o$ or $\overline{w}_o = -\frac{\pi}{\Phi}\varphi_o$, which yields $\overline{\eta}_o = -\tau_o \cos \overline{w}_o = -\tau_o \cos \frac{\pi}{\Phi}\varphi_o$.

Appendix C. Factorization of the scalar ξ. There are many ways to obtain the factorization of ξ in the $\overline{\eta}$-plane (see [22]). In this appendix we will obtain this factorization by inspection, by observing that in the $\overline{w}$-plane

$$\xi = \sqrt{k^2 - \eta^2} = -k \sin w = \left(-\sqrt{k} \sin \frac{1}{2}\overline{w}\right) \frac{\sqrt{k} \sin \frac{\Phi}{\pi}\overline{w}}{\sin \frac{1}{2}\overline{w}}. \tag{C1}$$

The factorization is thus accomplished. In fact

$$-\frac{\sqrt{k} \sin \frac{\Phi}{\pi}\overline{w}}{\sin \frac{1}{2}\overline{w}}$$

is a plus function since it is even in $\overline{w}$ and regular in $\overline{w} = 0$. Here also $\sqrt{k} \sin \frac{1}{2}\overline{w}$ is a minus function since it does not change when $\overline{w}$ is substituted by $-2\pi - \overline{w}$. In the $\overline{\eta}$-plane we have

$$\overline{\xi}_-(\overline{\eta}) = \sqrt{k} \sin \frac{1}{2}\overline{w} = \sqrt{k \frac{1 - \cos \overline{w}}{2}} = \sqrt{\frac{k + \overline{\eta}}{2}}. \tag{C2}$$

REFERENCES

[1] H. Poincaré, *Sur la polarization par diffraction*, Acta Math., 16 (1892), pp. 297–339.

[2] A. Sommerfeld, *Mathematische theorie der diffraktion*, Math. Ann., 47 (1896), pp. 317–341.

[3] A.V. Osipov and A.N. Norris, *The Malyuzhinets theory for scattering from wedge boundaries: A review*, Wave Motion, 29 (1999), pp. 313–340.

[4] B. Noble, *Methods Based on the Wiener–Hopf Technique*, Pergamon Press, London, 1958.

[5] B. Budaev, *Diffraction by Wedges*, Longman Scientific and Technical, Harlow, UK, 1995.

[6] B.V. Budaev and D.B. Bogy, *Rayleigh wave scattering by an elastic wedge*, Wave Motion, 22 (1995), pp. 239–257.

[7] B.V. Budaev and D.B. Bogy, *Rayleigh wave scattering by an elastic wedge* II, Wave Motion, 24 (1996), pp. 307–314.

[8] B.V. Budaev and D.B. Bogy, *Scattering of Rayleigh and Stonely waves by two adhering elastic wedges*, Wave Motion, 33 (2001), pp. 321–337.

[9] K. Fujii, *Rayleigh-wave scattering at various corners: Investigation in the wider range of wedge angles*, Bull. Seismol. Soc. Amer., 84 (1994), pp. 1916–1924.

[10] A.K. Gautesen, *Scattering of a Rayleigh wave by an elastic wedge whose angle is greater than* 180 *degrees*, ASME J. Appl. Mech., 68 (2001), pp. 476–479.

[11] V. Daniele and R. Zich, *An approximate factorization for the kernel involved in the scattering by a wedge at skew incidence*, in Proceedings of the 9th International Conference on Mathematical Methods in Electromagnetic Theory (MMET 02), Kiev, Ukraine, 2002, IEEE Electron Devices Society, 2002, pp. 130–135.

[12] T.B.A. Senior, *Some problems involving imperfectly half planes*, in Electromagnetic Scattering, P.L.E. Uslenghi, ed., Academic Press, New York, 1978, pp. 185–219.

[13] E. Lüneburg and A.H. Serbest, *Diffraction of a obliquely incident plane wave by a two-face impedance half plane: Wiener–Hopf approach*, Radio Sci., 35 (2000), pp. 1361–1374.

[14] R.A. Hurd and E. Lüneburg, *Diffraction by an anisotropic impedance half-plane*, Canad. J. Phys., 63 (1985), pp. 1135–1140.

[15] T.B.A. Senior and S.R. Legault, *Diffraction by an anisotropic impedance half-plane at skew incidence*, Electromagnetics, 18 (1998), pp. 207–225.
[16] V.G. Daniele, *On the solution of two coupled Wiener–Hopf equations*, SIAM J. Appl. Math., 44 (1984), pp. 667–680.
[17] V.G. Daniele, *On the solution of vector Wiener–Hopf equations occurring in scattering problems*, Radio Sci., 19 (1984), pp. 1173–1178.
[18] T.B.A. Senior and J.L. Volakis, *Approximate Boundary Conditions in Electromagnetics*, The Institution of Electrical Engineers, London, 1995.
[19] R. Sendag and H. Serbest, *Scattering at the junction formed by PEC half-plane and a half-plane with anisotropic conductivity*, Electromagnetics, 21 (2001), pp. 415–434.
[20] A. Büyükaksoy and A.H. Serbest, *Matrix Wiener–Hopf factorization methods and applications to some diffraction problems*, in Analytical and Numerical Methods in Electromagnetic Wave Theory, M. Hashimoto, M. Idemen, and O.A. Tretyakov, eds., Science House, Tokyo, 1993, pp. 257–315.
[21] V. Daniele, *Generalized Wiener–Hopf technique for wedge shaped regions of arbitrary angles*, in Proceedings of the 8th International Conference on Mathematical Methods in Electromagnetic Theory (MMET 2000), Kharkov, Ukraine, 2000, pp. 432–434.
[22] V. Daniele, *New analytical methods for wedge problems*, in Proceedings of the 2001 International Conference on Electromagnetics in Advanced Applications (ICEAA01), 2001, Torino, Italy, IEEE, New York, Washington, pp. 385–393.
[23] V.G. Daniele, *Two new methods for wedge problems*, Atti della Fondazione Giorgio Ronchi, 57 (2002), pp. 629–650.
[24] L.B. Felsen and N. Marcuvitz, *Radiation and Scattering of Waves*, Prentice–Hall, Englewood Cliffs, NJ, 1973.
[25] I.D. Abrahams, *The application of Padé approximants to Wiener–Hopf factorization*, IMA J. Appl. Math., 65 (2000), pp. 257–281.
[26] I.D. Abrahams, *On the non-commutative factorization of Wiener–Hopf kernels of Khrapkov type*, Proc. Roy. Soc. London Ser. A, 454 (1998), pp. 1719–1743.

SIAM J. APPL. MATH.
Vol. 63, No. 4, pp. 1461–1473

MYRIAD RADIAL CAVITATING EQUILIBRIA IN NONLINEAR ELASTICITY*

JEYABAL SIVALOGANATHAN† AND SCOTT J. SPECTOR‡

For Donald E. Carlson on the occasion of his 65th birthday

Abstract. It is shown that every bounded strictly increasing smooth positive function of sufficiently slow growth is the Jacobian of a radial hole creating equilibrium deformation for an appropriately constructed compressible nonlinearly elastic energy.

Key words. cavitation, elastic, equilibrium, singular minimizers

AMS subject classifications. 74B20, 74G70, 35J55

PII. S0036139901397005

1. Introduction.

Explicit solutions of model equations can be useful in gaining insight concerning the qualitative behavior of solutions to more general problems. Unfortunately, the explicit construction of radial equilibrium deformations that create new holes in a compressible nonlinearly elastic body has proven to be unexpectedly complicated. Consequently, although cavitation is a common occurrence in rubbery polymers, explicit solutions that exhibit this phenomenon are rare. The only such solutions (modulo a radial null-Lagrangian; see Horgan [3] and Steigmann [15]) that appear in the literature are for an elastic fluid (see, e.g., [3, 6]); for the Blatz–Ko constitutive relation for foam rubbers, which was obtained, in two dimensions, by Horgan and Abeyaratne [4] and, in three dimensions, by Tian-hu [18]; for a compressible neo-Hookean material, which was obtained in [11] (see also [1, section 7.6]); and for the generalized Carroll material, which was obtained by Murphy and Biwa [7] (see also Shang and Cheng [12]).

The usual method of obtaining an explicit solution is to solve the differential equation for a postulated model problem. In this paper we take a different approach; we first posit deformations that, based upon prior results, have desired properties, and then construct differential equations that have these deformations as solutions. We show, in particular, that every function in a certain class of radial cavitating deformations will satisfy an equilibrium equation that is appropriately chosen for that particular function. The radial deformations we use are those for which the Jacobian is an increasing radial function. The appropriate stored energy is then constructed as the sum of two terms. The first is a homogeneous isotropic strongly elliptic stored-energy function, while the second is a function of the Jacobian of the chosen radial deformation. This second function is constructed so that the chosen deformation will automatically satisfy the radial equilibrium equation. Further information on radial cavitation is contained in the survey article by Horgan and Polignone [5].

*Received by the editors October 24, 2001; accepted for publication (in revised form) January 7, 2003; published electronically May 29, 2003. This work was supported in part by the National Science Foundation under grant 0072414.
http://www.siam.org/journals/siap/63-4/39700.html
†Department of Mathematical Sciences, University of Bath, Bath BA2 7AY, England (js@maths.bath.ac.uk).
‡Department of Mathematics, Southern Illinois University, Carbondale, IL 62901–4408 (sspector@math.siu.edu).

2. The constitutive relation. Let $\Psi \in C^2((0,\infty)^n)$ be a symmetric function. We assume that the stored-energy function for the material is given by

$$W(\mathbf{F}) = \Phi(\nu_1, \nu_2, \dots, \nu_n) := \Psi(\nu_1, \nu_2, \dots, \nu_n) + h(\nu_1 \nu_2 \dots \nu_n) \tag{2.1}$$

for all $n \times n$ matrices $\mathbf{F}$ with positive determinant, where $\nu_1, \nu_2, \dots, \nu_n$ are the principal stretches, i.e., the eigenvalues of the square root of $\mathbf{F}\mathbf{F}^T$, and $h \in C^2((0,\infty))$ is a function to be determined.

The problem of interest is to determine stationary points of the energy

$$E(\mathbf{w}) = \int_{B_\mathrm{o}} W(\nabla \mathbf{w}(\mathbf{x}))\, d\mathbf{x} \tag{2.2}$$

among orientation-preserving injective $\mathbf{w}$ that satisfy the boundary condition $\mathbf{w}(\mathbf{x}) = \lambda \mathbf{x}$ for $\mathbf{x} \in \partial B_\mathrm{o}$, where $B_\mathrm{o} := B(\mathbf{0}, R_\mathrm{o}) \subset \mathbb{R}^n$ is the ball of radius R_o centered at the origin. For a radial deformation

$$\mathbf{w}(\mathbf{x}) = \frac{r(R)}{R}\mathbf{x}, \qquad R := |\mathbf{x}|,$$

$r : [0, R_\mathrm{o}] \to [0,\infty)$, the principal stretches at any point $\mathbf{x} \in B_\mathrm{o}$ are given by (see, e.g., [1]) $\nu_1(\mathbf{x}) = r'(R)$ and $\nu_i(\mathbf{x}) = r(R)/R$ for $i = 2, 3, \dots, n$. Thus (2.2) reduces to[1]

$$\overline{E}(r) = \int_0^{R_\mathrm{o}} \Phi\left(r'(R), \frac{r(R)}{R}, \frac{r(R)}{R}, \dots, \frac{r(R)}{R}\right) R^{n-1}\, dR \tag{2.3}$$

among those $r : [0, R_\mathrm{o}] \to [0,\infty)$ that satisfy $r' > 0$ a.e. and $r(R_\mathrm{o}) = \lambda R_\mathrm{o}$. A stationary point $\mathbf{w}$ of E corresponds to a solution r of the radial equilibrium equation

$$\frac{d}{dR}\left[R^{n-1}\Phi_{,1}\right] = (n-1)R^{n-2}\Phi_{,2}\,, \tag{2.4}$$

where

$$\Phi_{,i} = \Phi_{,i}\left(r'(R), \frac{r(R)}{R}, \frac{r(R)}{R}, \dots, \frac{r(R)}{R}\right)$$

and $\Phi_{,i}\,(v_1, v_2, \dots, v_n)$ denotes differentiation of Φ with respect to its ith argument (see [1, Theorem 7.3]). Also, if $r(0) > 0$, then the deformed ball contains a spherical cavity, and r must satisfy the natural boundary condition

$$T(R) := \left[\frac{R}{r(R)}\right]^{n-1} \Phi_{,1}\left(r'(R), \frac{r(R)}{R}, \frac{r(R)}{R}, \dots, \frac{r(R)}{R}\right) \to 0 \ \text{ as } R \to 0^+, \tag{2.5}$$

which corresponds to the radial component of the Cauchy stress vanishing on the cavity surface.

For energies of the form (2.1) the radial equilibrium equation (2.4) becomes

$$\frac{d}{dR}\left[R^{n-1}\Psi_{,1}\right] - (n-1)R^{n-2}\Psi_{,2} = -r(R)^{n-1}\frac{d}{dR}h'\left(r'(R)\left[\frac{r(R)}{R}\right]^{n-1}\right), \tag{2.6}$$

[1]The energy E is equal to $\overline{E}$ multiplied by the volume of the unit ball in $\mathbb{R}^n$.

and the natural boundary condition (2.5) reduces to

$$\lim_{R\to 0^+}\left[h'\left(r'(R)\left[\frac{r(R)}{R}\right]^{n-1}\right)+\left[\frac{R}{r(R)}\right]^{n-1}\Psi_{,1}\left(r'(R),\frac{r(R)}{R},\dots,\frac{r(R)}{R}\right)\right]=0. \tag{2.7}$$

The main idea in this paper is that, given Ψ and r, (2.6) can be used to define the function h. In order to accomplish this, we will need the following hypotheses on the energy:

(En1) for all $q>0$ and $t>0$

$$\Psi_{,11}(q,t,t,\dots,t)>0;$$

(En2) there exists $\lambda_*>0$ such that for every $\alpha\geq\lambda_*^n$

$$\lim_{t\to+\infty} t^{1-n}\Psi_{,1}(\alpha t^{1-n},t,t,\dots,t)=0;$$

(En3) for every $L>\lambda_*^n$ there are constants $\beta\in[0,n-1)$ and $K>0$ such that

$$\left|\Psi_{,2}(\kappa t^{1-n},t,t,\dots,t)\right|\leq Kt^{\beta}$$

for all $\lambda_*<t<\infty$ and $\lambda_*^n\leq\kappa\leq L$;

(En4) for $q\neq t$ define

$$\mathcal{R}(q,t):=\frac{q\Psi_{,1}(q,t,t,\dots,t)-t\Psi_{,2}(q,t,t,\dots,t)}{q-t}. \tag{2.8}$$

Then for every $\mu>\lambda_*$ we assume that there exists a $B_\mu>0$ such that

$$|\mathcal{R}(q,t)|\leq B_\mu$$

for all $\lambda_*<t<\mu$ and $0<q<t$.

Remark 2.1. Hypothesis (En1) is a consequence of the strong-ellipticity of the energy Ψ. Hypotheses (En2)–(En4) are satisfied by many examples of stored energies (see [1, 16, 17]). In particular, Stuart [16, 17] requires a more stringent growth hypothesis than (En4): $0\leq\mathcal{R}(q,t)\leq A+Bt^{\beta}$ for $0<q<t$.

3. The construction. Let $\lambda_{\rm crit}>\lambda_0>0$, and suppose that $J\in C^1([0,\infty))$ is a strictly monotone increasing function that satisfies

$$J(0)=\lambda_0^n,\qquad \lim_{R\to+\infty}J(R)=\lambda_{\rm crit}^n,\qquad \int_0^\infty J'(t)t^n\,dt\leq 1. \tag{3.1}$$

Define $\rho:[0,\infty)\to[1,\infty)$ by

$$\rho(R)^n:=1+n\int_0^R J(t)t^{n-1}\,dt \tag{3.2}$$

so that

$$\rho'(R)\left[\frac{\rho(R)}{R}\right]^{n-1}=J(R)\quad\text{for}\quad 0<R<\infty. \tag{3.3}$$

For future reference we note the following properties of ρ.

LEMMA 3.1. *The function ρ given by* (3.1) *and* (3.2) *satisfies*

(i) $0 < \rho'(R)$,

(ii) $\frac{d}{dR}[\frac{\rho(R)}{R}] < 0$,

(iii) $\rho'(R) < \frac{\rho(R)}{R}$,

(iv) $0 < \rho''(R)$,

(v) $\lim_{R\to+\infty} \frac{\rho(R)}{R} = \lim_{R\to+\infty} \rho'(R) = \lambda_{\text{crit}}$.

Remark 3.2. Properties (i)–(v) are standard properties of radial minimizers and radial equilibrium deformations (see, e.g., [1, 10, 16, 17]). Since our proof shows that $(3.1)_3$ is necessary and sufficient for (iii), it is now clear that $(3.1)_3$ is also a standard property of such deformations. Note also that by (3.2), $\rho(0) = 1$.

THEOREM 3.3. *Let Ψ satisfy* (En1)–(En4) *and let ρ be given by* (3.1) *and* (3.2), *where $\lambda_0 > \lambda_*$. Suppose that $h \in C^2((0,\infty);(0,\infty))$ and satisfies*

$$\begin{aligned} h'(J(R)) = &\int_0^R \frac{(n-1)s^{n-2}}{\rho(s)^{n-1}} \hat{\Psi}_2\left(\rho'(s), \frac{\rho(s)}{s}\right) ds \\ &- R^{n-1}\hat{\Psi}_1\left(\rho'(R), \frac{\rho(R)}{R}\right)\left[\frac{1}{\rho(R)}\right]^{n-1} \\ &+ \int_0^R s^{n-1}\hat{\Psi}_1\left(\rho'(s), \frac{\rho(s)}{s}\right) \frac{d}{ds}\left[\frac{1}{\rho(s)^{n-1}}\right] ds \end{aligned} \tag{3.4}$$

for $R \in (0,\infty)$, where $\hat{\Psi}_i(\rho'(s), \frac{\rho(s)}{s}) := \Psi_{,i}\,(\rho'(s), \frac{\rho(s)}{s}, \frac{\rho(s)}{s}, \ldots, \frac{\rho(s)}{s})$. Then ρ is a solution of the radial equilibrium equation (2.6) *on $(0,\infty)$ and satisfies the natural boundary condition* (2.7).

Remark 3.4. It follows from the proof of the above result (see (4.4)) that $h'(\lambda_0^n) = 0$.

Finally, we use ρ to construct a family of equilibrium deformations of B_{o}. For any $R_{\text{o}} > 0$ and $\delta > 0$

$$\mathbf{u}_\delta(\mathbf{x}) := \frac{r_\delta(|\mathbf{x}|)}{|\mathbf{x}|}\mathbf{x}, \qquad r_\delta(R) := \frac{\rho(\delta R)}{\delta} \tag{3.5}$$

is an orientation-preserving injective radial deformation of B_{o}, and, in view of (3.3), (4.1), and [1, Lemma 4.1], $J(\delta|\mathbf{x}|)$ is the Jacobian of $\mathbf{u}_\delta$ at any point $\mathbf{x} \neq \mathbf{0}$. Clearly, each $\mathbf{u}_\delta$ is a cavitating deformation that creates a new hole of radius $1/\delta$ at the center of the ball and satisfies the boundary condition

$$\mathbf{u}_\delta(\mathbf{x}) = \lambda\mathbf{x}, \qquad \lambda^n = \lambda(\delta, R_{\text{o}})^n := \frac{[1 + n\int_0^{\delta R_{\text{o}}} J(t)t^{n-1}\,dt]}{(\delta R_{\text{o}})^n} \tag{3.6}$$

for $\mathbf{x} \in \partial B_{\text{o}}$. Moreover, results in [1] show that each of these deformations is contained in the Sobolev space $W^{1,p}(B_{\text{o}};\mathbb{R}^n)$ for every $p \in [1,n)$, while Theorem 3.3 shows that $\mathbf{u}_\delta$ is a stationary point for the energy.

THEOREM 3.5. *Let W be given by* (2.1) *and satisfy* (En1)–(En4). *Let $\lambda > \lambda_{\text{crit}} > \lambda_0 > \lambda_*$. Then there exists a unique $\delta = \delta(\lambda)$ such that $\mathbf{u}_\delta$, given by* (3.1), (3.2), (3.5), *and* (3.6), *is a stationary point of the energy* (2.2) *and satisfies the boundary condition $\mathbf{u}_\delta(\mathbf{x}) = \lambda\mathbf{x}$ for $\mathbf{x} \in \partial B_{\text{o}}$.*

Proof. Let $\lambda > \lambda_{\text{crit}}$. Then since $\rho(0) = 1$, it is clear from Lemma 3.1(ii) and (v) that there exists a unique $R_\lambda > 0$ such that $\rho(R_\lambda) = \lambda R_\lambda$. Define $\delta = R_\lambda/R_{\text{o}}$. Then

by (3.5)

$$\mathbf{u}_\delta(\mathbf{x}) := \frac{\rho(\delta R_{\rm o})}{\delta R_{\rm o}}\mathbf{x} = \frac{\rho(R_\lambda)}{R_\lambda}\mathbf{x} = \lambda\mathbf{x} \quad \text{for} \quad |\mathbf{x}| = R_{\rm o}. \qquad \square$$

4. Proofs for the construction.

Proof of Lemma 3.1. We first note that (i) is clear from (3.2), (3.3), and the nonnegativity of J. Next, if we divide (3.2) by R^n and use the quotient rule to differentiate the result with respect to R, we find that

$$\begin{aligned}\left[\frac{\rho(R)}{R}\right]^{n-1}\frac{d}{dR}\left[\frac{\rho(R)}{R}\right] &= \frac{\left(nJ(R)R^{n-1}\right)R^n - nR^{n-1}(1+n\int_0^R J(t)t^{n-1}\,dt)}{nR^{2n}}\\ &= \frac{J(R)R^n - (1+n\int_0^R J(t)t^{n-1}\,dt)}{R^{n+1}}\\ &= \frac{-1+\int_0^R J'(t)t^n\,dt}{R^{n+1}},\end{aligned} \tag{4.1}$$

where an integration by parts has been used to deduce $(4.1)_3$ from $(4.1)_2$. It is now clear from (4.1) that $(3.1)_3$ is necessary and sufficient for (ii).

Next, if we differentiate $\rho(R)/R$ with respect to R, we see that

$$\frac{d}{dR}\left[\frac{\rho(R)}{R}\right] = \frac{1}{R}\left[\rho'(R) - \frac{\rho(R)}{R}\right], \tag{4.2}$$

and consequently (iii) is equivalent to (ii). Similarly, if we differentiate (3.3) with respect to R, we discover that

$$\rho''(R)\left[\frac{\rho(R)}{R}\right]^{n-1} = J'(R) - \rho'(R)(n-1)\left[\frac{\rho(R)}{R}\right]^{n-2}\frac{d}{dR}\left[\frac{\rho(R)}{R}\right], \tag{4.3}$$

and thus (iv) follows from (i), (ii), and the fact that J is increasing.

Finally, to obtain (v) we first note that (ii)–(iv) imply that both limits exist and are finite. Thus, if we divide (3.2) by R^n and take the limit as $R \to +\infty$, we find, using L'Hôpital's rule and $(3.1)_2$, that

$$\lim_{R\to+\infty}\left[\frac{\rho(R)}{R}\right]^n = \lim_{R\to+\infty} J(R) = \lambda_{\rm crit}^n,$$

which together with (3.3) yields (v). $\square$

Proof of Theorem 3.3. Assume for the moment that $s \mapsto s^{n-2}\hat{\Psi}_2(\rho'(s), \frac{\rho(s)}{s})$ and $s \mapsto s^{n-1}\hat{\Psi}_1(\rho'(s), \frac{\rho(s)}{s})$ are integrable on $(0,R)$, so that the right-hand side of (3.4) is well defined on $(0,\infty)$. Then, if we differentiate (3.4) with respect to R, it is clear that ρ satisfies the radial equilibrium equation (2.6) on $(0,\infty)$.

In order to show that $s \mapsto s^{n-2}\hat{\Psi}_2(\rho'(s), \frac{\rho(s)}{s})$ is integrable on $(0,R)$ we use (3.3), (En3), $\rho' \geq 0$, and the fact that $J(s) \in [\lambda_0^n, \lambda_{\rm crit}^n]$ for each s to conclude that

$$\begin{aligned}s^{n-2}\left|\hat{\Psi}_2\left(\rho'(s), \frac{\rho(s)}{s}\right)\right| &= \rho(s)^{n-2}\left[\frac{s}{\rho(s)}\right]^{n-2}\left|\hat{\Psi}_2\left(J(s)\left[\frac{\rho(s)}{s}\right]^{1-n}, \frac{\rho(s)}{s}\right)\right|\\ &\leq K\rho(s)^{n-2}\left[\frac{\rho(s)}{s}\right]^{\beta-n+2} \leq K\rho(R)^\beta s^{n-2-\beta},\end{aligned}$$

which is clearly integrable on $(0,R)$ since $\beta < n-1$.

In order to prove that $s \mapsto s^{n-1}\hat{\Psi}_1(\rho'(s), \frac{\rho(s)}{s})$ is integrable on $(0,R)$, we will show that

$$\lim_{s\to 0^+} s^{n-1}\hat{\Psi}_1\left(\rho'(s), \frac{\rho(s)}{s}\right) = 0. \tag{4.4}$$

Now, by (3.3),

$$s^{n-1}\hat{\Psi}_1\left(\rho'(s), \frac{\rho(s)}{s}\right) = \rho(s)^{n-1}\left[\frac{s}{\rho(s)}\right]^{n-1}\hat{\Psi}_1\left(J(s)\left[\frac{\rho(s)}{s}\right]^{1-n}, \frac{\rho(s)}{s}\right). \tag{4.5}$$

Moreover, since $\lambda_0^n \le J(s) \le \lambda_{\text{crit}}^n$, hypothesis (En1) implies

$$\hat{\Psi}_1\left(\lambda_0^n t^{1-n}, t\right) \le \hat{\Psi}_1\left(J(s)t^{1-n}, t\right) \le \hat{\Psi}_1\left(\lambda_{\text{crit}}^n t^{1-n}, t\right), \quad t := \frac{\rho(s)}{s}. \tag{4.6}$$

Since $\rho \to 1$ and $t \to +\infty$ as $s \to 0^+$, (4.4) now follows from (En2), (4.5), and (4.6). In addition, the natural boundary condition (2.7) follows from (3.4), (4.4), together with the integrability of $s \mapsto s^{n-2}\hat{\Psi}_2(\rho'(s), \frac{\rho(s)}{s})$ and $s \mapsto s^{n-1}\hat{\Psi}_1(\rho'(s), \frac{\rho(s)}{s})$ on $(0,R)$.

Finally, we need to show that the right-hand side of (3.4) is bounded as $R \to \infty$ in order that h can be extended smoothly as a real-valued function on $(\lambda_{\text{crit}}^n, \infty)$. Let $R_1 > 0$. Then by (3.4)

$$\begin{aligned} h'(J(R)) - h'(J(R_1)) \;=\; & \int_{R_1}^{R} \frac{(n-1)s^{n-2}}{\rho(s)^{n-1}}\hat{\Psi}_2\left(\rho'(s), \frac{\rho(s)}{s}\right) ds \\ & - R^{n-1}\hat{\Psi}_1\left(\rho'(R), \frac{\rho(R)}{R}\right)\rho(R)^{1-n} \\ & + {R_1}^{n-1}\hat{\Psi}_1\left(\rho'(R_1), \frac{\rho(R_1)}{R_1}\right)\rho(R_1)^{1-n} \\ & + \int_{R_1}^{R} s^{n-1}\hat{\Psi}_1\left(\rho'(s), \frac{\rho(s)}{s}\right)\frac{d}{ds}\left[\rho(s)^{1-n}\right] ds \end{aligned} \tag{4.7}$$

for $R \in (R_1, \infty)$. Now, it is clear from Lemma 3.1(v) that $(4.7)_2$ is bounded as $R \to \infty$. Next, the sum of the integrals on the right-hand side of $(4.7)_1$, $(4.7)_4$ is equal to

$$(n-1)\int_{R_1}^{R} s^{-1}\left[\frac{s}{\rho(s)}\right]^n \left[\frac{\rho(s)}{s}\hat{\Psi}_2\left(\rho'(s), \frac{\rho(s)}{s}\right) - \rho'(s)\hat{\Psi}_1\left(\rho'(s), \frac{\rho(s)}{s}\right)\right] ds. \tag{4.8}$$

However, by Lemma 3.1, $\lambda_{\text{crit}} < \frac{\rho(s)}{s} \le \frac{\rho(R_1)}{R_1} =: \mu$ for $R_1 \le s < \infty$ and hence, in view of (En4), the absolute value of (4.8) is bounded by a constant times

$$\begin{aligned} \int_{R_1}^{R} \frac{1}{s}\left[\frac{\rho(s)}{s} - \rho'(s)\right] ds &= \int_{R_1}^{R} -\frac{d}{ds}\left[\frac{\rho(s)}{s}\right] ds \\ &= \frac{\rho(R_1)}{R_1} - \frac{\rho(R)}{R}, \end{aligned}$$

which is bounded as $R \to \infty$. ∎

5. The energy: Uniqueness. In this section we note that, whenever the function h is convex, the cavitating radial equilibrium solution we have constructed is the unique global minimizer of the energy among radial deformations. The following three results can be found in Sivaloganathan [13] (see also [14]).

PROPOSITION 5.1. *Assume that*

$$\hat{\Phi}_{11}(q,t) > 0 \tag{5.1}$$

for all $q > 0$ and $t > 0$. Let $r_c \in C^1([0,\infty)) \cap C^2((0,\infty))$ be a cavitating equilibrium solution*; i.e., r_c satisfies* (2.4) *on $(0,\infty)$,* (2.5)*, $r_c(0) > 0$, and $r_c' > 0$ a.e. Suppose that $R_o > 0$, and define $\lambda = r_c(R_o)/R_o$. Let $r \in \mathcal{A}_\lambda$,*

$$\mathcal{A}_\lambda := \{r \in W^{1,1}((0,R_o)) : r(R_o) = \lambda R_o,\ r(0) \geq 0,\ r' > 0 \text{ a.e.}\}, \tag{5.2}$$

satisfy

$$0 < \limsup_{R\to 0^+} \left[\frac{r(R)}{R}\right]. \tag{5.3}$$

Then

$$\overline{E}(r_c) < \overline{E}(r)$$

unless $r \equiv r_c$, where $\overline{E}$ is given by (2.3)*.*

COROLLARY 5.2. *Let $\lambda > 0$ and $R_o > 0$. Then, under the hypotheses of the previous proposition, there exists at most one cavitating equilibrium solution $r_c \in C^1([0,\infty)) \cap C^2((0,\infty))$ that satisfies $r_c(R_o) = \lambda R_o$.*

Remarks. 1. The statement of Theorem 6.8 in [13] actually requires that (5.3) be satisfied with lim sup replaced by lim inf. However, the remark after the proof of [13, Theorem 6.9] notes that the result remains valid under this weaker hypothesis.

2. Theorems 6.8 and 6.9 in [13] also appear to require the weakened Baker–Ericksen inequality $\mathcal{R}(q,t) \geq 0$, where $\mathcal{R}$ is given by (2.8). However, an examination of the proofs in [13, 14] shows that this inequality is used only to extend a solution of the radial equilibrium equation from $(0,R_o)$ to $(0,\infty)$, a step not needed in our presentation.

Proof of Corollary 5.2. Let $\lambda > 0$ and $R_o > 0$. If r_c is any cavitating equilibrium solution that satisfies $r_c(R_o) = \lambda R_o$, then $r_c \in \mathcal{A}_\lambda$ and r_c satisfies (5.3). Thus, by the previous proposition, two distinct cavitating equilibrium solutions r_{c_1} and r_{c_2} would satisfy $\overline{E}(r_{c_2}) < \overline{E}(r_{c_1})$ and $\overline{E}(r_{c_1}) < \overline{E}(r_{c_2})$, which is a contradiction. □

COROLLARY 5.3. *Let $\lambda > 0$ and $R_o > 0$. Suppose that $r_c \in C^1([0,\infty)) \cap C^2((0,\infty))$ is a cavitating equilibrium solution that satisfies $r_c(R_o) = \lambda R_o$. Then, under the hypotheses of the previous proposition, $\overline{E}(r_c) < \overline{E}(r_h)$, where $r_h(R) := \lambda R$.*

Proof. For any $\lambda > 0$ and $R_o > 0$ the homogeneous deformation $r_h(R) := \lambda R$ satisfies $r_h \in \mathcal{A}_\lambda$ and (5.3). The result then follows from Proposition 5.1. □

In order to make use of Proposition 5.1 we will need the following additional hypothesis on the energy:

(En5) there exist $\phi, \psi : (0,\infty) \to \mathbb{R}$ and $\Psi^* \in C((0,\infty)^n; \mathbb{R})$, with $\phi > 0, \psi \geq 0$, and $\Psi \geq 0$, that satisfy

$$\Psi(\nu_1,\nu_2,\ldots,\nu_n) = \sum_{i=1}^{n} \phi(\nu_i) + \sum_{i\neq j} \psi(\nu_i\nu_j) + \Psi^*(\nu_1,\nu_2,\ldots,\nu_n),$$

where $t \mapsto \Psi^*(t,t,\dots,t)$ is bounded on $(0,\lambda^*]$ and $\psi \equiv 0$ if $n=2$.

We now use Proposition 5.1 to show that if our equilibrium solutions are supersolutions for the energy Ψ, then they are energy minimizers among the radial deformations.

THEOREM 5.4. *Let Ψ satisfy* (En1)–(En5) *and let ρ be given by* (3.1) *and* (3.2)*, where $\lambda_0 > \lambda_*$ and $J' > 0$ on $[0,\infty)$. Suppose that*

$$\frac{d}{dR}\left[R^{n-1}\hat{\Psi}_1\left(\rho'(R),\frac{\rho(R)}{R}\right)\right] < (n-1)R^{n-2}\hat{\Psi}_2\left(\rho'(R),\frac{\rho(R)}{R}\right). \tag{5.4}$$

Finally, suppose in addition that $h \in C^2((0,\infty);(0,\infty))$ satisfies (3.4) *and $h''(s) \geq 0$ for $s \in (0,\lambda_0^n) \cup [\lambda_{\text{crit}}^n,\infty)$. Then the radial cavitating deformation $r_{\delta(\lambda)}$ given by Theorem* 3.5 *satisfies*

$$\overline{E}(r_{\delta(\lambda)}) < \overline{E}(r)$$

for every $r \in \mathcal{A}_\lambda$ (see (5.2)*).*

Proof. We first show that h'' is nonnegative on its domain of definition. Let $s \in (0,\infty)$. If $s \notin [\lambda_0^n,\lambda_{\text{crit}}^n)$, then $h''(s) \geq 0$, by hypothesis. If $s \in [\lambda_0^n,\lambda_{\text{crit}}^n)$, then by (3.1) there exists an $R \in [0,\infty)$ such that $J(R) = s$. Therefore, by (3.3) and the radial equilibrium equation (2.6),

$$\frac{d}{dR}\left[R^{n-1}\Psi_{,1}\right] - (n-1)R^{n-2}\Psi_{,2} = -\rho(R)^{n-1}h''\left(J(R)\right)J'(R). \tag{5.5}$$

The desired result now follows from (5.4), (5.5), and the assumed positivity of ρ and J'.

Now let $\lambda > \lambda_{\text{crit}}$, $r \in \mathcal{A}_\lambda$, and suppose that

$$0 < \limsup_{R\to 0^+}\left[\frac{r(R)}{R}\right]. \tag{5.6}$$

Then, by Proposition 5.1, all we need show is that (5.1) is satisfied. If we differentiate (2.1) twice with respect to ν_1 and set $\nu_1 = q$ and $\nu_2 = \nu_3 = \cdots = \nu_n = t$, we find that

$$\hat{\Phi}_{11}(q,t) = \hat{\Psi}_{11}(q,t) + t^{2n-2}h''(qt^{n-1}).$$

Consequently, in view of (En1), a sufficient condition for (5.1) is that h'' be nonnegative on its domain of definition, which has previously been shown.

Before proceeding further we note that the convexity of h, together with $h'(\lambda_0^n) = 0$ (see the remark following Theorem 3.3), implies that h is bounded below by $h(\lambda_0^n)$ on $(0,\infty)$.

Next, suppose alternatively that (5.6) is not satisfied and therefore that

$$0 = \lim_{R\to 0^+}\left[\frac{r(R)}{R}\right].$$

Then, in particular, $r(0) = 0$, and hence

$$\int_0^{R_o} nr'r^{n-1}dR = \int_0^{R_o}\frac{d}{dR}\left[r^n\right]dR = r(R_o)^n = \lambda^n R_o^n = n\int_0^{R_o}\lambda^n R^{n-1}dR. \tag{5.7}$$

Now, in view of the convexity of h,

$$h\left(r'(R)\left[\frac{r(R)}{R}\right]^n\right) \geq h(\lambda^n) + \left(r'(R)\left[\frac{r(R)}{R}\right]^n - \lambda^n\right)h'(\lambda^n),$$

and consequently, by (5.7),

$$\int_0^{R_o} h\left(r'(R)\left[\frac{r(R)}{R}\right]^{n-1}\right)R^{n-1}dR \geq \int_0^{R_o} h(\lambda^n)R^{n-1}dR. \tag{5.8}$$

If $\overline{E}(r) = +\infty$, we are done. If instead $\overline{E}(r) < \infty$, then by (En5) and since h is bounded below,

$$R \mapsto R^{n-1}\phi\left(\frac{r(R)}{R}\right) \in L^1((0, R_o)), \tag{5.9}$$

$$R \mapsto R^{n-1}\psi\left(\left[\frac{r(R)}{R}\right]^2\right) \in L^1((0, R_o)). \tag{5.10}$$

We claim that

$$0 = \liminf_{R\to 0^+} R^n\hat{\Psi}\left(\frac{r(R)}{R}, \frac{r(R)}{R}\right). \tag{5.11}$$

Otherwise, $R^n\hat{\Psi}\big(\frac{r(R)}{R}, \frac{r(R)}{R}\big) \geq K > 0$ for small R, and thus

$$R^{n-1}\hat{\Psi}\left(\frac{r(R)}{R}, \frac{r(R)}{R}\right) \geq \frac{K}{R}.$$

This inequality, together with (En5), implies

$$nR^{n-1}\phi\left(\frac{r(R)}{R}\right) + \frac{1}{2}n(n-1)R^{n-1}\psi\left(\left[\frac{r(R)}{R}\right]^2\right) \geq \frac{K}{R} - \hat{\Psi}^*\left(\frac{r(R)}{R}, \frac{r(R)}{R}\right),$$

which, since $r(R)/R$ and hence Ψ^* are bounded, contradicts (5.9) or (5.10).

Next, $\Psi_{,11} > 0$. Therefore $\hat{\Psi}(q,t) \geq \hat{\Psi}(t,t) + (q-t)\hat{\Psi}_{,1}(t,t)$, and hence

$$\begin{aligned}\hat{\Psi}\left(r'(R), \frac{r(R)}{R}\right) &\geq \hat{\Psi}\left(\frac{r(R)}{R}, \frac{r(R)}{R}\right) + \left(r'(R) - \frac{r(R)}{R}\right)\hat{\Psi}_{,1}\left(\frac{r(R)}{R}, \frac{r(R)}{R}\right)\\ &= \frac{1}{n}R^{1-n}\frac{d}{dR}\left[R^n\hat{\Psi}\left(\frac{r(R)}{R}, \frac{r(R)}{R}\right)\right],\end{aligned}$$

which, when multiplied by nR^{n-1} and integrated over (R_k, R_o), yields

$$\int_{R_k}^{R_o} n\hat{\Psi}\left(r'(R), \frac{r(R)}{R}\right)R^{n-1}dR \geq \left[R_o^n\hat{\Psi}(\lambda,\lambda)\right] - \left[R_k^n\hat{\Psi}\left(\frac{r(R_k)}{R_k}, \frac{r(R_k)}{R_k}\right)\right]. \tag{5.12}$$

In particular, choose a sequence $R_k \to 0^+$ as $k \to \infty$ so that $R_k^n\hat{\Psi}\big(\frac{r(R_k)}{R_k}, \frac{r(R_k)}{R_k}\big)$ converges to its lim inf, which is zero by (5.11). Then, if we let $k \to \infty$ in (5.12) and apply the dominated convergence theorem, we find that

$$\int_0^{R_o} \hat{\Psi}\left(r'(R), \frac{r(R)}{R}\right)R^{n-1}dR \geq \int_0^{R_o} \hat{\Psi}(\lambda,\lambda)\,R^{n-1}dR,$$

which, together with (2.1) and (5.8), yields $\overline{E}(r) \geq \overline{E}(r_h)$. The desired result now follows from Corollary 5.3. ∎

6. Examples.

6.1. Ogden materials. In order to illustrate the form that hypotheses (En1)–(En5) take for a well-analyzed class of materials, we now restrict our attention to three dimensions and consider materials whose constitutive relation is of the form

$$\Psi(\lambda_1,\lambda_2,\lambda_3)=\phi(\lambda_1)+\phi(\lambda_2)+\phi(\lambda_3)+\psi(\lambda_1\lambda_2)+\psi(\lambda_2\lambda_3)+\psi(\lambda_1\lambda_3), \tag{6.1}$$

where $\phi,\psi\in C^2((0,\infty))$. (Such constitutive relations were used by Ogden [8] to match theory with experiments.)

For such materials we make the following assumptions (cf. [1, 16, 13, 14, 9] and especially [17, 10]):

(Og1) for all $s>0$

$$\phi''(s)>0,\qquad \psi''(s)\geq 0;$$

(Og2) there exist $\beta\in[1,2)$, $\gamma\in[0,1)$, and $B>0$ such that for every $s>0$

$$|\phi'(s)|\leq B\left[s^{-\gamma}+s^{\beta}\right],\qquad |\psi'(s)|\leq B[s^{-\gamma}+s^{(\beta-1)/2}].$$

We now show that (Og1) and (Og2) imply (En1)–(En5). First, it is clear that (En5) is satisfied with $\Psi^*\equiv 0$. Next, we differentiate (6.1) with respect to λ_1 and let $\lambda_1=q$ and $\lambda_2=\lambda_3=t$ to get

$$\begin{aligned}\hat{\Psi}_1(q,t)&=\phi'(q)+2t\psi'(qt),\\ \hat{\Psi}_{11}(q,t)&=\phi''(q)+2t^2\psi''(qt).\end{aligned} \tag{6.2}$$

Then (Og1) and $(6.2)_2$ yield (En1). If we differentiate (6.1) with respect to λ_2 and let $\lambda_1=q$ and $\lambda_2=\lambda_3=t$, we get

$$\hat{\Psi}_2(q,t)=\phi'(t)+q\psi'(qt)+t\psi'(t^2), \tag{6.3}$$

and hence, when $q=\kappa t^{-2}$, we find that

$$\hat{\Psi}_2(\kappa t^{-2},t)=\phi'(t)+\kappa t^{-2}\psi'(\kappa t^{-1})+t\psi'(t^2). \tag{6.4}$$

In order to obtain (En3) we take the absolute value of (6.4) and use the triangle inequality and (Og2) to conclude that

$$|\hat{\Psi}_2(\kappa t^{-2},t)|\leq B[(t^{-\gamma}+t^{\beta})+(\kappa^{1-\gamma}t^{\gamma-2}+\kappa^{(\beta+1)/2}t^{-(\beta+3)/2})+(t^{1-2\gamma}+t^{\beta})],$$

which implies (En3). Similarly, if we take $q=\alpha t^{-2}$ in $(6.2)_1$ and use the triangle inequality and (Og2), we obtain

$$|t^{-2}\hat{\Psi}_1(\alpha t^{-2},t)|\leq B[\alpha^{-\gamma}t^{2(\gamma-1)}+\alpha^{\beta}t^{-2(\beta+1)}+2\alpha^{-\gamma}t^{\gamma-1}+2\alpha^{(\beta-1)/2}t^{-(\beta+1)/2}],$$

which approaches zero as $t\to\infty$ since $\gamma<1$. This implies (En2).

In order to obtain (En4) we first use $(6.2)_1$ and (6.3) to get

$$\mathcal{R}(q,t)=\frac{t\phi'(t)-q\phi'(q)}{t-q}+\frac{t^2\psi'(t^2)-qt\psi'(qt)}{t-q}. \tag{6.5}$$

We then fix $\mu>\lambda_*$ and consider two cases: $\frac{1}{2}\lambda_*\leq q<t<\mu$ and $0<q<\frac{1}{2}\lambda_*<\lambda_*<t<\mu$.

Case I. $0 < q < \frac{1}{2}\lambda_* < \lambda_* < t < \mu$. Then $|t-q|^{-1} \leq 2/\lambda_*$, and hence (6.5) together with (Og2) and the triangle inequality yield

$$|\mathcal{R}(q,t)| \leq \frac{2B}{\lambda_*}[t^{1-\gamma} + t^{\beta+1} + q^{1-\gamma} + q^{\beta+1} + t^{2(1-\gamma)} + t^{\beta+1} + (qt)^{1-\gamma} + (qt)^{(\beta+1)/2}],$$

which is bounded for $0 < q < t < \mu$ since $\gamma < 1$. This implies (En4) for small q.

Case II. $\frac{1}{2}\lambda_* \leq q < t < \mu$. Then (6.5) together with the mean-value theorem applied to the functions $\tilde{\phi}(s) := s\phi'(s)$ and $\tilde{\psi}(s) := s\psi'(s)$ yield

$$\mathcal{R}(q,t) = \tilde{\phi}'(c^*) + t\tilde{\psi}'(\hat{c}) \tag{6.6}$$

for some $c^* \in (q,t)$ and $\hat{c} \in (qt, t^2)$. Thus, since ϕ and ψ are C^2, the function $|\mathcal{R}|$ is bounded when $\frac{1}{2}\lambda_* \leq q < t < \mu$. Therefore (En4) is also valid for larger q.

6.2. Deformations. It is easy to construct radial cavitating *deformations*: the specification of a strictly monotone increasing radial Jacobian $J(R)$ that satisfies (3.1) suffices. The content of Theorem 3.3 is that such a deformation will satisfy the radial equilibrium equation for *every* stored energy function of the form (6.1) that satisfies (Og1) and (Og2), *provided* h is defined by (3.4). The difficulty is then ensuring that this equilibrium deformation is the unique radial minimizer of the energy, i.e., that the combination of deformation and stored energy satisfies (5.4).

One method of obtaining a family of such deformations is to perturb from a known solution. We illustrate this idea when the initial solution is isochoric. The resulting deformations will be nearly incompressible, as is expected in many elastomers (see, e.g., [2] or [8]), since the resulting energy will heavily penalize even small changes in volume. First, let's restrict our attention to the energy

$$\Psi(\lambda_1, \lambda_2, \lambda_3) := \phi(\lambda_1) + \phi(\lambda_2) + \phi(\lambda_3), \tag{6.7}$$

where ϕ satisfies (Og1), (Og2) and ϕ' is a convex function. (For example, $\phi(t) = t^{\beta+1}$ with $\beta \in [1,2)$.) Then (5.4) reduces to

$$\frac{d}{dR}\left[R^2\phi'(\rho'(R))\right] < 2R\phi'\left(\frac{\rho(R)}{R}\right)$$

or, equivalently,

$$R\phi''(\rho'(R))\rho''(R) < 2\left[\phi'\left(\frac{\rho(R)}{R}\right) - \phi'(\rho'(R))\right]. \tag{6.8}$$

However, the mapping $t \mapsto \phi'(t)$ is convex:

$$\phi'\left(\frac{\rho(R)}{R}\right) \geq \phi'(\rho'(R)) + \phi''(\rho'(R))\left[\frac{\rho(R)}{R} - \rho'(R)\right],$$

and so, in view of (6.8), a sufficient condition for (5.4) is that the function ρ satisfies

$$R\rho''(R) < 2\left[\frac{\rho(R)}{R} - \rho'(R)\right] \tag{6.9}$$

(or, equivalently, $\frac{d}{dR}\mathrm{trace}(\mathbf{F}) > 0$).

Now, let $\lambda_0 > 0$, and suppose that $\theta : [0,1] \to [0,1]$ is continuous with $\theta > 0$ on $(0,1]$ and that

$$\theta(s) = o(s^6) \quad \text{as} \quad s \to 0^+. \tag{6.10}$$

Then for each $s \in [0,1]$ define

$$\rho(R,s)^3 := 1 + 3\int_0^R J(t,s)t^2\,dt, \tag{6.11}$$

$$J(R,s) := \lambda_0^3 + \theta(s)(1 - e^{-sR}) \tag{6.12}$$

and note that

$$J_R(R,s) = s\theta(s)e^{-sR}, \qquad s\theta(s)\int_0^\infty t^3 e^{-st}\,dt = \frac{6\theta(s)}{s^3}, \tag{6.13}$$

where the subscript R denotes the partial derivative with respect to R. (If $\lambda_0 = 1$, the deformation at $s = 0$ is isochoric.) By (6.13), for each $s > 0$, $R \mapsto J(R,s)$ is strictly monotone increasing and satisfies $(3.1)_3$, provided $0 < \theta(s) \le s^3/6$, which is a consequence of (6.10) for all sufficiently small s.

Next, by (4.2), (4.3), and $(6.13)_1$,

$$R\rho''(R)\left[\frac{\rho(R)}{R}\right]^2 = s\theta(s)Re^{-sR} - 2\rho'(R)\left[\frac{\rho(R)}{R}\right]\left[\rho'(R) - \frac{\rho(R)}{R}\right],$$

which shows that (6.9) is equivalent to

$$s\theta(s)Re^{-sR} < 2\left[\frac{\rho(R)}{R}\right]\left[\frac{\rho(R)}{R} - \rho'(R)\right]^2. \tag{6.14}$$

However, in view of (4.1), (4.2), and $(6.13)_1$,

$$R^6\left[\frac{\rho(R)}{R}\right]^4\left[\frac{\rho(R)}{R} - \rho'(R)\right]^2 = \left[1 - s\theta(s)\int_0^R t^3e^{-st}\,dt\right]^2,$$

so that (6.14) is the same as

$$s\theta(s)R^7\left[\frac{\rho(R)}{R}\right]^3 < 2e^{sR}\left[1 - s\theta(s)\int_0^R t^3e^{-st}\,dt\right]^2. \tag{6.15}$$

Finally, we note that in view of (6.10), $\theta(s) \le s^3/12$ for s sufficiently small, and consequently, by $(6.13)_2$,

$$s\theta(s)\int_0^R t^3e^{-st}\,dt \le \frac{1}{2} \tag{6.16}$$

for small s. In addition, $e^t \ge (1+t^7)/7!$ and hence, upon multiplying (6.15) by $(sR)^{-7}$ and making use of (6.16), it suffices to show

$$2\frac{\theta(s)}{s^6}\left[\frac{\rho(R)}{R}\right]^3 < \frac{1}{7!}\left[(sR)^{-7} + 1\right] \tag{6.17}$$

in order to obtain (6.15). However, for small s, (6.17) is a consequence of (6.10)–(6.12), which completes the example.

REFERENCES

[1] J. M. Ball, *Discontinuous equilibrium solutions and cavitation in nonlinear elasticity*, Philos. Trans. Roy. Soc. London Ser. A, 306 (1982), pp. 557–611.

[2] A. N. Gent and P. B. Lindley, *Internal rupture of bonded rubber cylinders in tension*, Proc. Roy. Soc. London Ser. A, 249 (1958), pp. 195–205.

[3] C. O. Horgan, *Void nucleation and growth for compressible non-linearly elastic materials: An example*, Internat. J. Solids Structures, 29 (1992), pp. 279–291.

[4] C. O. Horgan and R. Abeyaratne, *A bifurcation problem for a compressible nonlinearly elastic medium: Growth of a microvoid*, J. Elasticity, 16 (1986), pp. 189–200.

[5] C. O. Horgan and D. A. Polignone, *Cavitation in nonlinearly elastic solids: A review*, Appl. Mech. Rev., 48 (1995), pp. 471–485.

[6] M. R. Lancia, P. Podio-Guidugli, and G. Vergara Caffarelli, *Gleanings of radial cavitation*, J. Elasticity, 44 (1996), pp. 183–192.

[7] J. G. Murphy and S. Biwa, *Nonmonotonic cavity growth in finite, compressible elasticity*, Internat. J. Solids Structures, 34 (1997), pp. 3859–3872.

[8] R. W. Ogden, *Large deformation isotropic elasticity: On the correlation of theory and experiment for compressible rubberlike solids*, Proc. Roy. Soc. London Ser. A, 328 (1972), pp. 567–583.

[9] K. A. Pericak-Spector and S. J. Spector, *Nonuniqueness for a hyperbolic system: Cavitation in elastodynamics*, Arch. Ration. Mech. Anal., 101 (1988), pp. 293–317.

[10] K. A. Pericak-Spector and S. J. Spector, *Dynamic cavitation with shocks in nonlinear elasticity*, Proc. Roy. Soc. Edinburgh Sect. A, 127 (1997), pp. 837–857.

[11] K. A. Pericak-Spector, J. Sivaloganathan, and S. J. Spector, *An explicit radial cavitation solution in nonlinear elasticity*, Math. Mech. Solids, 7 (2002), pp. 87–93.

[12] X.-C. Shang and C.-J. Cheng, *Exact solution for cavitated bifurcation for compressible hyperelastic materials*, Internat. J. Engrg. Sci., 39 (2001), pp. 1101–1117.

[13] J. Sivaloganathan, *A field theory approach to stability of radial equilibria in nonlinear elasticity*, Math. Proc. Cambridge Philos. Soc., 99 (1986), pp. 589–604.

[14] J. Sivaloganathan, *Uniqueness of regular and singular equilibria for spherically symmetric problems of nonlinear elasticity*, Arch. Ration. Mech. Anal., 96 (1986), pp. 97–136.

[15] D. J. Steigmann, *Cavitation in membranes—An example*, J. Elasticity, 28 (1992), pp. 277–287.

[16] C. A. Stuart, *Radially symmetric cavitation for hyperelastic materials*, Anal. Nonlinéaire, 2 (1985), pp. 33–66.

[17] C. A. Stuart, *Estimating the critical radius for radially symmetric cavitation*, Quart. Appl. Math., 51 (1993), pp. 251–263.

[18] H. Tian-hu, *A theory of the appearance and growth of the micro-spherical void*, Internat. J. Fracture, 43 (1990), pp. R51–R55.

SIAM J. APPL. MATH.
Vol. 63, No. 5, pp. 1475–1498

TIME REVERSAL AND REFOCUSING IN RANDOM MEDIA*

GUILLAUME BAL† AND LEONID RYZHIK‡

Abstract. In time reversal acoustics experiments, a signal is emitted from a localized source, recorded at an array of receivers, time reversed, and finally reemitted into the medium. A celebrated feature of time reversal experiments is that the refocusing of the reemitted signals at the location of the initial source is improved when the medium is heterogeneous. Contrary to intuition, multiple scattering enhances the spatial resolution of the refocused signal and allows one to beat the diffraction limit obtained in homogeneous media. This paper presents a quantitative explanation of time reversal and other more general refocusing phenomena for general classical waves in heterogeneous media. The theory is based on the asymptotic analysis of the Wigner transform of wave fields in the high frequency limit. Numerical experiments complement the theory.

Key words. waves in random media, time reversal, refocusing, radiative transfer equations, diffusion approximation

AMS subject classifications. 35F10, 35B40, 82D30

DOI. 10.1137/S0036139902401082

1. Introduction.

In time reversal experiments, acoustic waves are emitted from a localized source, recorded in time by an array of receiver-transducers, time reversed, and retransmitted into the medium, so that the signals recorded first are reemitted last and vice versa [7, 8, 13, 16, 18, 19]. The retransmitted signal refocuses at the location of the original source, with a modified shape that depends on the array of receivers. The salient feature of these time reversal experiments is that refocusing is much better when wave propagation occurs in complicated environments than in homogeneous media. Time reversal techniques with improved refocusing in heterogeneous media have found important applications in medicine, nondestructive testing, underwater acoustics, and wireless communications (see the above references). It has also been applied to imaging in weakly random media [4, 13].

A schematic depiction of the time reversal procedure is given in Figure 1.1. Early experiments in time reversal acoustics are described in [7]; see also the more recent papers [11, 12, 13]. A very qualitative explanation for the better refocusing observed in heterogeneous media is based on *multipathing*. Since waves can scatter off a larger number of heterogeneities, more paths coming from the source reach the recording array, thus more is known about the source by the transducers than in a homogeneous medium. The heterogeneous medium plays the role of a lens that widens the aperture through which the array of receivers sees the source. Refocusing is also qualitatively justified by ray theory (geometrical optics). The phase shift caused by multiple scattering is exactly compensated when the time reversed signal follows the same path back to the source location. This phase cancellation happens only at the source location. The phase shift along paths leading to other points in space is essentially

*Received by the editors January 16, 2002; accepted for publication (in revised form) November 4, 2002; published electronically June 12, 2003. This work was supported in part by ONR grant N00014-02-1-0089.

http://www.siam.org/journals/siap/63-5/40108.html

†Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY 10027 (gb2030@columbia.edu). The research of this author was supported in part by NSF grant DMS-0072008.

‡Department of Mathematics, University of Chicago, Chicago, IL 60637 (ryzhik@math.uchicago.edu). The research of this author was supported in part by NSF grant DMS-9971742.

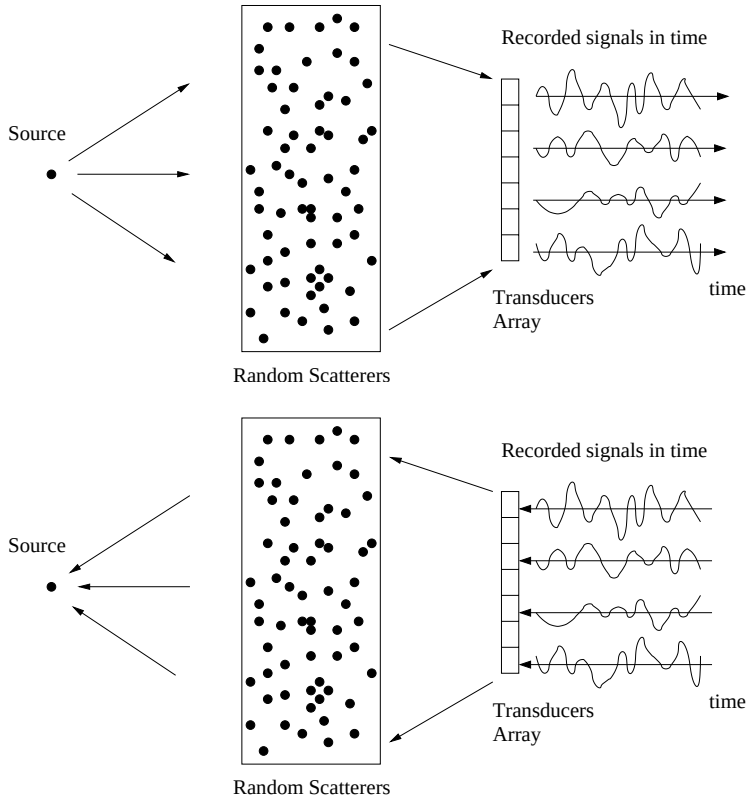


FIG. 1.1. *The time reversal procedure. Top: Propagation of signal and measurements in time. Bottom: Time reversal of recorded signals and back-propagation into the medium.*

random. The interference of multiple paths will thus be constructive at the source location and destructive anywhere else. This explains why refocusing at the source location is improved when the number of scatterers is large.

As convincing as they are, the above explanations remain qualitative and do not allow us to quantify how the refocused signal is modified by the time reversal procedure. Quantitative justifications require more careful analysis of wave propagation. The first quantitative description of time reversal was obtained in [5] in the framework of randomly layered media (see also the recent work [10]). That paper provides the first mathematical explanation of two of the most prominent features of time reversal: heterogeneities improve refocusing, and refocusing occurs for almost every realization of the random medium. The first multidimensional quantitative description of time reversal was obtained in [3] for the parabolic approximation, i.e., for waves that propagate in a privileged direction with no backscattering (see also [23, 24] for further analysis of time reversal in this regime). That paper shows that the random medium indeed plays the role of a lens. The back-propagated signal behaves as if the initial array were replaced by another one with a much bigger effective aperture. In a slightly different context, a recent paper [2] analyzes time reversal in ergodic cavities. There, wave mixing is created by reflection at the boundary of a chaotic cavity, which plays a role similar to that of the heterogeneities in a heterogeneous medium.

This paper generalizes the results of [3] to the case of general classical waves propagating in weakly fluctuating random media. The main results are briefly sum-

marized as follows. We first show that refocusing in time reversal experiments may be understood in the following three-step more general framework:

(i) A signal propagating from a localized source is recorded at a single time $T > 0$ by an array of receivers.
(ii) The recorded signal is processed at the array location.
(iii) The processed signal is emitted from the array and propagates in the *same* medium for the same time duration T.

The first main result is that the resulting signal will refocus at the location of the original source for a large class of waves and a large class of processings. The experiments described above correspond to the specific processing of acoustic waves in which pressure is kept unchanged and the sign of the velocity field is reversed.

The second main result is a quantitative description of the repropagated signal at time T. We show that the repropagated signal $\mathbf{u}^B(\boldsymbol{\xi})$ at a point $\boldsymbol{\xi}$ near the source location can be written in the high frequency limit as the following convolution of the original source $\mathbf{S}$:

$$\mathbf{u}^B(\boldsymbol{\xi}) = (F * \mathbf{S})(\boldsymbol{\xi}). \tag{1.1}$$

The kernel F depends on the location of the recording array and on the signal processing. The quality of the refocusing depends on the spatial decay of F. It turns out that F can be expressed in terms of the Wigner transform [25] of two wave fields. The decay properties of F depend on the smoothness of the Wigner transform in the phase space. The Wigner transform in random media has been extensively studied [9, 25, 27], especially in the high frequency regime, when the wavelength of the initial signal is small compared to the distance of propagation. It satisfies a radiative transport equation, which is used to describe the evolution of the energy density of waves in random media [17, 25, 26, 27]. The transport equations possess a smoothing effect so that the Wigner distribution becomes less singular in random media, which implies a stronger decay of the convolution kernel F and a better refocusing. The diffusion approximation to the radiative transport equations provides simple reconstruction formulas that can be used to quantify the refocusing quality of the back-propagated signal. This construction applies to a large class of classical waves—acoustic, electromagnetic, elastic, and others—and allows for a large class of signal processings at the recording array.

Some results of this paper have been announced in [1]. The concept of single-time time reversal emerged during early discussions with Knut Solna. We also stress that the important property of self-averaging of the time reversed signal (the refocused signal is almost independent of the realization of the random medium) is not analyzed in this paper. A formal explanation is given in [3, 23, 24] in the parabolic approximation. Self-averaging for classical waves will be addressed elsewhere.

This paper is organized as follows. Section 2 recalls the classical setting of time reversal and introduces single-time time reversal. The retransmitted signal and its relation to the Wigner transform are analyzed in section 3. A quantitative description of acoustic wave refocusing in weakly fluctuating random media is obtained by asymptotic analysis; see (3.42) and (3.43) for an explicit expression in the diffusion approximation. Section 4 generalizes the results in two ways. First, a more general signal processing at the recording array is allowed, such as recording only the pressure field of acoustic waves and not the velocity field. Second, the retransmission scheme is applied to more general waves, and the role of polarization and mode coupling is explained.

2. Classical time reversal and single-time time reversal. Propagation of acoustic waves is described by a system of equations for the pressure $p(t,\mathbf{x})$ and acoustic velocity $\mathbf{v}(t,\mathbf{x})$:

$$\begin{aligned}\rho(\mathbf{x})\frac{\partial \mathbf{v}}{\partial t}+\nabla p=0,\\ \kappa(\mathbf{x})\frac{\partial p}{\partial t}+\nabla\cdot\mathbf{v}=0,\end{aligned} \tag{2.1}$$

with suitable initial conditions and where $\rho(\mathbf{x})$ and $\kappa(\mathbf{x})$ are density and compressibility of the underlying medium, respectively. These equations can be recast as the following linear hyperbolic system:

$$A(\mathbf{x})\frac{\partial \mathbf{u}}{\partial t}+D^j\frac{\partial \mathbf{u}}{\partial x^j}=0,\quad \mathbf{x}\in\mathbb{R}^3, \tag{2.2}$$

with the vector $\mathbf{u}=(\mathbf{v},p)\in\mathbb{C}^4$. The matrix $A=\mathrm{Diag}(\rho,\rho,\rho,\kappa)$ is positive definite. The 4×4 matrices D^j, $j=1,2,3$, are symmetric and given by $D^j_{mn}=\delta_{m4}\delta_{nj}+\delta_{n4}\delta_{mj}$. We use the Einstein convention of summation over repeated indices.

The time reversal experiments in [7] consist of two steps. First, the direct problem

$$\begin{aligned}&A(\mathbf{x})\frac{\partial \mathbf{u}}{\partial t}+D^j\frac{\partial \mathbf{u}}{\partial x^j}=0,\quad 0\le t\le T,\\ &\mathbf{u}(0,\mathbf{x})=\mathbf{S}(\mathbf{x}),\end{aligned} \tag{2.3}$$

with a localized source $\mathbf{S}$ centered at a point $\mathbf{x}_0$, is solved. The signal is recorded during the period of time $0\le t\le T$ by an array of receivers located at $\Omega\subset\mathbb{R}^3$. Second, the signal is time reversed and reemitted into the medium. Time reversal is described by multiplying $\mathbf{u}=(\mathbf{v},p)$ by the matrix $\Gamma=\mathrm{Diag}(-1,-1,-1,1)$. The back-propagated signal solves

$$\begin{aligned}&\frac{\partial \mathbf{u}}{\partial t}+A^{-1}(\mathbf{x})D^j\frac{\partial \mathbf{u}}{\partial x^j}=\frac{1}{T}\mathbf{R}(2T-t,\mathbf{x}),\quad T\le t\le 2T,\\ &\mathbf{u}(T,\mathbf{x})=0,\end{aligned} \tag{2.4}$$

with the source term

$$\mathbf{R}(t,\mathbf{x})=\Gamma\mathbf{u}(t,\mathbf{x})\chi(\mathbf{x}). \tag{2.5}$$

The function $\chi(\mathbf{x})$ is either the characteristic function of the set where the recording array is located, or some other function that allows for possibly space-dependent amplification of the retransmitted signal.

The back-propagated signal is then given by $\mathbf{u}(2T,\mathbf{x})$. We can decompose it as

$$\mathbf{u}(2T,\mathbf{x})=\frac{1}{T}\int_0^T ds\ \mathbf{w}(s,\mathbf{x};s), \tag{2.6}$$

where the vector-valued function $\mathbf{w}(t,\mathbf{x};s)$ solves the initial value problem

$$\begin{aligned}&A(\mathbf{x})\frac{\partial \mathbf{w}(t,\mathbf{x};s)}{\partial t}+D^j\frac{\partial \mathbf{w}(t,\mathbf{x};s)}{\partial x^j}=0,\quad 0\le t\le s,\\ &\mathbf{w}(0,\mathbf{x};s)=\mathbf{R}(s,\mathbf{x}).\end{aligned}$$

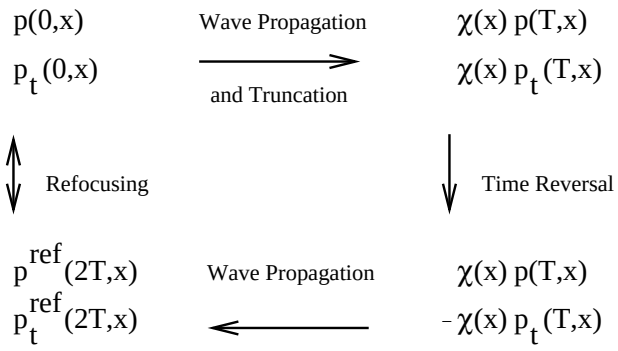


FIG. 2.1. *The single-time time reversal procedure. Here, p_t denotes $\frac{\partial p}{\partial t}$.*

We deduce from (2.6) that it is sufficient to analyze the refocusing properties of $\mathbf{w}(s,\mathbf{x};s)$ for $0 \le s \le T$ to obtain those of $\mathbf{u}(2T,\mathbf{x})$. For a fixed value of s, we call the construction of $\mathbf{w}(s,\mathbf{x};s)$ single-time time reversal.

We define single-time time reversal more generally as follows. The direct problem (2.3) is solved until time $t = T$ to yield $\mathbf{u}(T^-,\mathbf{x})$. At time T, the signal is recorded and processed. The processing is modeled by an amplification function $\chi(\mathbf{x})$, a blurring kernel $f(\mathbf{x})$, and a (possibly spatially varying) time reversal matrix Γ. After processing, we have

$$\mathbf{u}(T^+,\mathbf{x}) = \Gamma(f * (\chi\mathbf{u}))(T^-,\mathbf{x})\chi(\mathbf{x}). \tag{2.7}$$

The processed signal then propagates for the same time duration T:

$$\begin{aligned} &A(\mathbf{x})\frac{\partial \mathbf{u}}{\partial t} + D^j\frac{\partial \mathbf{u}}{\partial x^j} = 0, \quad T \le t \le 2T, \\ &\mathbf{u}(T^+,\mathbf{x}) = \Gamma(f * (\chi\mathbf{u}))(T^-,\mathbf{x})\chi(\mathbf{x}). \end{aligned} \tag{2.8}$$

The main questions are whether $\mathbf{u}(2T,\mathbf{x})$ refocuses at the location of the original source $\mathbf{S}(\mathbf{x})$ and how the original signal has been modified by the time reversal procedure. Notice that in the case of full ($\Omega = \mathbb{R}^3$) and exact ($f(\mathbf{x}) = \delta(\mathbf{x})$) measurements with $\Gamma = \mathrm{Diag}(-1,-1,-1,1)$, the time-reversibility of first-order hyperbolic systems implies that $\mathbf{u}(2T,\mathbf{x}) = \Gamma\mathbf{S}(\mathbf{x})$, which corresponds to exact refocusing. When only partial measurements are available, we shall see in the following sections that $\mathbf{u}(2T,\mathbf{x})$ is closer to $\Gamma\mathbf{S}(\mathbf{x})$ when propagation occurs in a heterogeneous medium than in a homogeneous medium.

The pressure field $p(t,\mathbf{x})$ satisfies the following scalar wave equation:

$$\frac{\partial^2 p}{\partial t^2} - \frac{1}{\kappa(\mathbf{x})}\nabla\cdot\left(\frac{1}{\rho(\mathbf{x})}\nabla p\right) = 0. \tag{2.9}$$

A schematic description of the single-time procedure for the wave equation is presented in Figure 2.1. This is the equation solved in the numerical experiments presented in this paper. The details of the numerical setting are described in the appendix. A numerical experiment for the single-time time reversal procedure is shown in Figure 2.2. In the numerical simulations, there is no blurring, $f(\mathbf{x}) = \delta(\mathbf{x})$, and the array of receivers is the domain $\Omega = (-1/6, 1/6)^2$. ($\chi(\mathbf{x})$ is the characteristic function of Ω.) Note that the truncated signal does not retain much information about the

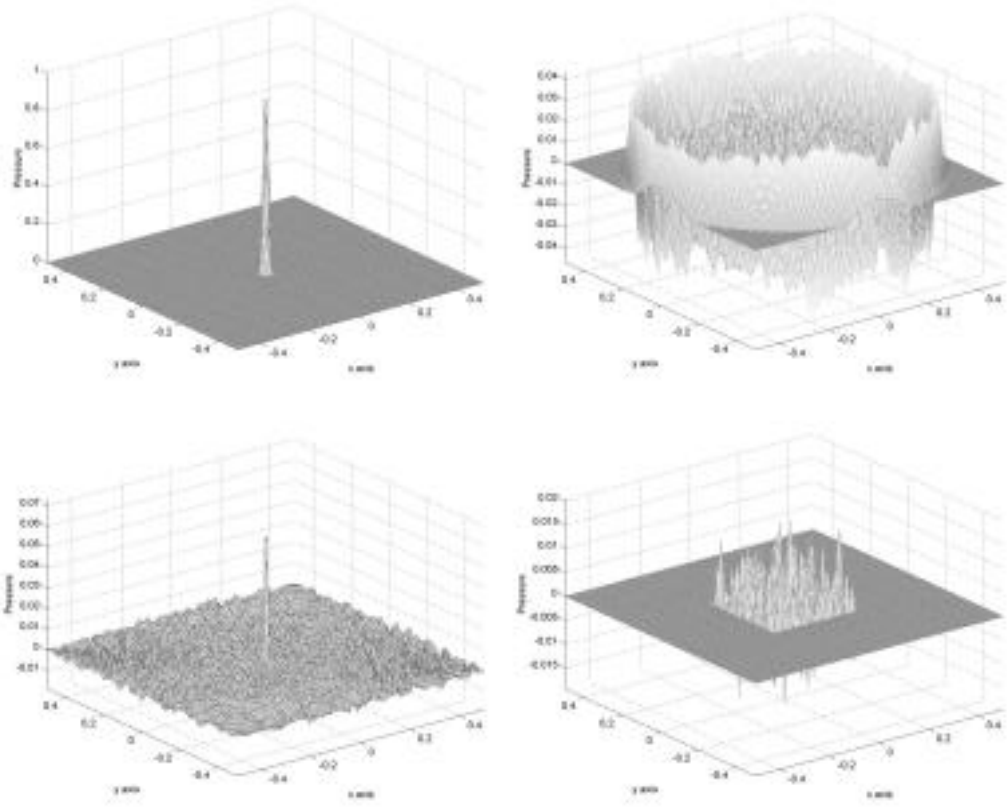


FIG. 2.2. *Numerical experiment using the single-time time reversal procedure. Top left: initial condition* $p(0,\mathbf{x})$*, a peaked Gaussian of maximal amplitude equal to* 1*. Top right: forward solution* $p(T^-,\mathbf{x})$*, of maximal amplitude* 0.04*. Bottom right: recorded solution* $p(T^+,\mathbf{x})$*, of maximal amplitude* 0.015 *on the domain* $\Omega = (-1/6,1/6)^2$*. Bottom left: back-propagated solution* $p(2T,\mathbf{x})$*, of maximal amplitude* 0.07*.*

ballistic part of the original wave (the part that propagates without scattering with the underlying medium). If we were in three space dimensions, the truncated signal in a homogeneous medium would even be identically zero, and no refocusing would be observed. The interesting aspect of time reversal is that a coherent signal emerges at time $2T$ out of a signal at time T^+ that seems to have no useful information.

3. Theory of time reversal in random media. Our objective now is to present a theory that explains in a quantitative manner the refocusing properties described in the preceding sections. We consider here the single-time time reversal for acoustic waves. Generalizations to other types of waves and more general processings in (2.8) are given in section 4.

3.1. Refocused signal. We recall that the single-time time reversal procedure consists of letting an initial pulse $\mathbf{S}(\mathbf{x})$ propagate according to (2.3) until time T,

$$\mathbf{u}(T^-,\mathbf{x}) = \int_{\mathbb{R}^3} G(T,\mathbf{x};\mathbf{z})\mathbf{S}(\mathbf{z})d\mathbf{z},$$

where $G(T,\mathbf{x};\mathbf{z})$ is the Green's matrix solution of

$$A(\mathbf{x})\frac{\partial G(t,\mathbf{x};\mathbf{y})}{\partial t}+D^j\frac{\partial G(t,\mathbf{x};\mathbf{y})}{\partial x^j}=0,\quad 0\le t\le T,\tag{3.1}$$
$$G(0,\mathbf{x};\mathbf{y})=I\delta(\mathbf{x}-\mathbf{y}).$$

At time T, the "intelligent" array reverses the signal. For acoustic pulses, this means keeping pressure unchanged and reversing the sign of the velocity field. The array of receivers is located in $\Omega\subset\mathbb{R}^3$. The amplification function $\chi(\mathbf{x})$ is an arbitrary bounded function supported in Ω, such as its characteristic function ($\chi(\mathbf{x})=1$ for $\mathbf{x}\in\Omega$ and $\chi(\mathbf{x})=0$ otherwise) when all transducers have the same amplification factor. We also allow for some blurring of the recorded data modeled by a convolution with a function $f(\mathbf{x})$. The case $f(\mathbf{x})=\delta(\mathbf{x})$ corresponds to exact measurements. Finally, the signal is time reversed, that is, the direction of the acoustic velocity is reversed. Here, the operator Γ in (2.7) is simply multiplication by the matrix

$$\Gamma=\mathrm{Diag}(-1,-1,-1,1).\tag{3.2}$$

The signal at time T^+ after time reversal then takes the form

$$\mathbf{u}(T^+,\mathbf{x})=\int_{\mathbb{R}^6}\Gamma G(T,\mathbf{y}';\mathbf{z})\chi(\mathbf{x})\chi(\mathbf{y}')f(\mathbf{x}-\mathbf{y}')\mathbf{S}(\mathbf{z})d\mathbf{z}d\mathbf{y}'.\tag{3.3}$$

The last step (2.8) consists of letting the time reversed field propagate through the random medium until time $2T$. To compare this signal with the initial pulse $\mathbf{S}$, we need to reverse the acoustic velocity once again and define

$$\mathbf{u}^B(\mathbf{x})=\Gamma\mathbf{u}(2T,\mathbf{x})=\int_{\mathbb{R}^9}\Gamma G(T,\mathbf{x};\mathbf{y})\Gamma G(T,\mathbf{y}';\mathbf{z})\chi(\mathbf{y})\chi(\mathbf{y}')f(\mathbf{y}-\mathbf{y}')\mathbf{S}(\mathbf{z})d\mathbf{y}d\mathbf{y}'d\mathbf{z}.\tag{3.4}$$

The time-reversibility of first-order hyperbolic systems implies that $\mathbf{u}^B(\mathbf{x})=\mathbf{S}(\mathbf{x})$ when $\Omega=\mathbb{R}^3$, $\chi\equiv 1$, and $f(\mathbf{x})=\delta(\mathbf{x})$, that is, when full and nondistorted measurements are available. It remains to understand which features of $\mathbf{S}$ are retained by $\mathbf{u}^B(\mathbf{x})$ when only partial measurement is available.

3.2. Localized source and scaling. We consider an asymptotic solution of the time reversal problem (2.3), (2.8) when the support λ of the initial pulse $\mathbf{S}(\mathbf{x})$ is much smaller than the distance L of propagation between the source and the recording array: $\varepsilon=\lambda/L\ll 1$. We also take the size a of the array comparable to L: $a/L=O(1)$. We assume that the time T between the emission of the original signal and recording is of order L/c_0, where c_0 is a typical speed of propagation of the acoustic wave. We consequently consider the initial pulse to be of the form

$$\mathbf{u}(0,\mathbf{x})=\mathbf{S}\left(\frac{\mathbf{x}-\mathbf{x}_0}{\varepsilon}\right)$$

in nondimensionalized variables $\mathbf{x}'=\mathbf{x}/L$ and $t'=t/(L/c_0)$. We drop primes to simplify notation. Here $\mathbf{x}_0$ is the location of the source. The transducers obviously have to be capable of capturing signals of frequency ε^{-1}, and blurring should happen on the scale of the source, so we replace $f(\mathbf{x})$ by $\varepsilon^{-3}f(\varepsilon^{-1}\mathbf{x})$. Finally, we are interested

in the refocusing properties of $\mathbf{u}^B(\mathbf{x})$ in the vicinity of $\mathbf{x}_0$. We therefore introduce the scaling $\mathbf{x} = \mathbf{x}_0 + \varepsilon\boldsymbol{\xi}$. With these changes of variables, expression (3.4) is recast as

$$\begin{aligned}\mathbf{u}^B(\boldsymbol{\xi};\mathbf{x}_0) &= \Gamma\mathbf{u}(2T,\mathbf{x}_0+\varepsilon\boldsymbol{\xi}) \\ &= \int_{\mathbb{R}^9} \Gamma G(T,\mathbf{x}_0+\varepsilon\boldsymbol{\xi};\mathbf{y})\Gamma G(T,\mathbf{y}';\mathbf{x}_0+\varepsilon\mathbf{z})\chi(\mathbf{y},\mathbf{y}')\mathbf{S}(\mathbf{z})d\mathbf{y}d\mathbf{y}'d\mathbf{z},\end{aligned} \tag{3.5}$$

where

$$\chi(\mathbf{y},\mathbf{y}') = \chi(\mathbf{y})\chi(\mathbf{y}')f\left(\frac{\mathbf{y}-\mathbf{y}'}{\varepsilon}\right). \tag{3.6}$$

In what follows we will also allow the medium to vary on a scale comparable to the source scale ε. Thus the Green's function G and the matrix A depend on ε. To simplify notation we do not make this dependence explicit. We are interested in the limit of $\mathbf{u}^B(\boldsymbol{\xi};\mathbf{x}_0)$ as $\varepsilon \to 0$. The scaling considered here is well adapted both to the physical experiments in [7] and the numerical experiments in Figure 2.2.

3.3. Adjoint Green's function. The analysis of the repropagated signal relies on the study of the two point correlation at nearby points of the Green's matrix in (3.5). There are two undesirable features in (3.5). First, the two nearby points $\mathbf{x}_0+\varepsilon\boldsymbol{\xi}$ and $\mathbf{x}_0+\varepsilon\mathbf{z}$ are terminal and initial points in their respective Green's matrices. Second, one would like to have the product of two Green's functions, with no matrix Γ in between. However, Γ and G do not commute. For these reasons, we introduce the *adjoint* Green's matrix, solution of

$$\begin{aligned}&\frac{\partial G_*(t,\mathbf{x};\mathbf{y})}{\partial t}A(\mathbf{x}) + \frac{\partial G_*(t,\mathbf{x};\mathbf{y})}{\partial x^j}D^j = 0, \\ &G_*(0,\mathbf{x};\mathbf{y}) = A^{-1}(\mathbf{x})\delta(\mathbf{x}-\mathbf{y}).\end{aligned} \tag{3.7}$$

We now prove that

$$G_*(t,\mathbf{x};\mathbf{y}) = \Gamma G(t,\mathbf{y};\mathbf{x})A^{-1}(\mathbf{x})\Gamma. \tag{3.8}$$

Note that for all initial data $\mathbf{S}(\mathbf{x})$, the solution $\mathbf{u}(t,\mathbf{x})$ of (2.3) satisfies

$$\mathbf{u}(t,\mathbf{x}) = \int_{\mathbb{R}^3} G(t-s,\mathbf{x};\mathbf{y})\mathbf{u}(s,\mathbf{y})d\mathbf{y}$$

for all $0 \le s \le t \le T$ since the coefficients in (2.3) are time-independent. Differentiating the above with respect to s and using (2.3) yields

$$0 = \int_{\mathbb{R}^3}\left(-\frac{\partial G(t-s,\mathbf{x};\mathbf{y})}{\partial t}\mathbf{u}(s,\mathbf{y}) - G(t-s,\mathbf{x};\mathbf{y})A^{-1}(\mathbf{y})D^j\frac{\partial\mathbf{u}(s,\mathbf{y})}{\partial y^j}\right)d\mathbf{y}.$$

Upon integrating by parts and letting $s=0$, we get

$$0 = \int_{\mathbb{R}^3}\left(-\frac{\partial G(t,\mathbf{x};\mathbf{y})}{\partial t} + \frac{\partial}{\partial y^j}\left[G(t,\mathbf{x};\mathbf{y})A^{-1}(\mathbf{y})D^j\right]\right)\mathbf{S}(\mathbf{y})d\mathbf{y}.$$

Since the above relation holds for all test functions $\mathbf{S}(\mathbf{y})$, we deduce that

$$\frac{\partial G(t,\mathbf{x};\mathbf{y})}{\partial t} - \frac{\partial}{\partial y^j}\left[G(t,\mathbf{x};\mathbf{y})A^{-1}(\mathbf{y})D^j\right] = 0. \tag{3.9}$$

Interchanging $\mathbf{x}$ and $\mathbf{y}$ in the above equation and multiplying it on the left and the right by Γ, we obtain that

$$\frac{\partial}{\partial t}\left[\Gamma G(t,\mathbf{y};\mathbf{x})A^{-1}(\mathbf{x})\right]A(\mathbf{x})\Gamma-\frac{\partial}{\partial x^j}\left[\Gamma G(t,\mathbf{y};\mathbf{x})A^{-1}(\mathbf{x})\right]D^j\Gamma=0. \tag{3.10}$$

We remark that with the choice of Γ in (3.2) we have

$$\Gamma D^j=-D^j\Gamma \qquad \text{and} \qquad \Gamma A(\mathbf{x})=A(\mathbf{x})\Gamma, \tag{3.11}$$

so that

$$\frac{\partial}{\partial t}\left[\Gamma G(t,\mathbf{y};\mathbf{x})A^{-1}(\mathbf{x})\Gamma\right]A(\mathbf{x})+\frac{\partial}{\partial x^j}\left[\Gamma G(t,\mathbf{y};\mathbf{x})A^{-1}(\mathbf{x})\Gamma\right]D^j=0$$

with $\Gamma G(0,\mathbf{y};\mathbf{x})A^{-1}(\mathbf{x})\Gamma=A^{-1}(\mathbf{x})\delta(\mathbf{x}-\mathbf{y})$. Thus (3.8) follows from the uniqueness of the solution to the above hyperbolic system with given initial conditions. We can now recast (3.5) as

$$\begin{aligned}\mathbf{u}^B(\boldsymbol{\xi};\mathbf{x}_0)\;=\;&\int_{\mathbb{R}^9}\Gamma G(T,\mathbf{x}_0+\varepsilon\boldsymbol{\xi};\mathbf{y})G_*(T,\mathbf{x}_0+\varepsilon\mathbf{z};\mathbf{y}')\Gamma\\&\times\chi(\mathbf{y})\chi(\mathbf{y}')f\left(\frac{\mathbf{y}-\mathbf{y}'}{\varepsilon}\right)A(\mathbf{x}_0+\varepsilon\mathbf{z})\mathbf{S}(\mathbf{z})d\mathbf{y}d\mathbf{y}'d\mathbf{z}.\end{aligned} \tag{3.12}$$

One further simplifies (3.12) with the help of the auxiliary matrix-valued functions $Q(t,\mathbf{x};\mathbf{q})$ and $Q_*(t,\mathbf{x},\mathbf{q})$ defined by

$$\begin{aligned}Q(T,\mathbf{x};\mathbf{q})&=\int_{\mathbb{R}^3}G(T,\mathbf{x};\mathbf{y})\chi(\mathbf{y})e^{i\mathbf{q}\cdot\mathbf{y}/\varepsilon}d\mathbf{y},\\Q_*(T,\mathbf{x};\mathbf{q})&=\int_{\mathbb{R}^3}G_*(T,\mathbf{x};\mathbf{y})\chi(\mathbf{y})e^{-i\mathbf{q}\cdot\mathbf{y}/\varepsilon}d\mathbf{y}.\end{aligned} \tag{3.13}$$

They solve the hyperbolic equations (2.3) and (3.7) with initial conditions given by $Q(0,\mathbf{x};\mathbf{q})=\chi(\mathbf{x})e^{i\mathbf{q}\cdot\mathbf{x}/\varepsilon}I$ and $Q_*(0,\mathbf{x};\mathbf{q})=A^{-1}(\mathbf{x})\chi(\mathbf{x})e^{-i\mathbf{q}\cdot\mathbf{x}/\varepsilon}$, respectively. Thus (3.12) becomes

$$\mathbf{u}^B(\boldsymbol{\xi};\mathbf{x}_0)=\int_{\mathbb{R}^6}\Gamma Q(T,\mathbf{x}_0+\varepsilon\boldsymbol{\xi};\mathbf{q})Q_*(T,\mathbf{x}_0+\varepsilon\mathbf{z};\mathbf{q})\Gamma A(\mathbf{x}_0+\varepsilon\mathbf{z})\mathbf{S}(\mathbf{z})\hat{f}(\mathbf{q})\frac{d\mathbf{q}d\mathbf{z}}{(2\pi)^3}, \tag{3.14}$$

where $\hat{f}(\mathbf{q})=\int_{\mathbb{R}^3}e^{-i\mathbf{q}\cdot\mathbf{x}}f(\mathbf{x})d\mathbf{x}$ is the Fourier transform of $f(\mathbf{x})$.

3.4. Wigner transform. The back-propagated signal in (3.14) now has the form suitable to be analyzed in the Wigner transform formalism [14, 25]. We define

$$W_\varepsilon(t,\mathbf{x},\mathbf{k})=\int_{\mathbb{R}^3}\hat{f}(\mathbf{q})U_\varepsilon(t,\mathbf{x},\mathbf{k};\mathbf{q})d\mathbf{q}, \tag{3.15}$$

where

$$U_\varepsilon(t,\mathbf{x},\mathbf{k};\mathbf{q})=\int_{\mathbb{R}^3}e^{i\mathbf{k}\cdot\mathbf{y}}Q\left(t,\mathbf{x}-\frac{\varepsilon\mathbf{y}}{2};\mathbf{q}\right)Q_*\left(t,\mathbf{x}+\frac{\varepsilon\mathbf{y}}{2};\mathbf{q}\right)\frac{d\mathbf{y}}{(2\pi)^3}. \tag{3.16}$$

Taking the inverse Fourier transform, we verify that

$$Q(t,\mathbf{x};\mathbf{q})Q_*(t,\mathbf{y};\mathbf{q})=\int_{\mathbb{R}^3}e^{-i\mathbf{k}\cdot(\mathbf{y}-\mathbf{x})/\varepsilon}U_\varepsilon\left(t,\frac{\mathbf{x}+\mathbf{y}}{2},\mathbf{k};\mathbf{q}\right)d\mathbf{k};$$

hence

$$\mathbf{u}^B(\boldsymbol{\xi};\mathbf{x}_0)=\int_{\mathbb{R}^6} e^{i\mathbf{k}\cdot(\boldsymbol{\xi}-\mathbf{z})}\Gamma W_\varepsilon\left(T,\mathbf{x}_0+\varepsilon\frac{\mathbf{z}+\boldsymbol{\xi}}{2},\mathbf{k}\right)\Gamma A(\mathbf{x}_0+\varepsilon\mathbf{z})\mathbf{S}(\mathbf{z})\frac{d\mathbf{z}d\mathbf{k}}{(2\pi)^3}. \tag{3.17}$$

We have thus reduced the analysis of $\mathbf{u}(\boldsymbol{\xi};\mathbf{x}_0)$ as $\varepsilon\to 0$ to that of the asymptotic properties of the Wigner transform W_ε. The Wigner transform has been used extensively in the study of wave propagation in random media, especially in the derivation of radiative transport equations modeling the propagation of high frequency waves. We refer to [14, 21, 25]. Note that in the usual definition of the Wigner transform, one has the adjoint matrix Q^* in place of Q_* in (3.16). This difference is not essential since Q_* and Q^* satisfy the same evolution equation, though with different initial data.

The main reason for using the Wigner transform in (3.17) is that W_ε has a weak limit W as $\varepsilon\to 0$. Its existence follows from simple a priori bounds for $W_\varepsilon(t,\mathbf{x},\mathbf{k})$. Let us introduce the space $\mathcal{A}$ of matrix-valued functions $\phi(\mathbf{x},\mathbf{k})$ bounded in the norm $\|\cdot\|_{\mathcal{A}}$ defined by

$$\|\phi\|_{\mathcal{A}}=\int_{\mathbb{R}^3}\sup_{\mathbf{x}}\|\tilde{\phi}(\mathbf{x},\mathbf{y})\|d\mathbf{y},\qquad\text{where}\quad\tilde{\phi}(\mathbf{x},\mathbf{y})=\int_{\mathbb{R}^3}e^{-i\mathbf{k}\cdot\mathbf{y}}\phi(\mathbf{x},\mathbf{k})d\mathbf{k}.$$

We denote by $\mathcal{A}'$ its dual space, which is a space of distributions large enough to contain matrix-valued bounded measures, for instance. We then have the following result.

LEMMA 3.1. *Let $\chi(\mathbf{x})\in L^2(\mathbb{R}^3)$ and $\hat{f}(\mathbf{q})\in L^1(\mathbb{R}^3)$. Then there is a constant $C>0$ independent of $\varepsilon>0$ and $t\in[0,\infty)$ such that for all $t\in[0,\infty)$ we have $\|W_\varepsilon(t,\mathbf{x},\mathbf{k})\|_{\mathcal{A}'}<C$.*

The proof of this lemma is essentially contained in [14, 21]; see also [1]. One may actually get L^2-bounds for W_ε in our setting because of the regularizing effect of $\hat{f}$ in (3.15), but this is not essential for the purposes of this paper. We therefore obtain the existence of a subsequence $\varepsilon_k\to 0$ such that W_{ε_k} converges weakly to a distribution $W\in\mathcal{A}'$. Moreover, an easy calculation shows that at time $t=0$ we have

$$W(0,\mathbf{x}_0,\mathbf{k})=|\chi(\mathbf{x}_0)|^2A_0^{-1}(\mathbf{x}_0)\hat{f}(\mathbf{k}). \tag{3.18}$$

Here, $A_0=A$ when A is independent of ε, and $A_0=\lim_{\varepsilon\to 0}A_\varepsilon$ if we assume that the family of diagonal matrices $A_\varepsilon(\mathbf{x})$ is uniformly positive definite, bounded, and continuous with limit A_0 in $\mathcal{C}(\mathbb{R}^d)$. These assumptions on A_ε are sufficient to deal with the radiative transport regime we will consider in section 3.7. Under the same assumptions on A_ε, we have the following result.

PROPOSITION 3.2. *The back-propagated signal $\mathbf{u}^B(\boldsymbol{\xi};\mathbf{x}_0)$ given by (3.17) converges weakly in $\mathcal{S}'(\mathbb{R}^3\times\mathbb{R}^3)$ as $\varepsilon\to 0$ to the limit*

$$\mathbf{u}^B(\boldsymbol{\xi};\mathbf{x}_0)=\int_{\mathbb{R}^6}e^{i\mathbf{k}\cdot(\boldsymbol{\xi}-\mathbf{z})}\Gamma W(T,\mathbf{x}_0,\mathbf{k})\Gamma A_0(\mathbf{x}_0)\mathbf{S}(\mathbf{z})\frac{d\mathbf{z}d\mathbf{k}}{(2\pi)^3}. \tag{3.19}$$

The proof of this proposition is based on taking the duality product of $\mathbf{u}^B(\boldsymbol{\xi};\mathbf{x}_0)$ with a vector-valued test function $\boldsymbol{\phi}(\boldsymbol{\xi};\mathbf{x}_0)$ in $\mathcal{S}(\mathbb{R}^3\times\mathbb{R}^3)$. After a change of variables we obtain $\langle\mathbf{u}^B,\boldsymbol{\phi}\rangle=\langle W_\varepsilon,Z_\varepsilon\rangle$. Here the duality product for matrices is given by the trace $\langle A,B\rangle=\sum_{i,k}\langle A_{ik},B_{ik}\rangle$, and

$$Z_\varepsilon(\mathbf{x}_0,\mathbf{k})=\int_{\mathbb{R}^6}e^{i\mathbf{k}\cdot(\mathbf{z}-\boldsymbol{\xi})}\Gamma\boldsymbol{\phi}\left(\boldsymbol{\xi},\mathbf{x}_0-\varepsilon\frac{\mathbf{z}+\boldsymbol{\xi}}{2}\right)\mathbf{S}^*(\mathbf{z})A_\varepsilon\left(\mathbf{x}_0+\varepsilon\frac{\mathbf{z}-\boldsymbol{\xi}}{2}\right)\Gamma\frac{d\mathbf{z}d\boldsymbol{\xi}}{(2\pi)^3}. \tag{3.20}$$

Defining Z as the limit of Z_ε as $\varepsilon \to 0$ by formally replacing ε by 0 in the above expression, (3.19) follows from showing that $\|Z_\varepsilon - Z\|_{\mathcal{A}} \to 0$ as $\varepsilon \to 0$. This is straightforward, and we omit the details.

The above proposition tells us how to reconstruct the back-propagated solution in the high frequency limit from the limit Wigner matrix W. Notice that we have made almost no assumptions on the medium described by the matrix $A_\varepsilon(\mathbf{x})$. At this level, the medium can be either homogeneous or heterogeneous. Without any further assumptions, we can also obtain some information about the matrix W. Let us define the dispersion matrix for the system (2.3) as (see [25])

$$L(\mathbf{x},\mathbf{k}) = A_0^{-1}(\mathbf{x})k_j D^j. \tag{3.21}$$

This is given explicitly by

$$L(\mathbf{x},\mathbf{k}) = \begin{pmatrix} 0 & 0 & 0 & k_1/\rho(\mathbf{x}) \\ 0 & 0 & 0 & k_2/\rho(\mathbf{x}) \\ 0 & 0 & 0 & k_3/\rho(\mathbf{x}) \\ k_1/\kappa(\mathbf{x}) & k_2/\kappa(\mathbf{x}) & k_3/\kappa(\mathbf{x}) & 0 \end{pmatrix}.$$

The matrix L has a double eigenvalue $\omega_0 = 0$ and two simple eigenvalues $\omega_\pm(\mathbf{x},\mathbf{k}) = \pm c(\mathbf{x})|\mathbf{k}|$, where $c(\mathbf{x}) = 1/\sqrt{\rho(\mathbf{x})\kappa(\mathbf{x})}$ is the speed of sound. The eigenvalues $\omega_\pm$ are associated with eigenvectors $\mathbf{b}_\pm(\mathbf{x},\mathbf{k})$, and the eigenvalue $\omega_0 = 0$ is associated with the eigenvectors $\mathbf{b}_j(\mathbf{x},\mathbf{k})$, $j = 1, 2$. The eigenvectors are given by

$$\mathbf{b}_\pm(\mathbf{x},\mathbf{k}) = \begin{pmatrix} \pm\dfrac{\hat{\mathbf{k}}}{\sqrt{2\rho(\mathbf{x})}} \\ \dfrac{1}{\sqrt{2\kappa(\mathbf{x})}} \end{pmatrix}, \qquad \mathbf{b}_j(\mathbf{x},\mathbf{k}) = \begin{pmatrix} \dfrac{\mathbf{z}^j(\mathbf{k})}{\sqrt{\rho(\mathbf{x})}} \\ 0 \end{pmatrix}, \tag{3.22}$$

where $\hat{\mathbf{k}} = \mathbf{k}/|\mathbf{k}|$ and $\mathbf{z}^1(\mathbf{k})$ and $\mathbf{z}^2(\mathbf{k})$ are chosen so that the triple $(\hat{\mathbf{k}}, \mathbf{z}^1(\mathbf{k}), \mathbf{z}^2(\mathbf{k}))$ forms an orthonormal basis. The eigenvectors are normalized so that

$$(A_0(\mathbf{x})\mathbf{b}_j(\mathbf{x},\mathbf{k}) \cdot \mathbf{b}_k(\mathbf{x},\mathbf{k})) = \delta_{jk} \tag{3.23}$$

for all $j, k \in J = \{+, -, 1, 2\}$. The space of 4×4 matrices is clearly spanned by the basis $\mathbf{b}_j \otimes \mathbf{b}_k$. We then have the following result.

PROPOSITION 3.3. *There exist scalar distributions $a_\pm$ and a_0^{mn}, $m, n = 1, 2$, so that the limit Wigner distribution matrix can be decomposed as*

$$\begin{aligned} W(t,\mathbf{x},\mathbf{k}) = &\sum_{j,m=1}^{2} a_0^{jm}(t,\mathbf{x},\mathbf{k})\mathbf{b}_j(\mathbf{x},\mathbf{k}) \otimes \mathbf{b}_m(\mathbf{x},\mathbf{k}) \\ &+ a_+(t,\mathbf{x},\mathbf{k})\mathbf{b}_+(\mathbf{x},\mathbf{k}) \otimes \mathbf{b}_+(\mathbf{x},\mathbf{k}) + a_-(t,\mathbf{x},\mathbf{k})\mathbf{b}_-(\mathbf{x},\mathbf{k}) \otimes \mathbf{b}_-(\mathbf{x},\mathbf{k}). \end{aligned} \tag{3.24}$$

The main result of this proposition is that the cross terms $\mathbf{b}_j \otimes \mathbf{b}_k$ with $\omega_j \neq \omega_k$ do not contribute to the limit W. The proof of this proposition can be found in [14] and a formal derivation in [25].

The initial conditions for the amplitudes a_j are calculated using the identity

$$A_0^{-1}(\mathbf{x}) = \sum_{j \in J} \mathbf{b}_j(\mathbf{x},\mathbf{k}) \otimes \mathbf{b}_j(\mathbf{x},\mathbf{k}).$$

Then (3.18) implies that $a_0^{12}(0,\mathbf{x},\mathbf{k}) = a_0^{21}(0,\mathbf{x},\mathbf{k}) = 0$ and

$$a_0^{jj}(0,\mathbf{x},\mathbf{k}) = a_\pm(0,\mathbf{x},\mathbf{k}) = |\chi(\mathbf{x})|^2 f(\mathbf{k}), \quad j = 1, 2. \tag{3.25}$$

3.5. Mode decomposition and refocusing. We can use the above result to recast (3.19) as

$$\mathbf{u}^B(\boldsymbol{\xi};\mathbf{x}_0) = (F(T,\cdot;\mathbf{x}_0) * \mathbf{S})(\boldsymbol{\xi}), \tag{3.26}$$

where

$$\begin{aligned} F(T,\boldsymbol{\xi};\mathbf{x}_0) = &\sum_{m,n=1}^{2} \int_{\mathbb{R}^3} e^{i\mathbf{k}\cdot\boldsymbol{\xi}} a_0^{mn}(T,\mathbf{x}_0;\mathbf{k}) \Gamma \mathbf{b}_m(\mathbf{x}_0,\mathbf{k}) \otimes \mathbf{b}_n(\mathbf{x}_0,\mathbf{k}) A_0(\mathbf{x}_0) \Gamma \frac{d\mathbf{k}}{(2\pi)^3} \\ &+ \int_{\mathbb{R}^3} e^{i\mathbf{k}\cdot\boldsymbol{\xi}} a_+(T,\mathbf{x}_0;\mathbf{k}) \Gamma \mathbf{b}_+(\mathbf{x}_0,\mathbf{k}) \otimes \mathbf{b}_+(\mathbf{x}_0,\mathbf{k}) A_0(\mathbf{x}_0) \Gamma \frac{d\mathbf{k}}{(2\pi)^3} \\ &+ \int_{\mathbb{R}^3} e^{i\mathbf{k}\cdot\boldsymbol{\xi}} a_-(T,\mathbf{x}_0;\mathbf{k}) \Gamma \mathbf{b}_-(\mathbf{x}_0,\mathbf{k}) \otimes \mathbf{b}_-(\mathbf{x}_0,\mathbf{k}) A_0(\mathbf{x}_0) \Gamma \frac{d\mathbf{k}}{(2\pi)^3}. \end{aligned} \tag{3.27}$$

This expression can be used to assess the quality of the refocusing. When $F(T,\boldsymbol{\xi};\mathbf{x}_0)$ has a narrow support in $\boldsymbol{\xi}$, refocusing is good. When its support in $\boldsymbol{\xi}$ grows larger, its quality degrades. The spatial decay of the kernel $F(t,\boldsymbol{\xi};\mathbf{x}_0)$ in $\boldsymbol{\xi}$ is directly related to the smoothness in $\mathbf{k}$ of its Fourier transform in $\boldsymbol{\xi}$:

$$\begin{aligned} \hat{F}(T,\mathbf{k};\mathbf{x}_0) = &\sum_{m,n=1}^{2} a_0^{mn}(T,\mathbf{x}_0;\mathbf{k}) \Gamma \mathbf{b}_m(\mathbf{x}_0,\mathbf{k}) \otimes \mathbf{b}_n(\mathbf{x}_0,\mathbf{k}) A_0(\mathbf{x}_0) \Gamma \frac{d\mathbf{k}}{(2\pi)^3} \\ &+ \Gamma \left[a_+(T,\mathbf{x}_0;\mathbf{k}) \mathbf{b}_+(\mathbf{x}_0,\mathbf{k}) \otimes \mathbf{b}_+(\mathbf{x}_0,\mathbf{k}) + a_-(T,\mathbf{x}_0;\mathbf{k}) \mathbf{b}_-(\mathbf{x}_0,\mathbf{k}) \otimes \mathbf{b}_-(\mathbf{x}_0,\mathbf{k}) \right] A_0(\mathbf{x}_0) \Gamma. \end{aligned}$$

Namely, for F to decay in $\boldsymbol{\xi}$, one needs $\hat{F}(\mathbf{k})$ to be smooth in $\mathbf{k}$. However, the eigenvectors $\mathbf{b}_j$ are singular at $\mathbf{k}=0$ as can be seen from the explicit expressions (3.22). Therefore, a priori $\hat{F}$ is not smooth at $\mathbf{k}=0$. This means that, in order to obtain good refocusing, one needs the original signal to have no low frequencies: $\hat{S}(\mathbf{k})=0$ near $\mathbf{k}=0$. Low frequencies in the initial data will not refocus well.

We can further simplify (3.26)–(3.27) if we assume that the initial condition is irrotational. Taking the Fourier transform of both sides in (3.26), we obtain that

$$\hat{\mathbf{u}}^B(\mathbf{k};\mathbf{x}_0) = \sum_{j,n\in J} a_j(T,\mathbf{x}_0,\mathbf{k}) \hat{S}_n(\mathbf{k}) (A_0(\mathbf{x}_0)\Gamma \mathbf{b}_n(\mathbf{x}_0,\mathbf{k}) \cdot \mathbf{b}_j(\mathbf{x}_0,\mathbf{k})) \Gamma \mathbf{b}_j(\mathbf{x}_0,\mathbf{k}), \tag{3.28}$$

where we have defined

$$\hat{\mathbf{S}}(\mathbf{k}) = \sum_{n\in J} \hat{S}_n(\mathbf{k}) \mathbf{b}_n(\mathbf{x}_0,\mathbf{k}). \tag{3.29}$$

Irrotationality of the initial condition means that $\hat{S}_1$ and $\hat{S}_2$ identically vanish, or equivalently that

$$\mathbf{S}(\mathbf{x}) = \begin{pmatrix} \nabla\phi(\mathbf{x}) \\ p(\mathbf{x}) \end{pmatrix} \tag{3.30}$$

for some pressure $p(\mathbf{x})$ and potential $\phi(\mathbf{x})$. Remarking that $\Gamma\mathbf{b}_\pm = -\mathbf{b}_\mp$ and by irrotationality that $(A_0(\mathbf{x}_0)\hat{\mathbf{S}}(\mathbf{k}) \cdot \mathbf{b}_{1,2}(\mathbf{k})) = 0$, we use (3.23) to recast (3.28) as

$$\hat{\mathbf{u}}^B(\mathbf{k};\mathbf{x}_0) = a_-(T,\mathbf{x}_0,\mathbf{k}) \hat{S}_+(\mathbf{k}) \mathbf{b}_+(\mathbf{x}_0,\mathbf{k}) + a_+(T,\mathbf{x}_0,\mathbf{k}) \hat{S}_-(\mathbf{k}) \mathbf{b}_-(\mathbf{x}_0,\mathbf{k}). \tag{3.31}$$

Decomposing the initial condition $\mathbf{S}(\mathbf{x})$ as

$$\mathbf{S}(\mathbf{x}) = \mathbf{S}_+(\mathbf{x}) + \mathbf{S}_-(\mathbf{x}) \qquad \text{such that} \quad \hat{\mathbf{S}}_\pm(\mathbf{k}) = \hat{S}_\pm(\mathbf{k})\mathbf{b}_\pm(\mathbf{x}_0, \mathbf{k}),$$

the back-propagated signal takes the form

$$\mathbf{u}^B(\boldsymbol{\xi}; \mathbf{x}_0) = (\hat{a}_-(T, \mathbf{x}_0, \cdot) * \mathbf{S}_+(\cdot))(\boldsymbol{\xi}) + (\hat{a}_+(T, \mathbf{x}_0, \cdot) * \mathbf{S}_-(\cdot))(\boldsymbol{\xi}), \tag{3.32}$$

where $\hat{a}_\pm$ is the Fourier transform of $a_\pm$ in $\mathbf{k}$. This form is much more tractable than (3.26)–(3.27). It is also almost as general. Indeed, rotational modes do not propagate in the high frequency regime. Therefore, they are exactly back-propagated when $\chi(\mathbf{x}_0) = 1$ and $f(\mathbf{x}) = \delta(\mathbf{x})$, and not back-propagated at all when $\chi(\mathbf{x}_0) = 0$. All the refocusing properties are thus captured by the amplitudes $a_\pm(T, \mathbf{x}_0, \mathbf{k})$. Their evolution equation characterizes how waves propagate in the medium and their initial conditions characterize the recording array.

3.6. Homogeneous media. In homogeneous media with $c(\mathbf{x}) = c_0$ the amplitudes $a_\pm(T, \mathbf{x}, \mathbf{k})$ satisfy the free transport equation [14, 25]

$$\frac{\partial a_\pm}{\partial t} \pm c_0 \hat{\mathbf{k}} \cdot \nabla_{\mathbf{x}} a_\pm = 0 \tag{3.33}$$

with initial data $a_\pm(0, \mathbf{x}, \mathbf{k}) = |\chi(\mathbf{x})|^2 f(\mathbf{k})$ as in (3.25). They are therefore given by

$$a_\pm(t, \mathbf{x}_0, \mathbf{k}) = |\chi(\mathbf{x}_0 \mp c_0 \hat{\mathbf{k}} t)|^2 \hat{f}(\mathbf{k}). \tag{3.34}$$

These amplitudes become more and more singular in $\mathbf{k}$ as time grows since their gradient in $\mathbf{k}$ grows linearly with time. The corresponding kernel F decays therefore more slowly in $\boldsymbol{\xi}$ as time grows. This implies that the quality of the refocusing degrades with time. For sufficiently large times, all the energy has left the domain Ω (assumed to be bounded), and the coefficients $a_\pm(t, \mathbf{x}_0, \mathbf{k})$ vanish. Therefore the back-propagated signal $\mathbf{u}^B(\boldsymbol{\xi}; \mathbf{x}_0)$ also vanishes, which means that there is no refocusing at all. The same conclusions could also be drawn by analyzing (3.4) directly in a homogeneous medium. This is the situation in the numerical experiment presented in Figure 2.2: in a homogeneous medium, the back-propagated signal would vanish.

3.7. Heterogeneous media and the radiative transport regime. The results of the preceding sections show how the back-propagated signal $\mathbf{u}^B(\boldsymbol{\xi}; \mathbf{x}_0)$ is related to the propagating modes $a_\pm(T, \mathbf{x}_0, \mathbf{k})$ of the Wigner matrix $W(T, \mathbf{x}_0, \mathbf{k})$. The form assumed by the modes $a_\pm(T, \mathbf{x}_0, \mathbf{k})$, and in particular their smoothness in $\mathbf{k}$, will depend on the hypotheses we make on the underlying medium, i.e., on the density $\rho(\mathbf{x})$ and compressibility $\kappa(\mathbf{x})$ that appear in the matrix $A(\mathbf{x})$. We have seen that partial measurements in homogeneous media yield poor refocusing properties. We now show that refocusing is much better in random media.

We consider here the radiative transport regime, also known as the weak coupling limit. There, the fluctuations in the physical parameters are weak and vary on a scale comparable to the scale of the initial condition. Density and compressibility assume the form

$$\rho(\mathbf{x}) = \rho_0 + \sqrt{\varepsilon}\rho_1\left(\frac{\mathbf{x}}{\varepsilon}\right) \qquad \text{and} \qquad \kappa(\mathbf{x}) = \kappa_0 + \sqrt{\varepsilon}\kappa_1\left(\frac{\mathbf{x}}{\varepsilon}\right). \tag{3.35}$$

The functions ρ_1 and κ_1 are assumed to be mean-zero spatially homogeneous processes. The average (with respect to realizations of the medium) of the propagating

amplitudes $a_\pm$, denoted by $\bar{a}_\pm$, satisfy in the high frequency limit $\varepsilon \to 0$ a radiative transfer equation (RTE), which is a linear Boltzmann equation of the form

$$\begin{aligned} &\frac{\partial \bar{a}_\pm}{\partial t} \pm c_0 \hat{\mathbf{k}} \cdot \nabla_{\mathbf{x}} \bar{a}_\pm = \int_{\mathbb{R}^3} \sigma(\mathbf{k},\mathbf{p})(\bar{a}_\pm(t,\mathbf{x},\mathbf{p}) - \bar{a}_\pm(t,\mathbf{x},\mathbf{k}))\delta(c_0(|\mathbf{k}|-|\mathbf{p}|))d\mathbf{p}, \\ &\bar{a}_\pm(0,\mathbf{x},\mathbf{k}) = |\chi(\mathbf{x})|^2 \hat{f}(\mathbf{k}). \end{aligned} \tag{3.36}$$

The scattering coefficient $\sigma(\mathbf{k},\mathbf{p})$ depends on the power spectra of ρ_1 and κ_1. We refer to [25] for the details of the derivation and explicit form of $\sigma(\mathbf{k},\mathbf{p})$. The above result remains formal for the wave equation and requires averaging over the realizations of the random medium, although this is not necessary in the physical and numerical time reversal experiments. A rigorous derivation of the linear Boltzmann equation (which also requires averaging over realizations) has been obtained only for the Schrödinger equation; see [9, 27]. Nevertheless, the above result formally characterizes the filter $F(T,\boldsymbol{\xi};\mathbf{x}_0)$ introduced in (3.27) and (3.32).

The transport equation (3.36) has a smoothing effect best seen in its integral formulation. Let us define the total scattering coefficient $\Sigma(\mathbf{k}) = \int_{\mathbb{R}^3} \sigma(\mathbf{k},\mathbf{p})\delta(c_0(|\mathbf{k}|-|\mathbf{p}|))d\mathbf{p}$. Then the transport equation (3.36) may be rewritten as

$$\begin{aligned} \bar{a}_\pm(t,\mathbf{x},\mathbf{k}) &= \bar{a}_\pm(0,\mathbf{x} \mp c_0\hat{\mathbf{k}}t,\mathbf{k})e^{-\Sigma(\mathbf{k})t} \\ &+ \frac{|\mathbf{k}|^2}{c_0}\int_0^t ds \int_{S^2} \sigma(\mathbf{k},|\mathbf{k}|\hat{\mathbf{p}})\bar{a}_\pm(s,\mathbf{x}\mp c_0(t-s)\hat{\mathbf{k}},|\mathbf{k}|\hat{\mathbf{p}})e^{-\Sigma(\mathbf{k})(t-s)}d\Omega(\hat{\mathbf{p}}). \end{aligned} \tag{3.37}$$

Here $\hat{\mathbf{p}} = \mathbf{p}/|\mathbf{p}|$ is the unit vector in the direction of $\mathbf{p}$, and $d\Omega(\hat{\mathbf{p}})$ is the surface element on the sphere S^2. The first term in (3.37) is the ballistic part that undergoes no scattering. It has no smoothing effect, and, moreover, if $a(0,\mathbf{x},\mathbf{k})$ is not smooth in $\mathbf{x}$, as may be the case for (3.25), the discontinuities in $\mathbf{x}$ translate into discontinuities in $\mathbf{k}$ at later times, as in (3.34) in a homogeneous medium. However, in contrast to the homogeneous medium case, the ballistic term decays exponentially in time and does not affect the refocused signal for sufficiently long times $t \gg 1/\Sigma$. The second term in (3.37) exhibits a smoothing effect. Namely, the operator $\mathcal{L}g$ defined by

$$\mathcal{L}g(t,\mathbf{x},\mathbf{k}) = \frac{|\mathbf{k}|^2}{c_0}\int_0^t ds \int_{S^2} \sigma(\mathbf{k},|\mathbf{k}|\hat{\mathbf{p}})g(s,\mathbf{x}\mp c_0(t-s)\hat{\mathbf{k}},|\mathbf{k}|\hat{\mathbf{p}})e^{-\Sigma(\mathbf{k})(t-s)}d\Omega(\hat{\mathbf{p}})$$

is regularizing, in the sense that the function $\tilde{g} = \mathcal{L}g$ has at least $1/2$ more derivatives than g (in the same space of a Sobolev scale). The precise formulation of this smoothing property is given by the averaging lemmas [15, 22] and will not be dwelt upon here. Iterating (3.37) n times, we obtain

$$\bar{a}_\pm(t,\mathbf{x},\mathbf{k}) = a^0_\pm(t,\mathbf{x},\mathbf{k}) + a^1_\pm(t,\mathbf{x},\mathbf{k}) + \cdots + a^n_\pm(t,\mathbf{x},\mathbf{k}) + \mathcal{L}^{n+1}\bar{a}_\pm(t,\mathbf{x},\mathbf{k}). \tag{3.38}$$

The terms $a^0_\pm, \ldots, a^n_\pm$ are given by

$$a^0_\pm(t,\mathbf{x},\mathbf{k}) = \bar{a}_\pm(0,\mathbf{x}\mp c_0\hat{\mathbf{k}}t,\mathbf{k})e^{-\Sigma(\mathbf{k})t}, \qquad a^j_\pm(t,\mathbf{x},\mathbf{k}) = \mathcal{L}a^{j-1}_\pm(t,\mathbf{x},\mathbf{k}).$$

They describe, respectively, the contributions from waves that do not scatter, scatter once, twice, ... It is straightforward to verify that all these terms decay exponentially in time and are negligible for times $t \gg 1/\Sigma$. The last term in (3.38) has at least $n/2$ more derivatives than the initial data a_0 or the solution (3.34) of the homogeneous transport equation. This leads to a faster decay in $\boldsymbol{\xi}$ of the Fourier transforms $\hat{a}_\pm(T,\mathbf{x}_0,\boldsymbol{\xi})$ of $\bar{a}_\pm(T,\mathbf{x}_0,\mathbf{k})$ in $\mathbf{k}$. This gives a qualitative explanation as to why refocusing is better in heterogeneous media than in homogeneous media. A more quantitative answer requires solving the transport equation (3.36).

3.8. Diffusion regime. It is known for times t much longer than the scattering mean free time $\tau_{sc} = 1/\Sigma$ and distances of propagation L very large compared to $l_{sc} = c_0\tau_{sc}$ that solutions to the radiative transport equation (3.36) can be approximated by solutions to a diffusion equation, provided that $c(\mathbf{x}) = c_0$ is independent of $\mathbf{x}$ [6, 20]. More precisely, we let $\delta = l_{sc}/L \ll 1$ be a small parameter and rescale time and space variables as $t \to t/\delta^2$ and $\mathbf{x} \to \mathbf{x}/\delta$. In this limit, the wave direction is completely randomized so that

$$\bar{a}_+(t,\mathbf{x},\mathbf{k}) \approx \bar{a}_-(t,\mathbf{x},\mathbf{k}) \approx a(t,\mathbf{x},|\mathbf{k}|),$$

where a solves

$$\begin{aligned} &\frac{\partial a(t,\mathbf{x},|\mathbf{k}|)}{\partial t} - D(|\mathbf{k}|)\Delta_{\mathbf{x}} a(t,\mathbf{x},|\mathbf{k}|) = 0, \\ &a(0,\mathbf{x},|\mathbf{k}|) = |\chi(\mathbf{x})|^2 \frac{1}{4\pi|\mathbf{k}|^2} \int_{\mathbb{R}^3} \hat{f}(\mathbf{q})\delta(|\mathbf{q}| - |\mathbf{k}|)d\mathbf{q}. \end{aligned} \tag{3.39}$$

The diffusion coefficient $D(|\mathbf{k}|)$ may be expressed explicitly in terms of the scattering coefficient $\sigma(\mathbf{k},\mathbf{p})$ and hence related to the power spectra of ρ_1 and κ_1. We refer to [25] for the details. For instance, let us assume for simplicity that the density is not fluctuating, $\rho_1 \equiv 0$, and that the compressibility fluctuations are delta-correlated, so that $\mathbb{E}\{\hat{\kappa}_1(\mathbf{p})\hat{\kappa}_1(\mathbf{q})\} = \kappa_0^2\hat{R}_0\delta(\mathbf{p}+\mathbf{q})$. Then we have

$$\sigma(\mathbf{k},\mathbf{p}) = \frac{\pi c_0^2|\mathbf{k}|^2\hat{R}_0}{2}, \qquad \Sigma(|\mathbf{k}|) = 2\pi^2 c_0|\mathbf{k}|^4\hat{R}_0 \tag{3.40}$$

and

$$D(|\mathbf{k}|) = \frac{c_0^2}{3\Sigma(|\mathbf{k}|)} = \frac{c_0}{6\pi^2|\mathbf{k}|^4\hat{R}_0}. \tag{3.41}$$

Let us assume that there are no initial rotational modes, so that the source $\mathbf{S}(\mathbf{x})$ is decomposed as in (3.30). Using (3.31), we obtain that

$$\hat{\mathbf{u}}^B(\mathbf{k};\mathbf{x}_0) = a(T,\mathbf{x}_0,|\mathbf{k}|)\hat{\mathbf{S}}(\mathbf{k}). \tag{3.42}$$

When $f(\mathbf{x})$ is isotropic so that $\hat{f}(\mathbf{k}) = \hat{f}(|\mathbf{k}|)$, and the diffusion coefficient is given by (3.41), the solution of (3.39) takes the form

(3.43)

$$a(T,\mathbf{x}_0,|\mathbf{k}|) = \hat{f}(|\mathbf{k}|)\left(\frac{3\pi|\mathbf{k}|^4\hat{R}_0}{2c_0T}\right)^{3/2}\int_{\mathbb{R}^3}\exp\left(-\frac{3\pi^2|\mathbf{k}|^4\hat{R}_0|\mathbf{x}_0-\mathbf{y}|^2}{2c_0T}\right)|\chi(\mathbf{y})|^2 d\mathbf{y}.$$

When $f(\mathbf{x}) = \delta(\mathbf{x})$, and $\Omega = \mathbb{R}^3$, so that $\chi(\mathbf{x}) \equiv 1$, we retrieve $a(T,\mathbf{x}_0,\mathbf{k}) \equiv 1$; hence the refocusing is perfect. When only partial measurement is available, the above formula indicates how the frequencies of the initial pulse are filtered by the single-time time reversal process. Notice that both the low and high frequencies are damped; the reason is that low frequencies scatter little from the underlying medium so that it takes a long time for them to be randomized. High frequencies strongly scatter with the underlying medium and consequently propagate little, so that the signal that reaches the recording array Ω is small unless recorders are also located at the source point: $\mathbf{x}_0 \in \Omega$. In the latter case they are very well measured and back-propagated, although this situation is not the most interesting physically. Expression (3.43) may be generalized to other power spectra of medium fluctuations in a straightforward manner using the formula for the diffusion coefficient in [25].

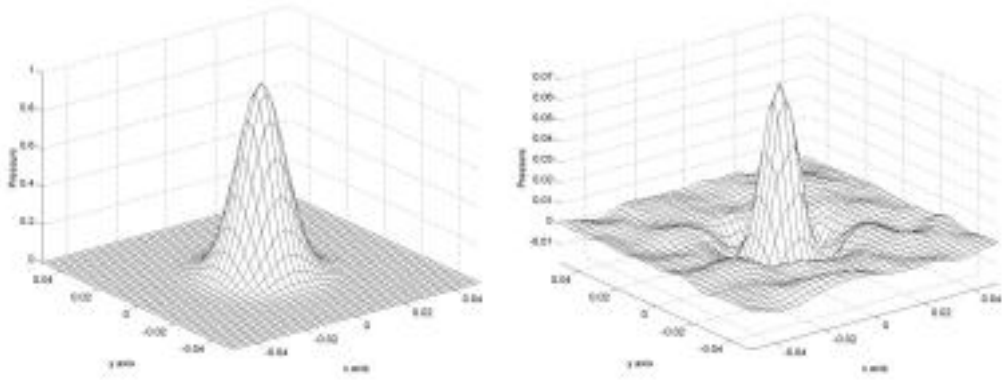


FIG. 3.1. *Zoom of the initial source and the refocused signal for the numerical experiment of Figure* 2.2.

3.9. Numerical results. The numerical results in Figure 2.2 show that some signal refocuses at the location of the initial source after the time reversal procedure. Based on the above theory, however, we do not expect the refocused signal to have exactly the same shape as the original one. Since the location of the initial source belongs to the recording array ($\chi(\mathbf{x}_0) = 1$) in our simulations, we expect from our theory that high frequencies will refocus well but that low frequencies will not. This is confirmed by the numerical results in Figure 3.1, where a zoom in the vicinity of $\mathbf{x}_0 = \mathbf{0}$ of the initial source and refocused signal are represented. Notice that the numerical simulations are presented here only to help in the understanding of the refocusing theory and do not aim at reproducing the theory in a quantitative manner. The random fluctuations are quite strong in our numerical simulations, and it is unlikely that the diffusive regime will be valid. The refocused signal in the right figure looks, however, like a high-pass filter of the signal in the left figure, as expected from theory.

4. Refocusing of classical waves. The theory presented in section 3 provides a quantitative explanation for the results observed in time reversal physical and numerical experiments. However, the time reversal procedure is by no means necessary to obtain refocusing. Time reversal is associated with the specific choice (3.2) for the matrix Γ in the preceding section, which reverses the direction of the acoustic velocity and keeps the pressure unchanged. Other choices for Γ are, however, possible. When nothing is done at time T, i.e., when we choose $\Gamma = I$, no refocusing occurs as one might expect. It turns out that $\Gamma = I$ is more or less the only choice of a matrix that prevents some sort of refocusing. Section 4.1 presents the theory of refocusing for acoustic waves, which is corroborated by numerical results presented in section 4.2. Sections 4.3 and 4.4 generalize the theory to other linear hyperbolic systems.

4.1. General refocusing of acoustic waves. In single-time time reversal, the action of the "intelligent" array is captured by the choice of the signal processing matrix Γ in (3.3). Time reversal is characterized by Γ given in (3.2). A passive array is characterized by $\Gamma = I$. This section analyzes the role of other choices for Γ, which we let depend on the receiver location so that each receiver may perform its own kind of signal processing.

The signal after time reversal is still given by (3.3), where $\Gamma(\mathbf{y}')$ is now arbitrary. At time $2T$, after back-propagation, we are free to multiply the signal by an arbitrary invertible matrix to analyze the signal. It is convenient to multiply the back-propagated signal by the matrix $\Gamma_0 = \mathrm{Diag}(-1,-1,-1,1)$ as in classical time reversal. The reconstruction formula (3.5) is then replaced by

$$\mathbf{u}^B(\boldsymbol{\xi};\mathbf{x}_0) = \int_{\mathbb{R}^9} \Gamma_0 G(T,\mathbf{x}_0+\varepsilon\boldsymbol{\xi};\mathbf{y})\Gamma(\mathbf{y}')G(T,\mathbf{y}';\mathbf{x}_0+\varepsilon\mathbf{z})\chi(\mathbf{y},\mathbf{y}')\mathbf{S}(\mathbf{z})d\mathbf{y}d\mathbf{y}'d\mathbf{z}, \tag{4.1}$$

with $\chi(\mathbf{y},\mathbf{y}')$ defined by (3.6). To generalize the results of section 3, we need to define an appropriate adjoint Green's matrix G_*. As before, this will allow us to remove the matrix Γ between the two Green's matrices in (4.1) and to interchange the order of points in the second Green's matrix. We define the new adjoint Green's function $G_*(t,\mathbf{x};\mathbf{y})$ as the solution to

$$\begin{aligned} &\frac{\partial G_*(t,\mathbf{x};\mathbf{y})}{\partial t}A(\mathbf{x}) + \frac{\partial G_*(t,\mathbf{x};\mathbf{y})}{\partial x^j}D^j = 0,\\ &G_*(0,\mathbf{x};\mathbf{y}) = \Gamma(\mathbf{x})\Gamma_0 A^{-1}(\mathbf{x})\delta(\mathbf{x}-\mathbf{y}). \end{aligned} \tag{4.2}$$

Following the steps of section 3.3, we show that

$$G_*(t,\mathbf{x},\mathbf{y}) = \Gamma(\mathbf{y})G(t,\mathbf{y};\mathbf{x})A^{-1}(\mathbf{x})\Gamma_0. \tag{4.3}$$

The only modification compared to the corresponding derivation of (3.8) is to multiply (3.9) on the left by $\Gamma(\mathbf{x})$ and on the right by Γ_0 so that $\Gamma(\mathbf{y})$ appears on the left in (3.10). The retransmitted signal may now be recast as

$$\mathbf{u}^B(\boldsymbol{\xi};\mathbf{x}_0) = \int_{\mathbb{R}^9} \Gamma_0 G(T,\mathbf{x}_0+\varepsilon\boldsymbol{\xi};\mathbf{y})G_*(T,\mathbf{x}_0+\varepsilon\mathbf{z};\mathbf{y}')\Gamma_0^{-1}A(\mathbf{x}_0+\varepsilon\mathbf{z})\chi(\mathbf{y},\mathbf{y}')\mathbf{S}(\mathbf{z})d\mathbf{y}d\mathbf{y}'d\mathbf{z}. \tag{4.4}$$

Therefore the only modification in the expression for the retransmitted signal compared to the time reversed signal (3.12) is in the initial data for (4.2), which is the only place where the matrix $\Gamma(\mathbf{x})$ appears.

The analysis in sections 3.3–3.7 requires only minor changes, which we now outline. The back-propagated signal may still be expressed in terms of the Wigner distribution (compare to (3.17))

$$\mathbf{u}^B(\boldsymbol{\xi};\mathbf{x}_0) = \int_{\mathbb{R}^6} e^{i\mathbf{k}\cdot(\boldsymbol{\xi}-\mathbf{z})}\Gamma_0 W_\varepsilon\left(T,\mathbf{x}_0+\varepsilon\frac{\mathbf{z}+\boldsymbol{\xi}}{2},\mathbf{k}\right)\Gamma_0 A(\mathbf{x}_0+\varepsilon\mathbf{z})\mathbf{S}(\mathbf{z})\frac{d\mathbf{z}d\mathbf{k}}{(2\pi)^3}. \tag{4.5}$$

The Wigner distribution is defined as before by (3.15) and (3.16). The function Q is defined as before as the solution of (2.3) with initial data $Q(0,x;\mathbf{q}) = \chi(\mathbf{x})e^{i\mathbf{q}\cdot\mathbf{x}/\varepsilon}I$, while Q_* solves (3.7) with the initial data $Q_*(0,\mathbf{x};\mathbf{q}) = \Gamma(\mathbf{x})\Gamma_0 A^{-1}(\mathbf{x})\chi(\mathbf{x})e^{-i\mathbf{q}\cdot\mathbf{x}/\varepsilon}$. The initial Wigner distribution is now given by

$$W(0,\mathbf{x},\mathbf{k}) = |\chi(\mathbf{x})|^2\Gamma(\mathbf{x})\Gamma_0 A^{-1}(\mathbf{x})\hat{f}(\mathbf{k}). \tag{4.6}$$

Lemma 3.1 and Proposition 3.2 also hold, and we obtain the analogue of (3.19):

$$\mathbf{u}(\boldsymbol{\xi};\mathbf{x}_0) = \int_{\mathbb{R}^6} e^{i\mathbf{k}\cdot(\boldsymbol{\xi}-\mathbf{z})}\Gamma_0 W(T,\mathbf{x}_0,\mathbf{k})\Gamma_0 A_0(\mathbf{x}_0)\mathbf{S}(\mathbf{z})d\mathbf{z}d\mathbf{k}. \tag{4.7}$$

The limit Wigner distribution $W(T, \mathbf{x}_0, \mathbf{k})$ admits the mode decomposition (3.24) as before. If we assume that the source $\mathbf{S}(\mathbf{x})$ has the form (3.30) so that no rotational modes are present initially, we recover the refocusing formula (3.31):

$$\hat{\mathbf{u}}^B(\mathbf{k}; \mathbf{x}_0) = a_-(T, \mathbf{x}_0, \mathbf{k})\hat{S}_+(\mathbf{k})\mathbf{b}_+(\mathbf{x}_0, \mathbf{k}) + a_+(T, \mathbf{x}_0, \mathbf{k})\hat{S}_-(\mathbf{k})\mathbf{b}_-(\mathbf{x}_0, \mathbf{k}). \tag{4.8}$$

The initial conditions for the amplitudes $a_\pm$ are replaced by

$$\begin{aligned} a_\pm(0, \mathbf{x}, \mathbf{k}) &= \mathrm{Tr}\left[A_0(\mathbf{x})W(0, \mathbf{x}, \mathbf{k})A_0(\mathbf{x})\mathbf{b}_\pm(\mathbf{x}_0, \mathbf{k})\mathbf{b}_\pm^*(\mathbf{x}_0, \mathbf{k})\right] \\ &= |\chi(\mathbf{x})|^2\hat{f}(\mathbf{k})(A_0(\mathbf{x})\Gamma(\mathbf{x})\mathbf{b}_\mp(\mathbf{x}, \mathbf{k}) \cdot \mathbf{b}_\pm(\mathbf{x}, \mathbf{k})). \end{aligned} \tag{4.9}$$

Observe that when $\Gamma(\mathbf{x}) = \Gamma_0$, we get back the results of section 3.7. When the signal is not changed at the array, so that $\Gamma = I$, the coefficients $a_\pm(0, \mathbf{x}, \mathbf{k}) \equiv 0$, by orthogonality (3.23) of the eigenvectors $\mathbf{b}_j$. Thus no refocusing occurs when the "intelligent" array is replaced by a passive array, as expected physically.

Another interesting example is when only pressure p is measured, so that the matrix $\Gamma = \mathrm{Diag}(0, 0, 0, 1)$. Then the initial data is

$$a_\pm(0, \mathbf{x}, \mathbf{k}) = \frac{1}{2}|\chi(\mathbf{x})|^2\hat{f}(\mathbf{k}),$$

which differs by a factor $1/2$ from the full time reversal case (3.25). Therefore the retransmitted signal $\mathbf{u}^B$ also differs by only a factor $1/2$ from the latter case, and the quality of refocusing as well as the shape of the repropagated signal are exactly the same. The same observation applies to the measurement and reversal of the acoustic velocity only, which corresponds to the matrix $\Gamma = \mathrm{Diag}(-1, -1, -1, 0)$. The factor $1/2$ comes from the fact that only the potential energy or the kinetic energy is measured in the first and second cases, respectively. For high frequency acoustic waves, the potential and kinetic energies are equal; hence the factor $1/2$. We can also verify that when only the first component of the velocity field is measured, so that $\Gamma = \mathrm{Diag}(-1, 0, 0, 0)$, the initial data is

$$a_\pm(0, \mathbf{x}, \mathbf{k}) = |\chi(\mathbf{x})|^2\hat{f}(\mathbf{k})\frac{k_1^2}{2|\mathbf{k}|^2}. \tag{4.10}$$

As in the time reversal setting of section 3, the quality of the refocusing is related to the smoothness of the amplitudes $a_\pm$ in $\mathbf{k}$. In a homogeneous medium they satisfy the free transport equation (3.33) and are given by

$$\begin{aligned} &a_\pm(t, \mathbf{x}, \mathbf{k}) \\ &\quad = |\chi(\mathbf{x} - c_0\hat{\mathbf{k}}t)|^2\hat{f}(\mathbf{k})(A_0(\mathbf{x} - c_0\hat{\mathbf{k}}t)\Gamma(\mathbf{x} - c_0\hat{\mathbf{k}}t)\mathbf{b}_\mp(\mathbf{x} - c_0\hat{\mathbf{k}}t, \mathbf{k}) \cdot \mathbf{b}_\pm(\mathbf{x} - c_0\hat{\mathbf{k}}t, \mathbf{k})). \end{aligned}$$

Once again we observe that, in a uniform medium, $a_\pm$ become less regular in $\mathbf{k}$ as time grows; thus refocusing is poor.

The considerations of section 3.7 show that in the radiative transport regime the amplitudes $a_\pm$ become smoother in $\mathbf{k}$ also with initial data given by (4.9). This leads to a better refocusing as explained in section 3.5. Let us assume that the diffusion regime of section 3.8 is valid and that the kernel f is isotropic $\hat{f}(\mathbf{k}) = \hat{f}(|\mathbf{k}|)$. This requires in particular that $A_0(\mathbf{x})$ be independent of $\mathbf{x}$. We obtain that $a_\pm(T, \mathbf{x}_0, \mathbf{k}) = \tilde{a}(T, \mathbf{x}_0, |\mathbf{k}|)$; thus the refocusing formula (4.8) reduces to

$$\hat{\mathbf{u}}^B(\mathbf{k}; \mathbf{x}_0) = \tilde{a}(T, \mathbf{x}_0, |\mathbf{k}|)\hat{\mathbf{S}}(\mathbf{k}). \tag{4.11}$$

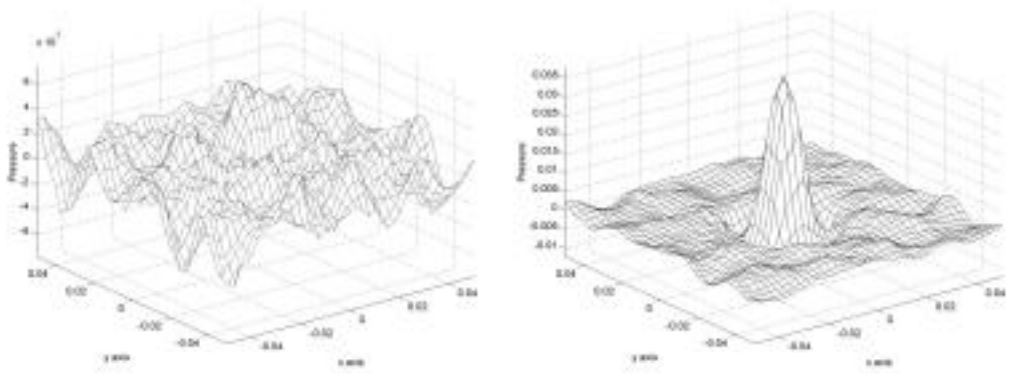


FIG. 4.1. *Zoom of the refocused signals for the numerical experiment of Figure* 2.2 *with processing* $\Gamma = I$ *(left), with a maximal amplitude of roughly* 4×10^{-3}, *and* $\Gamma = \mathrm{Diag}(0,0,0,1)$ *(right), with a maximal amplitude of roughly* 0.035.

The difference between this and the case treated in section 3.8 is that $\tilde{a}(T, \mathbf{x}, |\mathbf{k}|)$ solves the diffusion equation (3.39) with new initial conditions given by

$$\tilde{a}(0,\mathbf{x},|\mathbf{k}|) = \frac{|\chi(\mathbf{x})|^2}{4\pi|\mathbf{k}|^2}\int_{\mathbb{R}^3}\hat{f}(|\mathbf{q}|)(A_0\Gamma(\mathbf{x})\mathbf{b}_-(\mathbf{q})\cdot\mathbf{b}_+(\mathbf{q}))\delta(|\mathbf{q}|-|\mathbf{k}|)d\mathbf{q} \tag{4.12}$$
$$= \frac{|\chi(\mathbf{x})|^2}{4\pi|\mathbf{k}|^2}\int_{\mathbb{R}^3}\hat{f}(|\mathbf{q}|)(A_0\Gamma(\mathbf{x})\mathbf{b}_+(\mathbf{q})\cdot\mathbf{b}_-(\mathbf{q}))\delta(|\mathbf{q}|-|\mathbf{k}|)d\mathbf{q}.$$

When only the first component of the velocity field is measured, as in (4.10), the initial data for $\tilde{a}$ is

$$\tilde{a}(0,\mathbf{x},|\mathbf{k}|) = \frac{1}{6}|\chi(\mathbf{x})|^2\hat{f}(|\mathbf{k}|).$$

Therefore even time reversing only one component of the acoustic velocity field produces a repropagated signal that is equal to the full repropagated field up to a constant factor.

More generally, we deduce from (4.12) that a detector at $\mathbf{x}$ will contribute some refocusing for waves with wavenumber $|\mathbf{k}|$, provided that

$$\int_{S^2}\hat{f}(|\mathbf{k}|\hat{\mathbf{q}})(A_0\Gamma(\mathbf{x})\mathbf{b}_\mp(\hat{\mathbf{q}})\cdot\mathbf{b}_\pm(\hat{\mathbf{q}}))d\Omega(\hat{\mathbf{q}}) \neq 0.$$

When $f(\mathbf{x}) = f(|\mathbf{x}|)$ is radial, this property becomes independent of the wavenumber $|\mathbf{k}|$ and reduces to $\int_{S^2}(A_0\Gamma(\mathbf{x})\mathbf{b}_\mp(\hat{\mathbf{q}})\cdot\mathbf{b}_\pm(\hat{\mathbf{q}}))d\Omega(\hat{\mathbf{q}}) \neq 0$.

4.2. Numerical results. Let us come back to the numerical results presented in Figures 2.2 and 3.1. We now consider two different processings at the recording array. The first array is passive, corresponding to $\Gamma = I$, and the second array measures only pressure so that $\Gamma = \mathrm{Diag}(0,0,0,1)$. The zoom in the vicinity of $\mathbf{x}_0 = \mathbf{0}$ of the "refocused" signals is given in Figure 4.1. The left panel shows no refocusing, in accordance with physical intuition and theory. The right figure shows that refocusing indeed occurs when only pressure is recorded (and its time derivative is set to 0 in the solution of the wave equation presented in the appendix). Notice also that the

refocused signal is roughly one half the one obtained in Figure 3.1, as predicted by theory.

4.3. Refocusing of other classical waves. The preceding sections deal with the refocusing of acoustic waves. The theory can, however, be extended to more complicated linear hyperbolic systems of the form (2.3), with $A(\mathbf{x})$ a positive definite matrix, D^j symmetric matrices, and $\mathbf{u} \in \mathbb{C}^m$. These include electromagnetic and elastic waves. Their explicit representation in the form (2.3) and expressions for the matrices $A(\mathbf{x})$ and D^j in these cases may be found in [25]. For instance, the Maxwell equations

$$\frac{\partial \mathbf{E}}{\partial t} = \frac{1}{\epsilon(\mathbf{x})} \text{ curl } \mathbf{H},$$
$$\frac{\partial \mathbf{H}}{\partial t} = -\frac{1}{\mu(\mathbf{x})} \text{curl } \mathbf{E}$$

may be written in the form (2.3) with $\mathbf{u} = (\mathbf{E}, \mathbf{H}) \in \mathbb{C}^6$ and the matrix $A(\mathbf{x}) = \text{Diag}(\epsilon(\mathbf{x}), \epsilon(\mathbf{x}), \epsilon(\mathbf{x}), \mu(\mathbf{x}), \mu(\mathbf{x}), \mu(\mathbf{x}))$. Here $\epsilon(\mathbf{x})$ is the dielectric constant (not to be confused with the small parameter ε), and $\mu(\mathbf{x})$ is the magnetic permeability. The 6×6 dispersion matrix $L(\mathbf{x}, \mathbf{k})$ for the Maxwell equations is given by

$$L(\mathbf{x}, \mathbf{k}) = - \begin{pmatrix} 0 & 0 & 0 & 0 & -k_3/\epsilon(\mathbf{x}) & k_2/\epsilon(\mathbf{x}) \\ 0 & 0 & 0 & k_3/\epsilon(\mathbf{x}) & 0 & -k_1/\epsilon(\mathbf{x}) \\ 0 & 0 & 0 & -k_2/\epsilon(\mathbf{x}) & k_1/\epsilon(\mathbf{x}) & 0 \\ 0 & k_3/\mu(\mathbf{x}) & -k_2/\mu(\mathbf{x}) & 0 & 0 & 0 \\ -k_3/\mu(\mathbf{x}) & 0 & k_1/\mu(\mathbf{x}) & 0 & 0 & 0 \\ k_2/\mu(\mathbf{x}) & -k_1/\mu(\mathbf{x}) & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Generalization of our results for acoustic waves to such general systems is quite straightforward, so we concentrate only on the modifications that need be made. The time reversal procedure is exactly the same as before: a signal propagates from a localized source, is recorded, processed as in (3.3) with a general matrix $\Gamma(\mathbf{y}')$, and re-emitted into the medium. The retransmitted signal is given by (4.1). Furthermore, the equation for the adjoint Green's matrix (4.2), the definition of the Wigner transform in section 3.4, and the expression (4.7) for the repropagated signal still hold.

The analysis of the repropagated signal is reduced to the study of the Wigner distribution, which is now modified. The mode decomposition must be generalized. We recall that

$$L(\mathbf{x}, \mathbf{k}) = A_0^{-1}(\mathbf{x}) k_j D^j$$

is the $m \times m$ dispersion matrix associated with the hyperbolic system (2.3). Since $L(\mathbf{x}, \mathbf{k})$ is symmetric with respect to the inner product $\langle \mathbf{u}, \mathbf{v} \rangle_{A_0} = (A_0 \mathbf{u} \cdot \mathbf{v})$, its eigenvalues are real and its eigenvectors form a basis. We assume the existence of a time reversal matrix Γ_0 such that (3.11) holds with $\Gamma = \Gamma_0$ and such that $\Gamma_0^2 = I$. For example, for electromagnetic waves $\Gamma_0 = \text{Diag}(1, 1, 1, -1, -1, -1)$. Then the spectrum of L is symmetric about zero and the eigenvalues $\pm\omega^\alpha$ have the same multiplicity. We assume in addition that L is isotropic so that its eigenvalues have the form $\omega_\pm^\alpha(\mathbf{x}, \mathbf{k}) = \pm c^\alpha(\mathbf{x})|\mathbf{k}|$, where $c_\alpha(\mathbf{x})$ is the speed of mode α. We denote by r_α their respective multiplicities, assumed to be independent of $\mathbf{x}$ and $\mathbf{k}$ for $\mathbf{k} \neq 0$. The matrix L has a basis of eigenvectors $\mathbf{b}_\pm^{\alpha,j}(\mathbf{x}, \mathbf{k})$ such that

$$L(\mathbf{x}, \mathbf{k}) \mathbf{b}_\pm^{\alpha,j}(\mathbf{x}, \mathbf{k}) = \pm\omega^\alpha(\mathbf{x}, \mathbf{k}) \mathbf{b}_\pm^{\alpha,j}(\mathbf{x}, \mathbf{k}), \quad j = 1, \ldots, r_\alpha,$$

and $\mathbf{b}_{\pm}^{\alpha,j}$ form an orthonormal set with respect to the inner product $\langle\,,\rangle_{A_0}$. The different ω_α's correspond to different types of waves (modes). Various indices $1 \le j \le r_\alpha$ refer to different polarizations of a given mode. The eigenvectors $\mathbf{b}_{+}^{\alpha,j}$ and $\mathbf{b}_{-}^{\alpha,j}$ are related by

$$\Gamma_0 \mathbf{b}_{+}^{\alpha,j}(\mathbf{x},\mathbf{k}) = \mathbf{b}_{-}^{\alpha,j}(\mathbf{x},\mathbf{k}), \qquad \Gamma_0 \mathbf{b}_{-}^{\alpha,j}(\mathbf{x},\mathbf{k}) = \mathbf{b}_{+}^{\alpha,j}(\mathbf{x},\mathbf{k}). \tag{4.13}$$

Proposition 3.3 is then generalized as follows (see [14, 25]).

PROPOSITION 4.1. *There exist scalar functions* $a_{\pm}^{\alpha,jm}(t,\mathbf{x},\mathbf{k})$ *such that*

$$W(t,\mathbf{x},\mathbf{k}) = \sum_{\pm,\alpha,j,m} a_{\pm}^{\alpha,jm}(t,\mathbf{x},\mathbf{k}) \mathbf{b}_{\pm}^{\alpha,j}(\mathbf{x},\mathbf{k}) \otimes \mathbf{b}_{\pm}^{\alpha,m}(\mathbf{x},\mathbf{k}). \tag{4.14}$$

Here the sum runs over all possible values of $\pm$, α, *and* $1 \le j,m \le r_\alpha$.

The main content of this proposition is again that the cross terms $\mathbf{b}_{\pm}^{\alpha,j}(\mathbf{x},\mathbf{k}) \otimes \mathbf{b}_{\mp}^{\beta,m}(\mathbf{x},\mathbf{k})$ do not contribute, and neither do the terms $\mathbf{b}_{\pm}^{\alpha,j}(\mathbf{x},\mathbf{k}) \otimes \mathbf{b}_{\pm}^{\alpha',m}(\mathbf{x},\mathbf{k})$ when $\alpha \neq \alpha'$. This is because modes propagating with different speeds do not interfere constructively in the high frequency limit.

We may now insert expression (4.14) into (4.7) and obtain the following generalization of (4.8):

$$\begin{aligned}\hat{\mathbf{u}}^B(\mathbf{k};\mathbf{x}_0) = \sum_{\alpha,j,m} \Big[& a_{-}^{\alpha,mj}(T,\mathbf{x}_0,\mathbf{k}) \hat{S}_{+}^{\alpha,j}(\mathbf{x}_0,\mathbf{k}) \mathbf{b}_{+}^{\alpha,m}(\mathbf{x}_0,\mathbf{k}) \\ & + a_{+}^{\alpha,mj}(T,\mathbf{x}_0,\mathbf{k}) \hat{S}_{-}^{\alpha,j}(\mathbf{x}_0,\mathbf{k}) \mathbf{b}_{-}^{\alpha,m}(\mathbf{x}_0,\mathbf{k}) \Big],\end{aligned} \tag{4.15}$$

where $\hat{S}_{\pm}^{\alpha,j}(\mathbf{k}) = (A(\mathbf{x}_0)\hat{\mathbf{S}}(\mathbf{k}) \cdot \mathbf{b}_{\pm}^{\alpha,j}(\mathbf{x}_0,\mathbf{k}))$. This formula tells us that only the modes that are present in the initial source ($\hat{S}_{\pm}^{\alpha,j}(\mathbf{k}) \neq 0$) will be present in the back-propagated signal but possibly with a different polarization, that is, $j \neq m$.

The initial conditions for the modes $a_{\pm}^{\alpha,jm}$ are given by

$$a_{\pm}^{\alpha,jm}(0,\mathbf{x},\mathbf{k}) = |\chi(\mathbf{x})|^2 \hat{f}(\mathbf{k}) (A(\mathbf{x})\Gamma(\mathbf{x}) \mathbf{b}_{\mp}^{\alpha,m}(\mathbf{x},\mathbf{k}) \cdot \mathbf{b}_{\pm}^{\alpha,j}(\mathbf{x},\mathbf{k})), \tag{4.16}$$

which generalizes (4.9). When $\Gamma(\mathbf{x}) \equiv I$, we again obtain that $a_{\pm}^{\alpha,jm}(0,\mathbf{x},\mathbf{k}) \equiv 0$, i.e., there is no refocusing as physically expected. When $\Gamma(\mathbf{x}) \equiv \Gamma_0$, we have for all α that

$$a_{\pm}^{\alpha,jm}(0,\mathbf{x},\mathbf{k}) = |\chi(\mathbf{x})|^2 \hat{f}(\mathbf{k}) \delta_{jm}.$$

In a uniform medium the amplitudes $a_{\pm}^{\alpha,jm}$ satisfy an uncoupled system of free transport equations (3.33),

$$\frac{\partial a_{\pm}^{\alpha,jm}}{\partial t} \pm c_\alpha \hat{\mathbf{k}} \cdot \nabla_{\mathbf{x}} a_{\pm}^{\alpha,jm} = 0, \tag{4.17}$$

which have no smoothing effect, and hence refocusing in a homogeneous medium is still poor. When $f(\mathbf{x}) = \delta(\mathbf{x})$ and $\Omega = \mathbb{R}^3$, so that $\chi(\mathbf{x}) \equiv 1$, we still have that $a_{\pm}^{\alpha,jm}(T,\mathbf{x}_0,\mathbf{k}) = \delta_{jm}$ and refocusing is again perfect, that is, $\mathbf{u}^B(\boldsymbol{\xi};\mathbf{x}_0) = \mathbf{S}(\boldsymbol{\xi})$, as may be seen from (4.15).

4.4. The diffusive regime. The radiative transport regime holds when the matrices $A(\mathbf{x})$ have the form

$$A(\mathbf{x}) = A_0(\mathbf{x}) + \sqrt{\varepsilon} A_1\left(\frac{\mathbf{x}}{\varepsilon}\right),$$

as in (3.35). Then the $r_\alpha \times r_\alpha$ coherence matrices $w^\alpha_\pm$ with entries $w^\alpha_{\pm,jm} = a^{\alpha,jm}_\pm$ satisfy a system of matrix-valued radiative transport equations (see [25] for the details) similar to (3.36). The matrix transport equations simplify considerably in the diffusive regime, such as the one considered in section 3.8 when waves propagate over large distances and long times. We assume for simplicity that $A_0 = A_0(\mathbf{x})$ and $\Gamma = \Gamma(\mathbf{x})$ are independent of $\mathbf{x}$. Polarization is lost in this regime; that is, $a^{\alpha,jm}(t,\mathbf{x},\mathbf{k}) = 0$ for $j \neq m$, and wave energy is equidistributed over all directions. This implies that

$$a^{\alpha,jj}_+(t,\mathbf{x},\mathbf{k}) = a^{\alpha,jj}_-(t,\mathbf{x},\mathbf{k}) = a_\alpha(t,\mathbf{x},|\mathbf{k}|)$$

so that $a^{\alpha,jj}$ is independent of $j = 1,\ldots,r_\alpha$ and of the direction $\hat{\mathbf{k}} = \mathbf{k}/|\mathbf{k}|$. Furthermore, because of multiple scattering, a universal equipartition regime takes place so that

$$a_\alpha(t,\mathbf{x}_0,|\mathbf{k}|) = \phi(t,\mathbf{x}_0,c_\alpha|\mathbf{k}|), \tag{4.18}$$

where $\phi(t,\mathbf{x},\omega)$ solves a diffusion equation in $\mathbf{x}$ like (3.39) (see [25]). The diffusion coefficient $D(\omega)$ may be expressed explicitly in terms of the power spectra of the medium fluctuations [25]. Using (4.16) and (4.18), we obtain when f is isotropic the following initial data for the function ϕ:

$$\phi(0,\mathbf{x},\omega) = \frac{1}{4\pi}|\chi(\mathbf{x})|^2 \int_{S^2} \frac{2}{|\alpha|} \sum_{j,\omega_\alpha>0} \hat{f}\left(\frac{\omega}{c_\alpha}\right) (A_0\Gamma \mathbf{b}^{\alpha,j}_-(\hat{\mathbf{k}}), \mathbf{b}^{\alpha,j}_+(\hat{\mathbf{k}}))d\Omega(\hat{\mathbf{k}}), \tag{4.19}$$

where $|\alpha|$ is the number of nonvanishing eigenvalues of $L(\mathbf{x},\mathbf{k})$, and $d\Omega(\hat{\mathbf{k}})$ is the Lebesgue measure on the unit sphere S^2.

Let us assume that nonpropagating modes are absent in the initial source $\mathbf{S}(\mathbf{x})$; that is, $\hat{S}^j_0(\mathbf{k}) = 0$, with the subscript zero referring to modes corresponding to $\omega_0 = 0$. Then (4.15) becomes

$$\hat{\mathbf{u}}(\mathbf{k};\mathbf{x}_0) = \sum_{\alpha,j} \phi(T,\mathbf{x}_0,c_\alpha|\mathbf{k}|)\left[\hat{S}^{\alpha,j}_+(\mathbf{k})\mathbf{b}^{\alpha,j}_+(\mathbf{x}_0,\mathbf{k}) + \hat{S}^{\alpha,j}_-(\mathbf{k})\mathbf{b}^{\alpha,j}_-(\mathbf{x}_0,\mathbf{k})\right]. \tag{4.20}$$

This is an explicit expression for the repropagated signal in the diffusive regime, where ϕ solves the diffusion equation (3.39) with initial conditions (4.19).

5. Conclusions. This paper presents a theory that quantitatively describes the refocusing phenomena in time reversal acoustics as well as for more general processings of acoustic and other classical waves. We show that the back-propagated signal may be expressed as the convolution (1.1) of the original source $\mathbf{S}$ with a filter F. The quality of the refocusing is therefore determined by the spatial decay of the kernel F. For acoustic waves, the explicit expression (3.27) relates F to the Wigner distribution of certain solutions of the wave equation. The decay of F is related to the smoothness in the phase space of the amplitudes $a_j(t,\mathbf{x},\mathbf{k})$ defined in Proposition 3.3. The latter satisfy free transport equations in homogeneous media, which sharpens the gradients

of a_j and leads to poor refocusing. In contrast, the amplitudes a_j satisfy the radiative transport equation (3.36) in heterogeneous media, which has a smoothing effect. This leads to a rapid spatial decay of the filter F and a better refocusing. For longer times, a_j satisfies a diffusion equation. This allows for an explicit expression (3.42)–(3.43) of the time reversed signal. The same theory holds for more general waves and more general processing procedures at the recording array, which allows us to describe the refocusing of electromagnetic waves when only one component of the electric field is measured, for instance.

Appendix. This appendix presents the details of the numerical simulation of (2.9). We assume that ρ is constant and that only $\kappa(\mathbf{x})$ fluctuates. We can therefore recast (2.9) as

$$\frac{\partial^2 p}{\partial t^2} - c^2(\mathbf{x})\Delta p = 0.$$

The above wave equation is discretized using a second-order scheme (three point stencil in every variable) both in time and space. The resolution in time is explicit and time reversible; i.e., the equation that yields $p(t_{n+1})$ from $p(t_{n-1})$ and $p(t_n)$ can be used to retrieve $p(t_{n-1})$ exactly from $p(t_n)$ and $p(t_{n+1})$. We write $c^2(\mathbf{x}) = c_0^2 + c_1^2(\mathbf{x})$. The average velocity is $c_0^2 = 1$. The random part c_1^2 has been constructed as follows. Let $2N \times 2N$ be the number of spatial grid points and $c^2_{1;n,m}$ be the value of c_1^2 at the grid point (n, m). The values $c^2_{1;2n,2m}$ have been chosen independently and uniformly on $(-r, r)$ with $r < 1/2$. The value of c_1^2 is then set constant on four adjacent pixels by enforcing that $c^2_{1;2n-1,2m} = c^2_{1;2n-1,2m-1} = c^2_{1;2n,2m-1} = c^2_{1;2n,2m}$ for $1 \le n, m \le N$. In all simulations, we have $N = 200$, which generates a grid of $400^2 = 1.6 \times 10^4$ points. The time step has been chosen so that the CFL condition $\delta t < \min_{\mathbf{x}} c(\mathbf{x})/(2N)$ is ensured. The fluctuations of the velocity field have been chosen larger than those in the weak-fluctuation regime analyzed in this paper. This is to ensure that sufficient mixing occurs on the limited 400×400 grid that fits on a personal computer. The domain of truncation has also been centered to maximize the mixing of the recorded signal. Off-centered domains of truncation give very similar results, albeit with a smaller signal-to-noise ratio.

Acknowledgments. We would like to thank Knut Solna for fruitful discussions during the preparation of this work. We are indebted to George Papanicolaou for his contributions to the analysis of time reversal, which lie at the core of this paper. This work would also not have been possible without the numerous exchanges we benefited from at the Stanford MGSS summer school.

REFERENCES

[1] G. Bal and L. Ryzhik, *Time reversal for classical waves in random media*, C. R. Acad. Sci. Paris Sér. I Math., 333 (2001), pp. 1041–1046.

[2] C. Bardos and M. Fink, *Mathematical foundations of the time reversal mirror*, Asymptot. Anal., 29 (2002), pp. 157–182.

[3] P. Blomgren, G. Papanicolaou, and H. Zhao, *Super-resolution in time-reversal acoustics*, J. Acoust. Soc. Amer., 11 (2002), pp. 230–248.

[4] L. Borcea, C. Tsogka, G. Papanicolaou, and J. Berryman, *Imaging and time reversal in random media*, Inverse Problems, 18 (2002), pp. 1247–1279.

[5] J. F. Clouet and J. P. Fouque, *A time-reversal method for an acoustical pulse propagating in randomly layered media*, Wave Motion, 25 (1997), pp. 361–368.

[6] R. Dautray and J.-L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology*, Vol. 6, Springer-Verlag, Berlin, 1993.

[7] A. Derode, P. Roux, and M. Fink, *Robust acoustic time-reversal with high-order multiple-scattering*, Phys. Rev. Lett., 75 (1995), pp. 4206–4209.
[8] D. R. Dowling and D. R. Jackson, *Narrow-band performance of phase-conjugate arrays in dynamic random media*, J. Acoust. Soc. Amer., 91 (1992), pp. 3257–3277.
[9] L. Erdös and H. T. Yau, *Linear Boltzmann equation as the weak coupling limit of a random Schrödinger equation*, Comm. Pure Appl. Math., 53 (2000), pp. 667–735.
[10] N. Ewodo, *Refocusing of a time-reversed acoustic pulse propagating in randomly layered media*, J. Statist. Phys., 104 (2001), pp. 1253–1272.
[11] M. Fink, *Time reversed acoustics*, Physics Today, 50 (1997), pp. 34–40.
[12] M. Fink, *Chaos and time-reversed acoustics*, Phys. Scripta, 90 (2001), pp. 268–277.
[13] M. Fink and C. Prada, *Acoustic time-reversal mirrors*, Inverse Problems, 17 (2001), pp. R1–R38.
[14] P. Gérard, P. A. Markowich, N. J. Mauser, and F. Poupaud, *Homogenization limits and Wigner transforms*, Comm. Pure Appl. Math., 50 (1997), pp. 323–380.
[15] F. Golse, P.-L. Lions, B. Perthame, and R. Sentis, *Regularity of the moments of the solution of a transport equation*, J. Funct. Anal., 76 (1988), pp. 110–125.
[16] W. Hodgkiss, H. Song, W. Kuperman, T. Akal, C. Ferla, and D. Jackson, *A long-range and variable focus phase-conjugation experiment in shallow water*, J. Acoust. Soc. Amer., 105 (1999), pp. 1597–1604.
[17] A. Ishimaru, *Wave Propagation and Scattering in Random Media*, Academics, New York, 1978.
[18] S. R. Khosla and D. R. Dowling, *Time-reversing array retrofocusing in noisy environments*, J. Acoust. Soc. Amer., 109 (2001), pp. 538–546.
[19] W. Kuperman, W. Hodgkiss, H. Song, T. Akal, C. Ferla, and D. Jackson, *Phase-conjugation in the ocean*, J. Acoust. Soc. Amer., 102 (1997), pp. 1–16.
[20] E. W. Larsen and J. B. Keller, *Asymptotic solution of neutron transport problems for small mean free paths*, J. Math. Phys., 15 (1974), pp. 75–81.
[21] P.-L. Lions and T. Paul, *Sur les mesures de Wigner*, Rev. Mat. Iberoamericana, 9 (1993), pp. 553–618.
[22] M. Mokhtar-Kharroubi, *Mathematical Topics in Neutron Transport Theory*, World Scientific, Singapore, 1997.
[23] G. Papanicolaou, L. Ryzhik, and K. Solna, *The parabolic wave approximation and time reversal*, Mat. Contemp., 23 (2002), pp. 139–159.
[24] G. Papanicolaou, L. Ryzhik, and K. Solna, *Statistic stability in time reversal*, SIAM J. Appl. Math., to appear.
[25] L. Ryzhik, G. Papanicolaou, and J. B. Keller, *Transport equations for elastic and other waves in random media*, Wave Motion, 24 (1996), pp. 327–370.
[26] P. Sheng, *Introduction to Wave Scattering, Localization and Mesoscopic Phenomena*, Academic Press, New York, 1995.
[27] H. Spohn, *Derivation of the transport equation for electrons moving through random impurities*, J. Statist. Phys., 17 (1977), pp. 385–412.

SIAM J. APPL. MATH.
Vol. 63, No. 5, pp. 1499–1519

HEAT-INDUCED STRETCHING OF A GLASS TUBE UNDER TENSION: APPLICATION TO GLASS MICROELECTRODES*

HUAXIONG HUANG†, ROBERT M. MIURA‡, WILLIAM P. IRELAND§, AND ERNEST PUIL¶

Abstract. Deformation of glass using heat occurs in many industrial and artistic applications, including the manufacturing of laboratory glass products, drawing of fiber optics, and hand-blown artistic creations. The formation of glass objects is an art, but the trial-and-error aspect of the procedures can be reduced by development of a systematic theory, especially when the objects are formed using mechanical means. Glass microelectrodes are ubiquitous in experimental studies of the electrophysiology of biological cells and their membranes, and the "pulling" of these electrodes is based on trial-and-error. To make this process more systematic, we derive a model for glass microelectrode formation using a coil heater with a gravity-forced electrode puller, assuming that the glass tube is an incompressible, viscous fluid. The model is one-dimensional, and the effects of thermal radiation from the coil heater are essential in the formation process. A breaking stress criterion is imposed to fracture the glass tube, forming the electrode tip. The difficulty with the moving free end is avoided by introducing a quasi-Lagrangian coordinate system. The model equations are solved using an adaptive moving grid to account for the local stretching of the glass. A number of examples using a double-pull paradigm have been computed to illustrate the dependence of the electrode shape and tip diameter on the heater temperature and the ratio between the inner and outer radii.

Key words. glass tube, microelectrodes, viscous incompressible fluid, radiation heat transfer, finite-difference method, adaptive grid

AMS subject classifications. 76D99, 80A20, 65M06, 65M50

DOI. 10.1137/S0036139901393469

1. Introduction.

Glass objects are routinely produced in industry and in some technical arts. Such objects include laboratory products, fiber optics, and hand-blown glassware. In some applications, heat is applied to soften the glass during the formation of these glass objects, using trial-and-error procedures. However, when these procedures involve mechanical devices, the use of mathematical models provides a more systematic approach. For example, the pulling of glass fiber optics has been studied by Fitt et al. [1] and other researchers [4, 12]. Most of these studies focus on isothermal or prescribed temperature conditions and on given pulling velocities.

Glass microelectrodes have played an essential role in cell electrophysiology for decades and will continue to be an important tool in the future. These micropipettes are used to inject electric current and dyes into cells and to measure membrane potentials by insertion through cellular membranes or formation of a patch clamp of

*Received by the editors August 7, 2001; accepted for publication (in revised form) December 2, 2002; published electronically June 12, 2003. The research of the first and second authors was supported in part by grants from the Natural Sciences and Engineering Research Council (NSERC) of Canada.

http://www.siam.org/journals/siap/63-5/39346.html

†Department of Mathematics and Statistics, York University, Toronto, ON, Canada M3J 1P3 (hhuang@yorku.ca).

‡Department of Mathematics, Institute of Applied Mathematics, and Department of Pharmacology and Therapeutics, University of British Columbia, Vancouver, BC, Canada V6T 1Z2. Current address: Department of Mathematical Sciences, New Jersey Institute of Technology, Newark, NJ 07102 (miura@njit.edu).

§Department of Anatomy and Physiology, University of Prince Edward Island, Charlottetown, PE, Canada C1A 4P3 (ireland@upei.ca).

¶Department of Pharmacology and Therapeutics, University of British Columbia, Vancouver, BC, Canada V6T 1Z3 (puil@neuro.pharmacology.ubc.ca).

the membrane. The data collected give information about membrane electrical properties in the presence of voltage-gated and receptor-gated ion channels and with the application of drugs. Laboratories using these microelectrodes usually make them on a daily basis, using commercially available glass tubes and electrode pullers that use coil heaters for softening the glass tubes during the stretching procedure.

There are four experimentally relevant parameters that can be measured on glass microelectrodes. They are tip length, tip diameter, electrode resistance, and electrode capacitance. Tip length is significant because this determines the physical strength of the electrode and how easily it will penetrate tissue and cells. A short tip with steep taper is robust but does not penetrate tissue easily. The converse is true of long, gently tapering tips.

Tip diameter is important because it determines whether the electrode is suitable for intracellular recording or for patch recording. Intracellular recordings are made with electrodes having very narrow tips (approximately 0.1 micron), which can be obtained using a single-pull electrode puller. Generally, patch clamping requires a larger tip diameter, of approximately 1 micron outside diameter and 0.5 micron inside diameter, obtained using a double-pull electrode puller. There is a correlation between tip length and distal tip diameter; for example, using a specific puller with varying heater widths and geometries, long tips were of narrower gauge at the ends than short tips [2].

Capacitance and resistance are properties of the electrode that determine its confounding effect on measurements taken from neurons. In order to get accurate data, one needs to measure these properties and compensate for them. These properties are functions of the physical form of the electrode, the properties of the glass used to make it, and the electrolyte used to fill it. Therefore, since the manufacturing process determines this physical form, the process should be understood so that it can be tailored to create electrodes fitting the experimentalist's needs.

Capacitance and electrode resistance as a function of the pulling parameters can be measured but not at the same time as precise measurement of tip diameter or shank geometry. The procedure for measuring tip size accurately involves using the scanning electron microscope (SEM), which is time-consuming and expensive. There are means to do this nondestructively [3]. Other methods have been explored, including measuring the rate of flow of a solution down the shank of an electrode and using a mathematical model to find tip size [9]. Mittman et al. [7] estimated tip diameter from the pressure needed to force bubbles from a microelectrode immersed in methanol. From knowledge of the tip diameter and length of the electrode's tapering shank, capacitance also has been computed using the approximation that the shank is a cone [11].

However, the exact relationship between the variables in the actual manufacturing process (heater geometry, rate of pulling the glass tube, length of first pull (for a patch electrode), rate of the second pull, etc.) and electrode properties is usually determined empirically by a method of trial-and-error. Though some work has been done on the influence of heater geometry and width on electrode form (see [2]), in general the process is not well determined. In this paper, our objective is to develop a basic mathematical model for the formation process of these glass microelectrodes and, through computer simulations, understand the complex interaction of variables in the manufacturing process with the properties of the resulting electrode. This has the advantages that many different types of pullers can be simulated, the effects of many parameters can be explored rapidly, and predictions to guide the formation of microelectrodes using existing pullers as well as future design of electrode pullers can

be made. There are several types of pullers in common use today. The differences among them include heater size and shape, method of application of the force (gravity, electromagnetic) to extend the glass tube, use of single and double pulls, and auxiliary cooling methods (e.g., puffs of air).

The main focus of this paper is on the pulling of glass microelectrodes. We will not analyze the rupture of the glass tube at the end of the pulling process. Instead, we will impose a breaking stress criterion to terminate the stretching of the glass tube when the viscous stress inside the glass tube exceeds the critical value of the stress, i.e., the breaking stress. However, the theory developed here can be generalized and applied to a broader class of problems. From the fluid dynamics point of view, the breaking of the glass tube may be due to the collapse of the glass wall. Furthermore, it also is possible that the inner radius of the glass tube shrinks to zero, resulting in failure of this production method to produce a functional electrode. Three major factors that determine these outcomes are the surface tension, viscous stress (tension), and pressure, which are all affected by the temperature of the glass. Previous studies of the stability of free surfaces of isothermal fluid jets as well as the effects of pressure and temperature may be used to predict the behavior of the surfaces in the stretched section of the glass tube. This will be the subject of a future paper.

The paper is organized as follows. In section 2, the assumptions for the model and derivation of the model equations are given. Details of the control-volume approach used in this derivation are specified in Appendix A. The effects of thermal radiation from the coil heater on the glass tube require the determination of the geometric factors from the coil and from the background to the glass; see Appendix B. Also described are the breaking stress criterion, which is used to terminate the numerical computations, and the glass properties used in the model. The added complication of a moving boundary due to the stretching of the glass tube is avoided by introducing a quasi-Lagrangian coordinate system. The actual length of the tube evolves according to a system of ordinary differential equations.

The finite-difference method is described in section 3 and is used to compute the solutions numerically in space and time. In section 4, several different cases of parameter values are given as examples for the numerical computations, and results for a double pull case are described. The paper closes with a discussion in section 5 on the limitation of the present model and future work for improving it.

2. A model for glass microelectrode formation.

2.1. Derivation of the model. In the model equations developed here for glass microelectrode formation, we account for certain types of electrode pullers which are capable of single and multiple pulls. The microelectrode starts off as a glass tube that is held vertically, being clamped at the top and with a weight hanging from the bottom; see Figure 2.1(a). The tube then is heated nonuniformly in the longitudinal direction by radiation from an axially symmetric heated wire coil, which surrounds the tube near the middle at the initial time. Also, there is radiation loss to the background, and we assume there is no radiant energy passing through the glass. The tube heats up to the softening point and begins to stretch due to the weight; see Figure 2.1(b). Examination of microelectrodes formed in this way shows that symmetrical radial contraction of the glass tube occurs as longitudinal stretching takes place. Although this is a two-dimensional formation process, we treat the problem in one space dimension along the tube length by averaging over the cross-sectional area. This is justified because of the small diameter of the tube compared with its length.

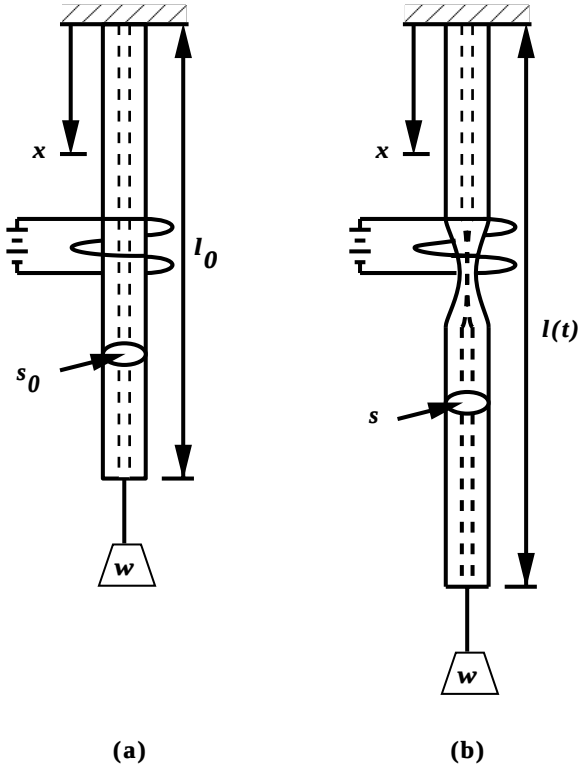


FIG. 2.1. *Schematics of the glass tube before* (a) *and during* (b) *the stretching.*

In developing a mathematical model to describe the heating of a glass tube to form a microelectrode, we initially will be working with an Eulerian coordinate system, with x and t as the spatial and temporal coordinates, respectively. The mass of the tube is small compared to the hanging weight w and will be ignored in the following derivation. At $t = 0$, we assume a uniform glass tube of length ℓ_0 with a circular annular cross section having constant outer radius R_0 and constant inner radius r_0. The resulting area of the annular cross section is denoted by $s_0 = \pi(R_0^2 - r_0^2)$, and the coordinate along the glass tube is given by $0 \le x \le \ell_0$.

We assume that the stretching glass tube maintains a circular annular cross section with outer radius $R(x,t)$ and inner radius $r(x,t)$ at a given location x and time t (thus, $R_0 = R(x,0)$ and $r_0 = r(x,0)$). The cross-sectional area of the glass tube is given by $s(x,t) = \pi(R^2 - r^2)$ (thus, $s_0 = s(x,0)$). The length of the tube at time t is denoted by $\ell(t)$ (thus, $\ell_0 = \ell(0)$), and we refer to the location $x = \ell(t)$ as the "free end." The velocity of the glass at x at time t is given by $u(x,t)$. Thus the extensional strain rate is given by $\partial u/\partial x$, which corresponds to the spatial variation of the glass velocity.

From experiments, the dependence of density changes on temperature is negligible. Therefore, to simplify the model equations, we assume that the glass is an incompressible fluid with constant density and that the stretching is axial with concomitant shrinking of the cross-sectional area, $s(x,t)$. Using a control-volume approach (cf. Appendix A), we have the continuity equation

$$\frac{\partial \gamma}{\partial t} + u\frac{\partial \gamma}{\partial x} + \gamma\frac{\partial u}{\partial x} = 0 \tag{2.1}$$

where $\gamma(x,t) = s(x,t)/s_0$.

The axial stress at each point in the glass tube is given by

$$\text{axial stress} = \frac{F}{s(x,t)} \tag{2.2}$$

where F is the axial force and is uniform over the length of the tube because we have neglected its mass.[1] To determine F, we use Newton's second law to describe the motion of the applied weight w,

$$w - F = \frac{wa}{g} \tag{2.3}$$

where the acceleration $a = d^2\ell/dt^2$ is measured at the end of the glass tube and g is the gravitational acceleration. Thus the glass satisfies the constitutive relation

$$\frac{w}{s(x,t)}\left(1 - \frac{a}{g}\right) = 3\mu\frac{\partial u}{\partial x} \tag{2.4}$$

where μ is the shear viscosity and 3μ is the coefficient of viscosity for axial tension. We have assumed that viscous stress is dominant over the inertia of the glass and the surface tension of the glass-air interface. Therefore, the Reynolds number, $Re = \rho\mathcal{U}\mathcal{L}/\bar{\mu}$, is small[2] where ρ is the glass density and $\mathcal{U}$, $\mathcal{L}$, and $\bar{\mu}$ are the characteristic velocity, length, and viscosity, respectively. Note that $\mu = \mu(\theta(x,t))$, where $\theta(x,t)$ is the temperature of the glass at (x,t). (The issue of surface tension is addressed briefly in section 5.) Combining (2.1) and (2.4) yields

$$\frac{\partial\gamma}{\partial t} + u\frac{\partial\gamma}{\partial x} = -\frac{\mathcal{P}}{\mu}\left(1 - \frac{a}{g}\right) \tag{2.5}$$

where $\mathcal{P} \equiv w/3s_0 = \text{constant}$.

The temperature distribution $\theta(x,t),\ 0 < x < \ell(t),\ t > 0$, is subject to the initial condition $\theta(x,0) = \theta_0(x),\ 0 \le x \le \ell(0)$. We apply a standard control-volume method for deriving the energy equation (see Appendix A), resulting in

$$\rho\left(\frac{\partial c_p\theta}{\partial t} + u\frac{\partial c_p\theta}{\partial x}\right) = \frac{1}{s}\frac{\partial}{\partial x}\left(s\kappa(\theta)\frac{\partial\theta}{\partial x}\right) + E_R \tag{2.6}$$

where c_p and κ are the specific heat and thermal conductivity of the glass, respectively, and E_R represents the transport of thermal energy to the glass tube by radiation. This radiation term is given by

$$\begin{aligned} E_R = 2k\sqrt{\frac{\pi}{s(1-\beta^2)}}\Bigg[&F_{hg}\frac{\varepsilon_h\alpha}{1-(1-\alpha)(1-\varepsilon_h)}\left(\theta_h^4 - \theta^4\right) \\ &+ F_{bg}\frac{\varepsilon_b\alpha}{1-(1-\alpha)(1-\varepsilon_b)}\left(\theta_b^4 - \theta^4\right)\Bigg] \end{aligned} \tag{2.7}$$

where k is the Boltzmann constant, α is the absorptivity of the glass to radiative thermal energy, ε_h and ε_b are the emissivities of the heater and background, respectively,

[1] We refer the reader to Appendix A for more details.

[2] This assumption will be justified in the next section.

and $\theta_h(x,t)$ and $\theta_b(x,t)$ are the temperatures of the heater and the background, respectively. The quantities F_{hg}, F_{bg}, and F_{bh} are geometric factors between the heater and the glass tube, the background and the glass tube, and the background and the heater, respectively. We note that $F_{gh} = F_{hg}$, and similarly for the other two geometric factors. (The derivation of the geometric factors is given in Appendix B.) Finally, $\beta = r/R$ is the ratio of the inner and outer radii; thus $s = \pi R^2(1-\beta^2)$. In general, β is a function of x and t and can be determined by treating the interfaces between the glass tube and the air as free boundaries. When the viscosity of the glass is small, i.e., the Reynolds number is large, it can be shown that β is approximately a constant, which is not the case for moderate and large values of the viscosity. Further discussion of these issues will be given in section 4.

The initial conditions are given by $\theta(x,0) = \theta_0$, $\gamma(x,0) = 1$, and $u(x,0) = 0$. At the boundaries $x = 0$ and $x = \ell(t)$, θ is fixed at θ_a, which is the ambient temperature. Since the boundary at $x = 0$ is fixed, the velocity $u(0,t) = 0$. At the moving boundary, $x = \ell(t)$, the velocity is given by $u(\ell(t),t) = d\ell(t)/dt$. In summary, the governing equations of our problem are the coupled equations, (2.5) and (2.6), with the specified initial and boundary conditions.

Finally, the criterion for determining when the glass tube breaks is the breaking stress for the glass. The maximal stress in the tube can be computed, and when it reaches the assumed breaking stress, the tube is considered broken. This procedure requires an accurate determination of the stress distribution in the tube both spatially and temporally because it varies rapidly as the tube thins down to small diameters. In the present study, we will find that, in some cases, the breaking stress is not achieved after the maximum allowed extension of the glass tube is reached, and this results in an extremely fine tip with a long shank, which is not usable in experiments. When the breaking stress is achieved, the glass tube will break into two pieces, and the resulting shapes of the two electrodes produced then are determined.

3. Numerical procedures. Since the governing equations are nonlinear, we seek the solution through numerical means. The main purpose of the numerical tests in this paper is to investigate the effects on the shape of the glass microelectrode tips obtained by changing the heater temperature θ_h. This parameter has been chosen since it is easy to adjust in the laboratory. We will simulate a typical two-pull electrode formation process used in the laboratory when the glass microelectrode is produced, which will be described briefly in the following.

3.1. Coordinate transformation. The moving boundary introduces an extra complication into the model even though the solution behavior is quite regular. To avoid this added complication in the numerical computations, we choose a quasi-Lagrangian coordinate system in which the moving boundary is fixed. However, this coordinate system does not follow the material motion as in true Lagrangian coordinates. This coordinate system can be obtained by a simple transformation from the Eulerian coordinates, which we now describe.

We derive the governing equations under the new coordinates (ξ,τ) defined by

$$\xi = \xi(x,t) \quad \text{or} \quad x = X(\xi,\tau), \tag{3.1}$$

$$\tau = t, \tag{3.2}$$

where $\xi(0,t) = 0$ and $\xi(\ell(t),t) = 1$, and with new dependent variables

$$u(x,t) = U(\xi,\tau), \quad \theta(x,t) = \Theta(\xi,\tau), \quad \ell(t) = L(\tau), \tag{3.3}$$

$$a(t) = A(\tau), \quad \gamma(x,t) = \Gamma(\xi,\tau), \quad \mu(x,t) = \nu(\xi,\tau). \tag{3.4}$$

The relationships between the derivatives in the new (ξ,τ) and the old (x,t) coordinates are

$$\frac{\partial}{\partial x} = G\frac{\partial}{\partial \xi}, \tag{3.5}$$

$$\begin{aligned}\frac{\partial}{\partial t} &= \frac{\partial}{\partial \tau} - V\frac{\partial}{\partial x} \\ &= \frac{\partial}{\partial \tau} - VG\frac{\partial}{\partial \xi}\end{aligned} \tag{3.6}$$

where $V = \partial X/\partial\tau$ is the "grid velocity," and $G = \partial\xi/\partial x$ is the "reciprocal grid stretching ratio," which is to be specified.

From (2.4), we have

$$\frac{\partial u}{\partial x} = \frac{\mathcal{P}}{\gamma\mu}\left(1 - \frac{a}{g}\right), \tag{3.7}$$

and in the new coordinates, this becomes

$$\frac{\partial U}{\partial \xi} = \frac{\mathcal{P}}{G\Gamma\nu}\left(1 - \frac{A}{g}\right) \tag{3.8}$$

where $\nu = \gamma\mu$. The constitutive equation for γ, (2.5), under the transformation can be written as

$$\frac{\partial \Gamma}{\partial \tau} + (U - V)G\frac{\partial \Gamma}{\partial \xi} = -\frac{\mathcal{P}}{\nu}\left(1 - \frac{A}{g}\right). \tag{3.9}$$

The energy equation, (2.6), now is written in the new coordinates as

$$\rho\left(\frac{\partial c_p\Theta}{\partial \tau} + G(U - V)\frac{\partial c_p\Theta}{\partial \xi}\right) = G\Gamma\frac{\partial}{\partial \xi}\left(\frac{\kappa(\Theta)G}{\Gamma}\frac{\partial \Theta}{\partial \xi}\right) + E_R. \tag{3.10}$$

In order to solve (3.8)–(3.9), we need to compute conditions at the free end, $L(\tau)$. Integrating (3.8) from $\xi = 0$ to $\xi = 1$ and using $W(\tau) = U(1,\tau) = dL/d\tau$ and $dW/d\tau = A(\tau) = d^2L/d\tau^2$, we obtain

$$\frac{dL}{d\tau} = \left(1 - \frac{1}{g}\frac{d^2L}{d\tau^2}\right)\int_0^1 \frac{\mathcal{P}}{G\Gamma\nu}d\xi. \tag{3.11}$$

The information for the moving boundary (the length $L(\tau)$ and velocity $W(\tau)$) is obtained by solving a system of ordinary differential equations derived from (3.11):

$$\frac{dW}{d\tau} = g - \frac{1}{\mathcal{I}}W, \tag{3.12}$$

$$\frac{dL}{d\tau} = W \tag{3.13}$$

where $\mathcal{I} = \int_0^1 \mathcal{P}/(\nu g G\Gamma)d\xi$.

3.2. Adaptive moving grid generation. A special feature of the problem being studied here is that the solutions (temperature, radius of the tube, etc.) vary rapidly in the region near the heater and much more slowly in regions away from the heater. A uniform grid results in grid points that are unnecessarily dense in some regions and not sufficiently dense in other regions. Therefore, an adaptive grid is desirable in the computation.

The approaches used in adaptive grid generation can be divided into two general categories. When the grid generation and the physical equations are dealt with separately, the approach is called *static regridding*. This approach is usually robust and relatively simple to use. If the grid is generated simultaneously with the solution of the physical equations, then the approach belongs to the *moving mesh* methods.

We use a simple moving mesh method by solving a partial differential equation for the coordinate transformation

$$\xi = \xi(x,t), \tag{3.14}$$

$$\tau = t \tag{3.15}$$

between the domain $\xi \in [0,1]$ and $x \in [0, L(\tau)]$. The differential equation, based on the equidistribution principle, can be written as

$$\frac{\partial X}{\partial \tau} = \frac{1}{\tau_r}\frac{\partial}{\partial \xi}\left(M\frac{\partial X}{d\xi}\right), \qquad 0 < \xi < 1. \tag{3.16}$$

Here τ_r is a (small) relaxation parameter and $M = \sqrt{(1+p\Gamma^{-2})/(1+\Gamma^{-2})}$ is the monitor function, which is chosen such that the grid is dense where Γ is small. The smoothing parameter p determines the ratio of finest and coarsest grid sizes and therefore prevents the distribution of too many points in the region when Γ is small. When solved numerically on a uniform grid $0 = \xi_0 < \xi_1 < \cdots < \xi_N = 1$, subject to the boundary conditions $X(0,\tau) = 0$ and $X_\tau(1,\tau) = U(1,\tau) = W(\tau)$ (or $X(1,\tau) = L(\tau)$), (3.16) generates a nonuniform grid $0 = x_0 < x_1 < \cdots < x_N = L(t)$.

We note that X must be a monotone function of ξ at any given time τ in order to be a coordinate transformation. This is guaranteed since (3.16) is a heat equation. Other equations and monitor functions can be used as well. For a detailed discussion, see [6] and references therein.

3.3. Finite-difference scheme. We solve the system (3.9)–(3.10) by a finite-difference method. We discretize (3.9) using a backward Euler scheme in time and an upwind difference in space, namely,

$$\begin{aligned}\Gamma_{i,n+1} - \Gamma_{i,n} = {} & -\delta\tau\frac{\mathcal{P}}{\nu_{i,n+1}}\left(1 - \frac{A_{n+1}}{g}\right) \\ & + \begin{cases} G_{i+1/2,n+1}(U_{i,n+1} - V_{i,n+1})(\Gamma_{i+1,n+1} - \Gamma_{i,n+1})\frac{\delta\tau}{\delta\xi} \\ \qquad \text{if} \quad U_{i,n+1} - V_{i,n+1} \le 0, \\ G_{i-1/2,n+1}(U_{i,n+1} - V_{i,n+1})(\Gamma_{i,n+1} - \Gamma_{i-1,n+1})\frac{\delta\tau}{\delta\xi} \\ \qquad \text{if} \quad U_{i,n+1} - V_{i,n+1} > 0 \end{cases}\end{aligned} \tag{3.17}$$

where $\delta\xi$ and $\delta\tau$ are the mesh and time step sizes, respectively, and i and n are the indices for the space coordinate and time level, respectively. For any grid function $f_{i,n}$, we define $f_{i+1/2,n} = (f_{i+1,n} + f_{i,n})/2$, $f_{i-1/2,n} = (f_{i-1,n} + f_{i,n})/2$. The reciprocal grid stretching ratio $G_{i,n}$ is computed using the standard central difference formula.

We now discretize (3.10) using a backward Euler scheme in time and central-differences in space; thus

$$\begin{aligned}\rho c_p\Theta_{i,n+1} = \rho c_p\Theta_{i,n} &- \frac{\rho c_p G_{i,n+1}\delta\tau}{2L_{n+1}\delta\xi}\left(U_{i,n+1}-V_{i,n+1}\right)\left(\Theta_{i+1,n+1}-\Theta_{i-1,n+1}\right)\\ &+ \frac{G_{i,n+1}\delta\tau}{\Gamma_{i,n+1}(L_{n+1})^2\delta\xi^2}\left[\kappa_{i+1/2,n+1}\Gamma_{i+1/2,n+1}\left(\Theta_{i+1,n+1}-\Theta_{i,n+1}\right)\right.\\ &\qquad\left.- \kappa_{i-1/2,n+1}\Gamma_{i-1/2,n+1}\left(\Theta_{i,n+1}-\Theta_{i-1,n+1}\right)\right]\\ &+ 2k\delta\tau\sqrt{\frac{\pi}{S_0\Gamma_{i,n+1}(1-\beta^2)}}\left[F_{hg}\frac{\varepsilon_h\alpha}{1-(1-\alpha)(1-\varepsilon_h)}\left(\Theta_h^4-\Theta_{i,n+1}^4\right)\right.\\ &\qquad\left.+ F_{bg}\frac{\varepsilon_b\alpha}{1-(1-\alpha)(1-\varepsilon_b)}\left(\Theta_b^4-\Theta_{i,n+1}^4\right)\right].\end{aligned} \tag{3.18}$$

The equations describing the tip motion, (3.12) and (3.13), are solved in a different fashion. First, (3.12) is integrated from time level n (τ) to $n+1$ ($\tau+\delta\tau$), assuming that $\mathcal{I}$ is a constant in that time interval, which yields

$$W_{n+1} = g\mathcal{I}_{n+1}\left[1-\exp\left(\frac{-\delta\tau}{\mathcal{I}_{n+1}}\right)\right] + W_n\exp\left(\frac{-\delta\tau}{\mathcal{I}_{n+1}}\right). \tag{3.19}$$

The acceleration then is obtained using (3.12) as

$$A_{n+1} = g - \frac{W_{n+1}}{\mathcal{I}_{n+1}}. \tag{3.20}$$

The tip length is obtained using a backward Euler method,

$$L_{n+1} = L_n + W_{n+1}\delta\tau. \tag{3.21}$$

The trapezoidal rule is used to evaluate the integral

$$\mathcal{I}_{n+1} = \delta\xi\sum_{i=1}^{N}\frac{\mathcal{P}}{2g}\left(\frac{1}{G_{i,n+1}\Gamma_{i,n+1}\nu_{i,n+1}}+\frac{1}{G_{i-1,n+1}\Gamma_{i-1,n+1}\nu_{i-1,n+1}}\right). \tag{3.22}$$

Finally, the velocity at each interior point is computed by discretizing (3.8) using the trapezoidal rule

$$\begin{aligned}U_{i,n+1} &= U_{i-1,n+1}\\ &+ \delta\xi\frac{\mathcal{P}}{2}\left(1-\frac{A_{n+1}}{g}\right)\left(\frac{1}{G_{i,n+1}\Gamma_{i,n+1}\nu_{i,n+1}}+\frac{1}{G_{i-1,n+1}\Gamma_{i-1,n+1}\nu_{i-1,n+1}}\right).\end{aligned} \tag{3.23}$$

The mesh adaptation also is done numerically. The finite-difference approximation of (3.16) is

$$\begin{aligned}X_{i,n+1} = X_{i,n} &+ \frac{\delta\tau}{\tau_r(\delta\xi)^2}\left[M_{i+1/2,n+1}(X_{i+1,n+1}-X_{i,n+1})\right.\\ &\qquad\left.- M_{i-1/2,n+1}(X_{i,n+1}-X_{i-1,n+1})\right].\end{aligned} \tag{3.24}$$

The boundary conditions for (3.24) are

$$X_{0,n+1} = 0, \qquad X_{N,n+1} = L_{n+1}.$$

3.4. Solution algorithm. For the initial conditions, we assume that the glass tube has uniform properties and constant temperature distribution with no motion. Since the equation for temperature is second-order in space, we need two boundary conditions, namely, the Dirichlet conditions that the temperature is held fixed at both ends. The equation for Γ is first-order in space, which means that normally one condition needs to be specified. However, $U-V=0$ at both ends, so the characteristic curve at each boundary point, $\xi=0$ and 1, is tangent to the boundary. Therefore, no explicit boundary conditions are needed for Γ. Furthermore, $U_{0,n+1}=0$, which is sufficient to solve for $U_{i,n+1}$.

We assume that the values of the variables at time level n are known, and we use an iterative procedure for solving the discrete equations for these variables at time level $n+1$ as follows:

1. the temperature Θ is computed using (3.18);
2. the value of $\mathcal{I}$ is calculated using (3.22);
3. then, W, A, and L are obtained from (3.19), (3.20), and (3.21), respectively;
4. the values of Γ are computed from (3.17);
5. the velocity U is determined from (3.23); and
6. after the physical equations are solved, we update the mesh by solving the mesh equation (3.24), and then move to the next time level.

Using these values of the variables, we update the coefficients, e.g., ν, etc., in the discrete equations, and repeat steps 1–6 until convergence is achieved. These computations are repeated until either the maximum extension length is reached or the breaking stress in the glass tube is exceeded.

We note that our numerical method is implicit and an iterative procedure is required due to the nonlinear nature of the physical and grid generation equations. In principle, a simpler explicit method also can be used to solve the set of equations listed above. However, such methods normally impose a severe constraint on the size of the time step even for linear problems. On the other hand, we can choose a relatively large step size by using the implicit method. The iterative procedure at each time step usually converges with only a few iterations.

Most of the results presented in the following section are obtained using 512 grid points, in the spatial discretization. We have experimented with more grid points, but the results are essentially the same. The size of the time step, which is allowed to vary in our computations, is chosen according to the velocity of the glass so that the free end of the glass moves less than 10^{-3} of the initial tube length ℓ_0.

4. Results. To illustrate the theory developed in this paper, we have carried out several representative computations of glass microelectrode formation. For patch-clamp experiments, appropriate strength in the shank of the electrode requires two separate pulls with different heater temperatures. Therefore, the first pull with heater temperature θ_h^1 is stopped when the glass tube is extended to a certain length and forms a "neck." The heater is switched off, and switched on again, usually at a lower temperature θ_h^2, after it has been moved to a location approximately at the smallest part of the neck and the glass tube has cooled down. The glass tube usually breaks during the second pull with the desired tip shape if the temperature of the heater is set properly. A maximum extension length, which is determined by the dimension of the puller, is set for the second pull. The adjustable parameters are the temperature, the location of the heater, and the force load on the end of the tube. The computations were carried out for two different values of θ_h^1 for the first pull, combined with various heater temperatures θ_h^2 for the second pull.

TABLE 4.1
List of the physical parameters used in the computations.

ρ g/cm^3	c_p Erg/Kg	κ cm^2/sec	k Erg/cm^{2}sec K^4	ε_h cm^2/g	ε_h cm^2/g	α cm^2/g
2.23	7.538×10^6	1.130×10^5	5.67×10^{-5}	1	1	0.4

Since the ratio of the inner and outer radii of the tube, $\beta = r/R$, is treated as an unknown function in our model, we considered two cases. The first set of computations were carried out assuming that β is a constant, which is valid when the viscosity is relatively small. For the second set of computations, we assumed that β is a linear function of the nondimensional area Γ,

$$\beta = \frac{1}{2} + \beta_0(\Gamma - 1) \tag{4.1}$$

where β_0 is a constant. We note that neither of these assumptions on β may be realistic. However, the computations based on these assumptions will provide useful information on the limitations and effects of β on the shape of the electrode when other parameters remain unchanged.

We carried out a number of numerical computations for the double-pull paradigm. The objective in the double-pull cases was to obtain tip diameters of approximately 1 micron. The important parameter was the heater temperature, which can be controlled in experiments and was varied in the computations. We first describe the properties of the glass tubes used in the numerical computations and then present the results of these computations for the glass shapes, temperature distributions, stress distributions, and their time evolutions.

4.1. Glass properties and geometrical parameters. We first describe the relevant physical and geometrical parameters. The nonlinear dependence on temperature of the coefficients in the governing equations is determined by experimental measurements. In order to follow the physical process as closely as possible, we estimate the parameter values from these data.

The most important parameter is the viscosity of the glass, μ. According to the measurements, the relationship between $\ln \mu$ and the temperature is piecewise linear. Therefore, we use power laws for the temperature dependence over certain intervals. A typical formula for the viscosity of the glass (in g/cm sec) is given by

$$\mu(\theta) = \begin{cases} 10^{9-c_1(\theta-293)}, & \theta \le 900, \\ 10^{3.612-c_2(\theta-900)}, & 900 \le \theta \le 1100, \\ 10^{3.38-c_3(\theta-1100)}, & 1100 \le \theta \le 1500. \end{cases} \tag{4.2}$$

Here $c_1 = 8.876 \times 10^{-3}$, $c_2 = 1.16 \times 10^{-3}$, and $c_3 = 7.355 \times 10^{-3}$, based on measurements for soda-lime [8]. The other physical parameters used in our computations also come from the experimental setup and are summarized in Table 4.1. The initial temperature of the glass tube is set to be the background temperature, or room temperature (assumed to be 20°C), i.e., $\theta_0 = \theta_b = 293$°K.

The geometrical parameters for the glass tube and heater are given in Table 4.2. The other two relevant parameters are ℓ_{p1} and ℓ_{p2}, which are the maximum lengths set for the first and second pulls, respectively. Due to the physical constraint of the puller, $\ell_{p1} + \ell_{p2}$ is usually fixed, while various combinations are allowed. In this

TABLE 4.2
List of the geometrical parameters used in the computations.

Glass Properties			Heater Properties		
ℓ_0 cm	R_0 cm	r_0 cm	x_h cm	ℓ_h cm	R_h cm
7.56	8.66×10^{-2}	4.33×10^{-2}	3.63	0.3	0.15

TABLE 4.3
NB stands for the cases in which the breaking stress criterion is not met; the superscript e denotes the values at the end of the second pull, and the superscript b denotes the values at which the maximum stress $\mathcal{S}_{max}$ first exceeds the breaking stress $\mathcal{S}_b$; R_{min} and r_{min} are the outer and inner radii, respectively, of the glass at the neck.

θ_h^1 (^{o}K)	1100			1500		
θ_h^2 (^{o}K)	900	1100	1500	900	1100	1500
$\mathcal{S}_b^e$ ($\times 10^9$ dyn/cm^3)	2.41	2.26	2.10	2.45	2.33	2.17
$\mathcal{S}_{max}^e$ ($\times 10^9$ dyn/cm^3)	4.89	1.88	.350	7.49	3.48	.726
R_{min}^e (μm)	12.4	6.67	5.35	5.98	3.75	2.71
r_{min}^e (μm)	6.19	3.33	2.68	2.99	1.88	1.35
$\mathcal{S}_b$ ($\times 10^9$ dyn/cm^3)	2.40	NB	NB	2.44	2.30	NB
$\mathcal{S}_{max}^b$ ($\times 10^9$ dyn/cm^3)	2.41	NB	NB	2.49	2.31	NB
R_{min}^b (μm)	51.2	NB	NB	55.3	43.0	NB
r_{min}^b (μm)	25.6	NB	NB	27.6	21.5	NB

study, however, we have chosen $\ell_{p1} = 0.58$ cm and $\ell_{p2} = 4$ cm. The initial area is $s_0 = \pi(R_0^2 - r_0^2) = \pi R_0^2(1 - \beta_0^2) = 1.767 \times 10^{-2}$.

During the pulling processes, the glass tube usually breaks in the location where the stress exceeds the "breaking stress." The breaking stress is a material-dependent parameter that also depends on the temperature. For example, for the glass used in this study, the breaking stress (in dyn/cm^3) is given by the empirical formula (see Scholze [10, pp. 255–272])

$$\mathcal{S}_b = \frac{5.12 \times 10^{10}}{\sqrt{\theta}}, \tag{4.3}$$

which indicates that it becomes easier to break this type of glass as the temperature increases. For our first set of computations with a constant β, we choose not to impose the breaking stress. Instead, we use the maximum length $\ell_{p1} + \ell_{p2}$ as a stopping criterion. The breaking stress using (4.3) is computed as a reference value. The breaking stress criterion is imposed for the second set of computations when β is a function of Γ (4.1).

4.2. Numerical results for constant β. In Table 4.3, we summarize the computational results based on the constant area ratio $\beta = 1/2$.

The table lists the minimum radii (the values at the "neck" of the tube) and the maximum stress at the ends of the first and the second pulls, and when the stress in the glass tube exceeds the breaking stress computed using formula (4.3). The heater temperatures for the two pulls are chosen to be a combination of 900°K, 1100°K, and 1500°K, since these are the critical values for the glass viscosity shown in (4.2).

TABLE 4.4

NB stands for the cases in which the breaking stress criterion is not met; the superscript b denotes the values at which the maximum stress $\mathcal{S}^b_{max}$ first exceeds the breaking stress $\mathcal{S}_b$; R^b_{min} and r^b_{min} are the outer and inner radii, respectively, at the neck; t is the time from the start of the second pull; and x_{min} is the location of the neck from the clamped end of the glass.

θ^1_h (°K)	1100				1500			
θ^2_h (°K)	900	1042	1044	1046	1100	1190	1192	1194
$\mathcal{S}_b$ ($\times 10^9$ dyn/cm^3)	2.40	2.28	2.28	NB	2.30	2.25	2.25	NB
$\mathcal{S}^b_{max}$ ($\times 10^9$ dyn/cm^3)	2.41	2.28	2.28	NB	2.31	2.25	2.25	NB
R^b_{min} (μm)	51.2	19.6	17.7	NB	43.0	18.1	15.6	NB
r^b_{min} (μm)	25.6	9.81	8.87	NB	21.5	9.01	7.78	NB
t^b (sec)	12.8	9.97	9.95	NB	3.05	2.80	2.80	NB
x^b_{min} (cm)	4.59	4.92	4.99	NB	4.31	4.58	4.65	NB

The first observation one can make is that critical heater temperatures exist, i.e., temperatures beyond which the stress in the glass tube never reaches the breaking stress. It also can be seen that radii at the end of the second pull do not vary significantly. However, the value of maximum stress that exceeds the breaking stress, for some of the cases listed, is affected by the heater temperature. This suggests that if the breaking stress criterion is imposed, then the glass tube may break and the minimum radius (at the neck) of the tube may vary. This is confirmed by the values of the radii at which the stress first exceeds the breaking stress. In all the cases, the radii at the end of the second pull are comparable to the values for the glass electrodes pulled in the laboratory. However, the radii when the stress exceeds the breaking stress on the first pull are much larger than those at the end of the second pull.

To further investigate the dependence of the minimum radii on the heater temperature, we have done simulations with careful variations of the temperature. The results are given in Table 4.4.

With small variations of the heater temperature, we numerically identified the critical temperatures for the second pull. For $\theta^1_h = 1100°$K, this critical temperature is about $\theta^2_h = 1044°$K, whereas for $\theta^1_h = 1500°$K, this critical temperature is about $\theta^2_h = 1192°$K. In general, the neck radii are sensitive to the heater temperature. However, when the glass breaks (breaking stress is reached) at the critical heater temperatures during the second pull, the radii of the neck appear to be less sensitive to the heater temperature of the first pull.

The time history of the stress and breaking stress offers a clearer picture of what happens near the critical temperature, as plotted in Figure 4.1 for $\theta^1_h = 1500°$K. It can be seen that, during the second pull, the breaking stress decreases initially, reaches a minimum value, and then starts to increase again. This apparently corresponds to the fact that the temperature at the neck rises first, reaches a maximum, and then decays when the neck moves away from the center of the heater. In the cases shown in Figures 4.1(a)–(c), at a relatively low heater temperature ($\theta^2_h = 1100°$K), the stress at the neck of the glass tube increases monotonically with time and passes the breaking stress during the second pull. At a higher heater temperature ($\theta^2_h = 1500°$K), the maximum stress increases initially, reaches a maximum value, and then

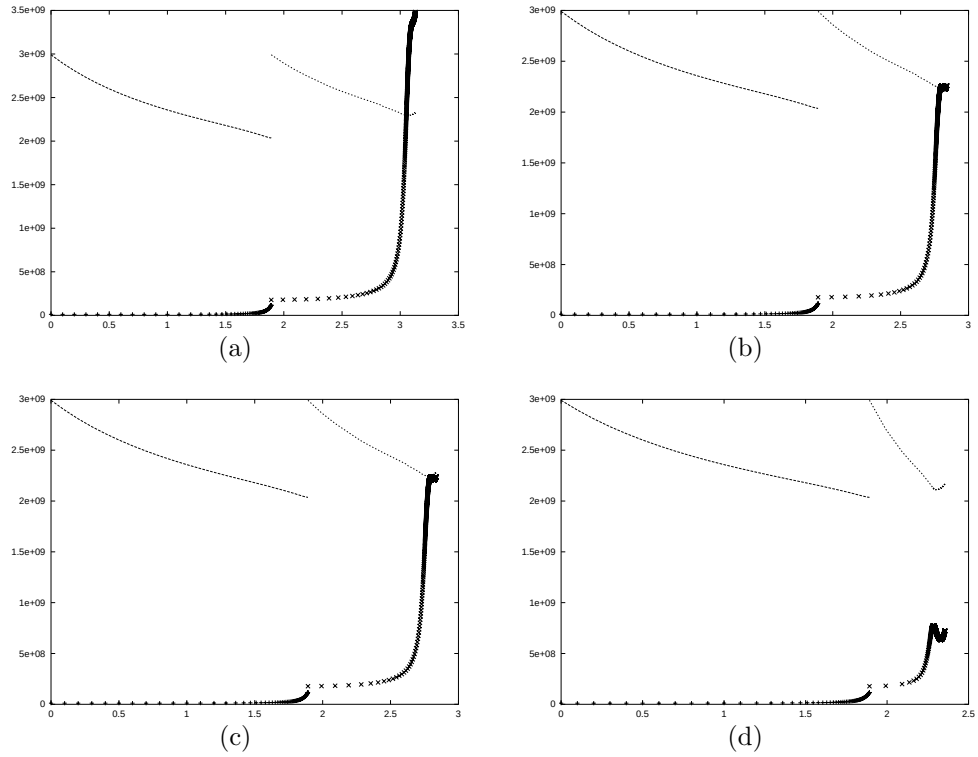


FIG. 4.1. *Stress and breaking stress (in dyn/cm^3) versus time with $\theta_h^1 = 1500°K$ and* (a) $\theta_h^2 = 1100°K$, (b) $\theta_h^2 = 1192°K$, (c) $\theta_h^2 = 1194°K$, (d) $\theta_h^2 = 1500°K$. *The dashed line denotes the breaking stress, and the symbols ($\times$ and $+$) are the maximum stresses.*

decreases. It never reaches the breaking stress (Figure 4.1(d)). Note that the peak of the maximum stress, just before breaking, decreases with increasing temperature during the second pull. Deceasing the heater temperature raises the maximum value of the stress (Figure 4.1(a)). As a result, the stress reaches and passes the breaking stress sooner for lower heater temperature during the second pull. Therefore, the neck radii are greater at lower heater temperature during the second pull when the breaking stress is met since there is insufficient time for the tube to stretch and reduce the neck radii. This is demonstrated in Figure 4.2, where the outer radius of the neck is plotted against time.

4.3. Numerical results for variable β. In order to gain some insight on the effect of β, the ratio of the inner and outer radii of the glass tube, we present some computations with $\theta_h^1 = 1500°$K and $\theta_h^2 = 1190°$K, with β given by (4.1) where β_0 varies from 0.3 to -0.2. The breaking stress criterion is imposed, and the glass breaks for every value of β_0. In Table 4.5, we have listed the outer and inner radii and the dimensionless area at the tip (neck) as well as the minimum value of β when the glass tube breaks. Note that γ is initially 1, so β is initially 0.5 for the first pull. However, the value of γ is negligible in determining β_{min} when the glass tube breaks; cf. (4.1) and Table 4.5. Clearly, the value of β_0 (therefore the value of β) has a visible effect

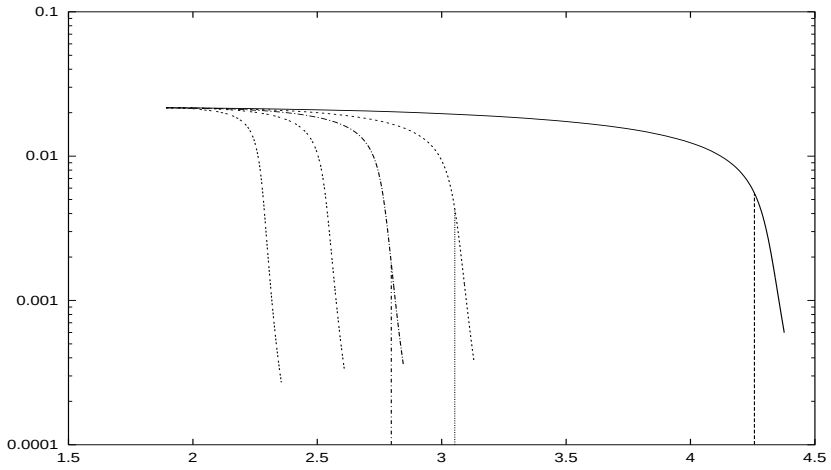


FIG. 4.2. *Neck outer radius (in cm) versus time (in sec) with the same heater temperature* $\Theta_H^1 = 1500^\circ K$ *for the first pull and* $\theta_h^2 = 900$, 1100, 1190, 1300, *and* $1500^\circ K$ *for the second pull, ordered from right to left. The times when the maximum stress exceeds the breaking stress and the corresponding values of the radii are marked. These times depend on the heater temperature during the second pull.*

TABLE 4.5

Tip geometry as a function of β. *The temperatures of the heater are* $1500^\circ K$ *for the first pull and* $1190^\circ K$ *for the second pull. Radii are given in* μm. *The minimum value of the reciprocal grid stretching ratio is given in the second column.*

β_0	G_{min}	R_{min}	r_{min}	β_{min}
0.3	1.20×10^{-3}	26.5	5.32	0.20
0.2	1.10×10^{-3}	26.1	7.84	0.30
0.1	9.33×10^{-4}	25.0	10.0	0.40
0.0	4.34×10^{-4}	18.1	9.03	0.50
–0.1	1.51×10^{-5}	3.64	2.19	0.60
–0.2	1.31×10^{-5}	3.80	2.66	0.70

on the values of the tip radii and area. It can be seen that the tip area decreases monotonically as β_0 decreases. However, the actual values of the tip outer and inner radii vary in a more complicated fashion. We note that the effect of β_0 (and β) is more effective in reducing R_{min} and r_{min} when $\beta_0 \leq 0$ ($\beta_{min} \geq 0.5$). There seems to exist a β_0 value between 0 and -0.2 (correspondingly $\beta_{\min}$ is between 0.5 and 0.7) for which both radii at the tip are reduced by about a factor of 3–4 (close to their respective local minima). This corresponds to the case in which the inner radius contracts more slowly than the outer radius, resulting in a smaller tip area. On the other hand, if the inner radius contracts more quickly than the outer one ($\beta < 0.5$), the effects of different β_0 (β) on the inner radius are clearly demonstrated, while the effects on the tip area and its outer radius are limited as β_0 increases (β_{min} decreases).

Finally in Figure 4.3, we plot the corresponding shape of the glass tube for three of the cases computed when the breaking stress is reached. It can be seen that smaller values of the tip radii, when $\beta_0 = -0.1$, are due mainly to the time available for the glass to stretch.

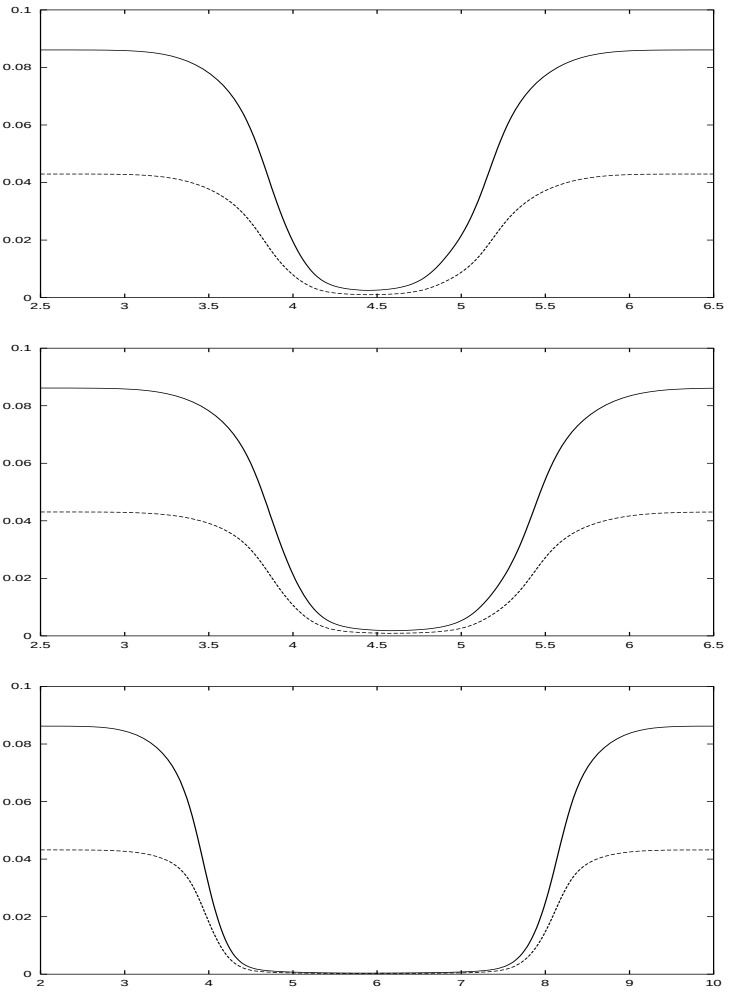


FIG. 4.3. *Shape (outer (solid) and inner (dashed) radii) of a glass tube when it breaks with* $\theta_h^1 = 1500°K$ *and* $\theta_h^2 = 1100°K$*: $\beta_0 = 0.1$ (top); $\beta_0 = 0$ (middle); $\beta_0 = -0.1$ (bottom).*

5. Concluding remarks. In this study, we present a mathematical model for the formation of glass microelectrodes using a double pull paradigm. The model is one-dimensional and is based on the assumption that the glass material is an incompressible viscous fluid. Numerical results indicate that the heater temperature plays a critical role in the formation process. In order to obtain glass electrodes with desirable tip shape and dimension, one needs a suitable combination of the heater temperature for the two pulls. Another important factor revealed by our model and the numerical simulations is the ratio of the inner and outer radii, β. Since our model does not take the free boundary into account, this ratio remains a free parameter. The effect of β is investigated by using a simple linear relationship between β and the dimensionless area γ, which can be viewed as the first-order approximation when γ is close to its initial value. It is shown that changing the value of β has little effect on the results when the inner radius contracts faster than the outer one. On the other hand, both

the inner and outer radii at the tip could be reduced by a factor of four or more if the inner radius contracts more slowly than the outer one. We also have investigated the effects of surface tension. It was found that under the physical conditions, surface tension is at least one order of magnitude smaller than the viscous stress. Therefore, excluding the effect of surface tension in the current model is justified.

Although our focus here has been on the specific manufacturing paradigm for glass microelectrodes in the laboratory, our mathematical model does allow us to investigate a variety of other possible changes in the parameters to obtain better shaped microelectrodes, e.g., the effects of changes in the weight, the ambient temperature, and different pulling strategies.

However, a more conclusive result only can be obtained from a more realistic model which takes the free boundary into account. Work is currently underway to incorporate the effects of surface tension and pressure changes due to temperature variations into the current model.

Appendix A. Derivation of the model. To derive both the continuity equation (2.1) and the energy equation (2.6) for the glass tube, we apply a control-volume approach. Figure A.1 is a schematic diagram of a typical infinitesimal section of the glass tube with length Δx. The control volume consists of four surfaces: two annuli with areas $s(x,t)$ on the left side and $s(x+\Delta x,t)$ on the right side, and outer and inner surfaces with areas given approximately by $s_{\text{out}} = 2\pi R(x,t)\Delta x$ and $s_{\text{in}} = 2\pi r(x,t)\Delta x$ where R and r are the radii of the outer and inner circles on the left side, respectively. We assume that the effect of the local slopes with respect to x of the outer and inner radii are negligible.

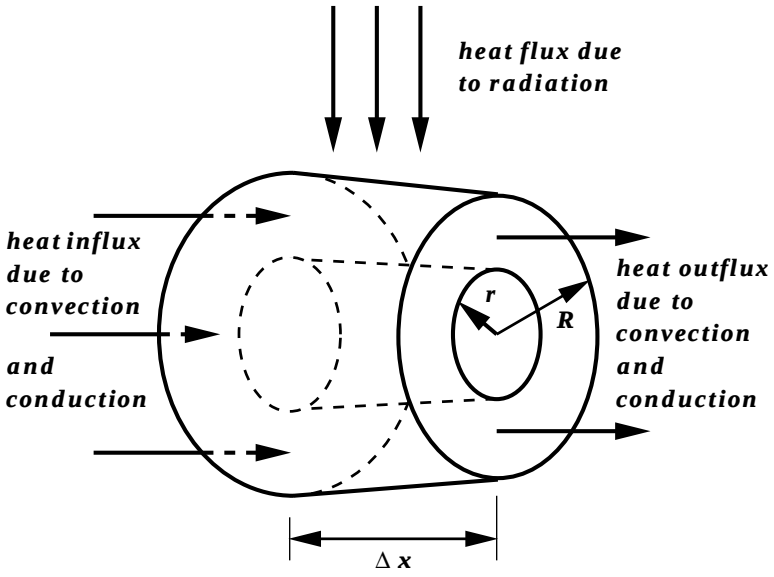


FIG. A.1. *Schematic of the control volume for the conservation of energy in the glass tube.*

The conservation of mass, i.e., the net change of the glass in the control volume due to the net flux of glass from the left and right sides, is given by the continuity equation

$$\frac{\partial(\rho s)}{\partial t} + \frac{\partial(\rho s u)}{\partial x} = 0 \tag{A.1}$$

where ρ is the glass density and u is the local velocity. With the incompressibility assumption, i.e., the material derivative of the density is $D\rho/Dt = \partial\rho/\partial t + u\partial\rho/\partial x = 0$, the continuity equation can be simplified to

$$\frac{\partial s}{\partial t} + \frac{\partial (su)}{\partial x} = 0. \tag{A.2}$$

Define $\gamma \equiv s/s_0$ where s_0 is the initial constant annular cross-sectional area of the tube. Then (A.2) becomes

$$\frac{\partial \gamma}{\partial t} + u\frac{\partial \gamma}{\partial x} + \gamma\frac{\partial u}{\partial x} = 0. \tag{A.3}$$

In order to write down the conservation law for the energy, we make the following observations:

1. On the left surface, the energy influx is due to convective flow of the glass and conduction of heat through the glass within a period Δt, which is given by

$$E^{\text{in}} = s\left(\rho c_p u\theta - \kappa\frac{\partial\theta}{\partial x}\right)\Delta t \tag{A.4}$$

where c_p, κ, and θ are the specific heat, thermal conductivity, and temperature of the glass, respectively. The first term on the right-hand side is the convection due to motion of the glass, and the second term is the heat conduction using Fourier's law.

2. On the right surface, the energy outflux can be written similarly as

$$E^{\text{out}} = -s\left(\rho c_p u\theta - \kappa\frac{\partial\theta}{\partial x}\right)\Delta t - \frac{\partial}{\partial x}\left[s\left(\rho c_p u\theta - \kappa\frac{\partial\theta}{\partial x}\right)\right]\Delta x\Delta t. \tag{A.5}$$

The minus signs reflect the fact that energy is being removed.

3. On the outer surface, the energy exchange mechanism is mainly radiation between the heater coil, glass pipette, and background. For a differential cross section of the tube, the radiative heat transfer absorbed at the surface is given by

$$E^{\text{rad}} = s_{\text{out}}\left[F_{hg}\frac{\varepsilon_h\alpha}{\varepsilon_h + \alpha - \varepsilon_h\alpha}(\theta_h^4 - \theta^4) + F_{bg}\frac{\varepsilon_b\alpha}{\varepsilon_b + \alpha - \varepsilon_b\alpha}(\theta_b^4 - \theta^4)\right], \tag{A.6}$$

where we have assumed the standard fourth-power law of surface-to-surface radiative heat transfer. The factors F_{hg} and F_{bg} are geometrical factors related to the relevant surfaces. In this case, they are the heater surface, the glass outer surface, and the background. Therefore, the subscripts hg and bg refer to the relationship between the heater and glass surfaces, and the background and glass surfaces, respectively. The detailed derivations of these factors are given in Appendix B. The parameters $\varepsilon_h, \varepsilon_b$, and α are the emissivities for the heater and the background, and the absorptivity of the glass, respectively. The first term is the amount of the net energy flux absorbed by the outer surface of the glass tube, due to the radiation from the heater, which is at the temperature $\theta_h \geq \theta$. The second term is the net energy radiated away from the glass tube towards the background, which is at the temperature $\theta_b \leq \theta$.

4. There is no net energy exchange through the inner surface of the glass.

5. The incremental change in the internal energy within the control volume of the glass tube is determined by the change of temperature during the time period Δt and is given by

$$E^{\text{int}} = [\rho c_p s(x, t+\Delta t)\theta(x, t+\Delta t) - \rho c_p s(x,t)\theta(x,t)]\Delta x. \tag{A.7}$$

With these assumptions, we can write down the equation for the temperature using the conservation of energy; i.e., the net increase in the internal energy is equal to the net flux of energy into the control volume, and thus (A.7) = (A.4) + (A.5) + (A.6). After rearranging the terms, we obtain

$$\begin{aligned}\frac{\partial \rho c_p s\theta}{\partial t} + \frac{\partial \rho c_p su\theta}{\partial x} &= \frac{1}{s}\frac{\partial}{\partial x}\left(s\kappa(\theta)\frac{\partial \theta}{\partial x}\right) \\ &\quad + 2\sqrt{\frac{\pi}{s(1-\beta^2)}}k\left[F_{hg}\frac{\varepsilon_h \alpha}{\varepsilon_h + \alpha - \varepsilon_h\alpha}(\theta_h^4 - \theta^4)\right. \\ &\quad \left. + F_{bg}\frac{\varepsilon_b \alpha}{\varepsilon_b + \alpha - \varepsilon_b\alpha}(\theta_b^4 - \theta^4)\right]. \end{aligned} \tag{A.8}$$

We have assumed that the ratio of the inner and outer radii, $\beta = r/R$, is a constant independent of t and x. Note that this equation is different from (2.6) used in our computations. Equation (A.8) is written in conservation form, while (2.6) is in nonconservative form. In (A.8), we can rewrite the two terms on the left-hand side as

$$\begin{aligned}\frac{\partial \rho c_p s\theta}{\partial t} + \frac{\partial \rho c_p su\theta}{\partial x} &= \rho s\left(\frac{\partial c_p\theta}{\partial t} + u\frac{\partial c_p\theta}{\partial x}\right) + c_p\theta\left(\frac{\partial(\rho s)}{\partial t} + \frac{\partial \rho su}{\partial x}\right) \\ &= \rho s\left(\frac{\partial c_p\theta}{\partial t} + u\frac{\partial c_p\theta}{\partial x}\right)\end{aligned}$$

where we have used the continuity equation (A.1). Therefore, the conservation form (A.8) and nonconservative form (2.6) of the energy equation are equivalent.

Appendix B. Derivation of geometric factors. The derivation of the equation for the transfer of thermal energy from the coil heater to the glass tube through radiation requires a detailed knowledge of the differential radiation from a differential area element on the inside surface of the heater to a differential area element on the outer surface of the glass tube. It is assumed that there is no radiation passing through the tube itself. The glass tube surface is assumed to have an absorptivity equal to α. Also, we take into account the variations of the inner and outer radii of the tube. However, we assume that the effects due to the local slope of the outer radius with respect to x are negligible. Also, we approximate the coil by a cylindrical surface with emissivity ε_h.

To account for the relative surface orientations of the heater and tube surfaces, we compute the geometric (or radiation configuration) factor needed to evaluate the thermal energy transfer to the tube (Howell [5]). The geometry of the heater and tube system and the notation are shown in Figure B.1. We evaluate the total radiative heat transfer from the heater to a differential cross section of the tube centered at x. The differential of the geometric factor for a differential area ds at x on the tube for a differential area ds_h on the heater is given by

$$dF_{hg} = \frac{\cos\phi \cos\phi_h ds ds_h}{\pi\sigma^2}$$

where ϕ and ϕ_h are the angles of the line (of length σ) connecting the differential area elements ds and ds_h with the outer normal of the glass tube and inner normal of the heater, respectively. The differential area element on the heater is

$$ds_h = R_h dy d\Phi$$

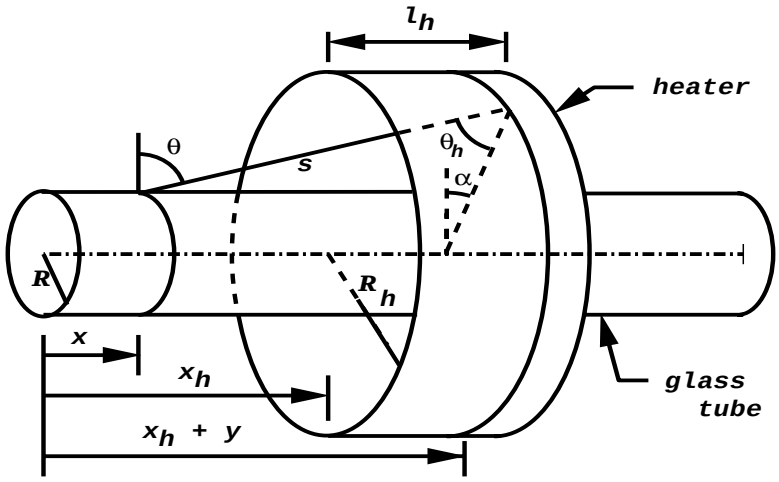


FIG. B.1. *Geometry for radiative heat transfer from the heater to the glass tube.*

where R_h is the heater radius, y is the axial distance between the differential area elements, and Φ is the angle between the two radial segments shown in Figure B.1. Thus the geometric factor from the heater to the area element ds is given by

$$\begin{aligned} F_{hg} &= \frac{1}{\pi}\iint \frac{\cos\phi\cos\phi_h}{\sigma^2}\,ds_h \\ &= \frac{1}{\pi}\int_{-\Phi_0}^{\Phi_0}\int_0^{\ell_h} \frac{\cos\phi\cos\phi_h}{\sigma^2} R_h\,dy\,d\Phi \end{aligned} \tag{B.1}$$

where

$$\begin{aligned} \cos\phi &= -\frac{R^2+\sigma^2-(\ell_h^2+R_h^2)}{2R\sigma}, \\ \cos\phi_h &= \frac{R_h^2+\sigma^2-(\ell_h^2+R_h^2)}{2R_h\sigma}, \\ \sigma^2 &= (y+x_h-x)^2 + R^2 + R_h^2 - 2RR_h\cos\Phi, \\ \Phi_0 &= \cos^{-1}\left(\frac{R}{R_h}\right). \end{aligned} \tag{B.2}$$

The double integral for the geometric factor (B.1) can be simplified to a single integral

$$\begin{aligned} F_{hg} = -\frac{1}{\pi}\int_0^{\Phi_0} & R_h(R_h - R\cos\Phi)(R - R_h\cos\Phi) \\ & \cdot\left[\frac{y+x_h-x}{\mathcal{R}^2[(y+x_h-x)^2+\mathcal{R}^2]} + \frac{\tan^{-1}\frac{y+x_h-x}{\mathcal{R}}}{\mathcal{R}^3}\right]_{y=0}^{y=\ell_h} d\Phi \end{aligned} \tag{B.3}$$

where

$$\mathcal{R}^2 = R^2 + R_h^2 - 2RR_h\cos\Phi.$$

The geometric factor between the glass tube and the background is equivalent to the geometric factor between the glass tube and an infinite cylindrical shell where the segment shielded by the heater is excluded. The formula for this geometric factor can be obtained more simply by replacing the cylinder with a spherical shell, again excluding the portion of the shell shielded by the heater. This formula is given by

$$F_{bg} = 1 - \frac{1}{2\pi}[2(\beta_2 - \beta_1) + \sin 2\beta_2 - \sin 2\beta_1] \tag{B.4}$$

where $\beta_1 = \arctan \frac{R_h}{x_h}$ and $\beta_2 = \arctan \frac{R_h}{x_h + \ell_h}$.

Acknowledgments. We wish to thank Dr. Hilton Ramsey for several useful discussions in the early part of this research. Also, we thank Ken Burkett and Bruce Walding of Corning Glass for supplying us with the thermal and mechanical properties of Corning glass.

REFERENCES

[1] A.D. Fitt, K. Furusawa, T.M. Monro, and C.P. Please, *Modeling the fabrication of hollow fibers: Capillary drawing,* J. Lightwave Technol., 19 (2001), pp. 1924–1931.

[2] D.G. Flaming and K.T. Brown, *Micropipette puller design, form of the heating filament and effects of filament width on tip length and diameter,* J. Neurosci. Methods, 6 (1982), pp. 91–102.

[3] D.M. Fry, *A scanning electron microscope method for the examination of glass microelectrode tips either before or after use,* Experientia, 31 (1975), pp. 695–697.

[4] P. Gospodinov and A.L. Yarin, *Draw resonance of optical microcapillaries in nonisothermal drawing,* Int. J. Multiphase Flow, 23 (1997), pp. 967–976.

[5] J.R. Howell, *A Catalog of Radiation Configuration Factors,* McGraw-Hill, New York, 1982.

[6] W. Huang, Y. Ren, and R.D. Russell, *Moving mesh partial differential equations (MMPDES) based on the equidistribution principle,* SIAM J. Numer. Anal., 31 (1994), pp. 709–730.

[7] S. Mittman, D.G. Flaming, D.R. Copenhagen, and J.H. Belgum, *Bubble pressure measurement of micropipette tip outer diameter,* J. Neurosci. Methods, 22 (1987), pp. 161–166.

[8] *Pyrex Glass Code* 7740*, Material Properties*, Brochure Pyrex B-87, Corning Glass, Corning, NY, 1987.

[9] G.R. Robinson and B.I.H. Scott, *A new method of estimating micropipette tip diameter,* Experientia, 29 (1973), pp. 1039–1040.

[10] H. Scholze, *Glass, Nature, Structure, and Properties*, translated by M.J. Lakin, Springer-Verlag, New York, 1990.

[11] E.M. Snell, *Some electrical properties of fine tipped pipette microelectrodes,* in Glass Microelectrodes, M. Lavallee, A.F. Shanne, and N.C. Hubert, eds., Wiley, New York, 1969, pp. 111–123.

[12] A.L. Yarin, P. Gospodinov, and V.I. Roussinov, *Stability loss and sensitivity in hollow fiber drawing,* Phys. Fluids, 6 (1994), pp. 1454–1463.

SIAM J. APPL. MATH.
Vol. 63, No. 5, pp. 1520–1538

PERIODIC TRAVELLING WAVE SELECTION BY DIRICHLET BOUNDARY CONDITIONS IN OSCILLATORY REACTION-DIFFUSION SYSTEMS*

JONATHAN A. SHERRATT†

Abstract. Periodic travelling waves are a fundamental solution form in oscillatory reaction-diffusion equations. Here I discuss the generation of periodic travelling waves in a reaction-diffusion system of the generic λ-ω form. I present numerical results suggesting that when this system is solved on a semi-infinite domain subject to Dirichlet boundary conditions in which the variables are fixed at zero, periodic travelling waves develop in the domain. The amplitude and speed of these waves are independent of the initial conditions, which I generate randomly in numerical simulations. Using a combination of numerical and analytical methods, I investigate the mechanism of periodic travelling wave selection. By looking for an appropriate similarity solution, I reduce the problem to an ODE system. Using this, I derive a formula for the selected speed and amplitude as a function of parameters. Finally, I discuss applications of this work to ecology.

Key words. periodic waves, wavetrains, reaction-diffusion, oscillatory systems

AMS subject classification. 35K57

DOI. 10.1137/S0036139902392483

1. Introduction.

Periodic travelling waves (PTWs) are a fundamental solution form in oscillatory reaction-diffusion equations, by which I mean reaction-diffusion systems whose kinetics have a stable limit cycle. PTWs are the one-dimensional analogue of spiral waves and target patterns, and underlie many observed behaviors in biology and chemistry (Bjørnstad, Ims, and Lambin (1999); Scott et al. (2000)). In 1973, Kopell and Howard published their seminal paper, which showed that a reaction-diffusion system develops a one-parameter family of PTWs as its kinetics pass through a Hopf bifurcation. Wave speed or amplitude are convenient parameters for this family, and an oscillatory reaction-diffusion equation has a PTW solution for any speed above a critical minimum value and for any amplitude below that of the limit cycle in the kinetics. Building on Kopell and Howard's work, periodic travelling waves were studied extensively in the 1970s and 1980s. This work focussed primarily on the existence and stability of the solutions. For instance, Maginu (1981) showed that PTWs of sufficiently high speed are stable in general systems, and Ermentrout (1981) demonstrated stable small amplitude waves in a particular reaction-diffusion system. More recent work includes nonlinear stability analysis (Kapitula (1994)), the application of symmetry methods (Romero, Gandarias, and Medina (2000)), and the generation of PTWs behind invasive fronts (Sherratt (1994a,b); Sneyd and Sherratt (1997); Ermentrout, Chen, and Chen (1997); Petrovskii and Malchow (1999), (2000); Ashwin et al. (2002)).

The simplest behavior of an oscillatory system is a spatially uniform oscillation. In many cases, this solution is stable on an infinite, spatially homogeneous domain: for instance, in reaction-diffusion systems, stability is guaranteed when the diffusion coefficients are sufficiently similar (Kopell and Howard (1973)). However, spatially

*Received by the editors May 20, 2002; accepted for publication (in revised form) January 31, 2003; published electronically June 12, 2003. This research was supported in part by an EPSRC Advanced Research Fellowship.

http://www.siam.org/journals/siap/63-5/39248.html

†Centre for Theoretical Modelling in Medicine, Department of Mathematics, Heriot-Watt University, Edinburgh EH14 4AS, UK (jas@ma.hw.ac.uk).

uniform oscillations are incompatible with spatial heterogeneities, which can arise via spatially varying parameter values or via conditions imposed on finite boundaries. Such situations provide a potential mechanism for the generation of PTWs in oscillatory chemical or biological systems. For instance, it is well known that in experiments with the oscillatory Belousov–Zhabotinskii chemical reaction, small impurities such as dust particles force target patterns or spiral waves rather than homogeneous oscillations (Nagashima (1991), Winfree (2001)). Mathematically, the effects of such heterogeneities have been studied most fully for systems of discrete coupled oscillators. In particular, the work of Ermentrout, Kopell, and colleagues gives a detailed account of the response of chains of weakly coupled oscillators to both boundary- and parameter-based heterogeneities (Ermentrout and Kopell (1984), (1986); Kopell, Ermentrout, and Williams (1991); Ren and Ermentrout (1998)). In oscillatory reaction-diffusion systems, there has been some study of periodic wave generation by spatial inhomogeneities in the domain (Hagan (1981a); Kopell (1981); Kay and Sherratt (2000)). However, heterogeneities imposed at the edges of a domain have received little attention, despite early work by Auchmuty and Nicolis (1976), who developed series solutions for the Brusselator model close to Hopf bifurcation on a finite domain with Neumann and Dirichlet end conditions.

In the present paper, I study the generation of PTWs by particular Dirichlet conditions at one edge of a semi-infinite domain. In section 2, I introduce this behavior with the results of numerical simulations. In section 3, I show that solutions of the observed form satisfy an ODE system with one free parameter, which corresponds to the temporal frequency of the oscillations. I then present a combination of analytical and numerical results suggesting that this ODE system has a solution satisfying appropriate end conditions for a countably infinite set of values of this parameter. In section 5, I discuss the hypothesis that in only one of these solutions does the amplitude vary monotonically in space, and that this determines the stability of the solutions. In section 6, I use a similarity solution to derive a formula for the speed and amplitude of the observed periodic wave. Finally, in section 7, I discuss extensions to two space dimensions and applications of the results.

2. Numerical simulations of PTW generation. All of the work in this paper involves the following oscillatory reaction-diffusion system:

$$\frac{\partial u}{\partial t} = \nabla^2 u + (1 - r^2)u - (\omega_0 - \omega_1 r^2)v, \tag{2.1a}$$

$$\frac{\partial v}{\partial t} = \nabla^2 v + (\omega_0 - \omega_1 r^2)u + (1 - r^2)v, \tag{2.1b}$$

where $r = \sqrt{u^2 + v^2}$. This belongs to the "λ-ω" class of equations introduced by Kopell and Howard (1973). The kinetics in (2.1) are the normal form of any oscillatory kinetics close to a supercritical Hopf bifurcation, and, as such, (2.1) is the natural system for studying generic behavior in systems in which each variable has the same diffusion coefficient. This system is often seen with $(1 - r^2)$ replaced by $(\lambda_0 - \lambda_1 r^2)$, but the coefficients λ_0 and λ_1 can easily be removed by rescaling. All of the work in sections 2–6 is in one space dimension; in section 7, two-dimensional behavior is discussed briefly.

The kinetics of (2.1) have an unstable equilibrium at $u = v = 0$ and a stable circular limit cycle centered at this equilibrium, of radius 1. Standard theory, due

originally to Kopell and Howard (1973), shows that the family of PTWs is given by

$$u = r^* \cos\left[\theta_0 \pm \sqrt{1-r^{*\,2}}x + \left(\omega_0 - \omega_1 r^{*\,2}\right)t\right], \tag{2.2a}$$

$$v = r^* \sin\left[\theta_0 \pm \sqrt{1-r^{*\,2}}x + \left(\omega_0 - \omega_1 r^{*\,2}\right)t\right], \tag{2.2b}$$

where r^* parameterizes the family and θ_0 is an arbitrary constant. The wave is stable as a solution of (2.1), provided that

$$r^* > r_{stab} \equiv \left(\frac{2+2\omega_1^2}{3+2\omega_1^2}\right)^{1/2} \tag{2.3}$$

(Kopell and Howard (1973)). In many situations, it is convenient to rewrite system (2.1) using r and $\theta = \tan^{-1}(v/u)$, which are polar coordinates in the u-v plane. In one space dimension, this gives

$$r_t = r_{xx} - r\theta_x^2 + r(1-r^2), \tag{2.4a}$$

$$\theta_t = \theta_{xx} + \frac{2r_x\theta_x}{r} + \omega_0 - \omega_1 r^2. \tag{2.4b}$$

Here and throughout the paper, the suffixes $_x$ and $_t$ denote derivatives. The PTW solutions (2.2) are of course given in terms of r and θ by

$$r = r^*, \qquad \theta = \theta_0 \pm \sqrt{1-r^{*\,2}}\,x + \left(\omega_0 - \omega_1 r^{*\,2}\right)t.$$

In fact, it is easy to show that any solution with r constant and < 1 is a PTW.

The starting point of my work is the following very simple situation. I consider (2.1) on a semi-infinite domain $x > 0$, say, with the boundary condition $u = v = 0$ at $x = 0$. Numerically this can be reproduced by solving on the finite domain $0 < x < X_\infty$, with X_∞ large and with $u_x = v_x = 0$ at $x = X_\infty$. I consider the solution that develops from random initial conditions, by which I mean that I use a random number generator to calculate u and v values, between ± 1, at points with an equal spacing of about $\Delta x = 5$ throughout the domain, and then join these random values by straight lines to give the initial condition.

For a wide range of values of the parameters ω_0 and ω_1, the numerical solutions of this problem show the same behavior (Figure 1). The solution changes rapidly from the random initial conditions to spatially uniform oscillations everywhere away from the $x = 0$ boundary. A transition wave then develops, which has homogeneous oscillations ahead of it and a PTW behind; this PTW is the long term solution form away from the $x = 0$ boundary. The development and persistence of PTWs in these solutions depends intrinsically on the boundary condition at $x = 0$. For example, if the boundary condition is switched to zero flux ($u_x = v_x = 0$ at $x = 0$), the PTWs disappear, to be replaced by spatially uniform oscillations (see Kay and Sherratt (1999)). Moreover, the speed/amplitude of the PTWs is independent of the seed in the random number generator used for the initial conditions. This suggests that the Dirichlet boundary condition robustly selects a particular member of the PTW family. The basic goal of the paper is to investigate the details of this selection process.

Before I begin analytical investigation of the solution shown in Figure 1, I mention one final and important result from the numerical simulations. The behavior illustrated in Figure 1 applies when $|\omega_1|$ is relatively small. For larger $|\omega_1|$, the long

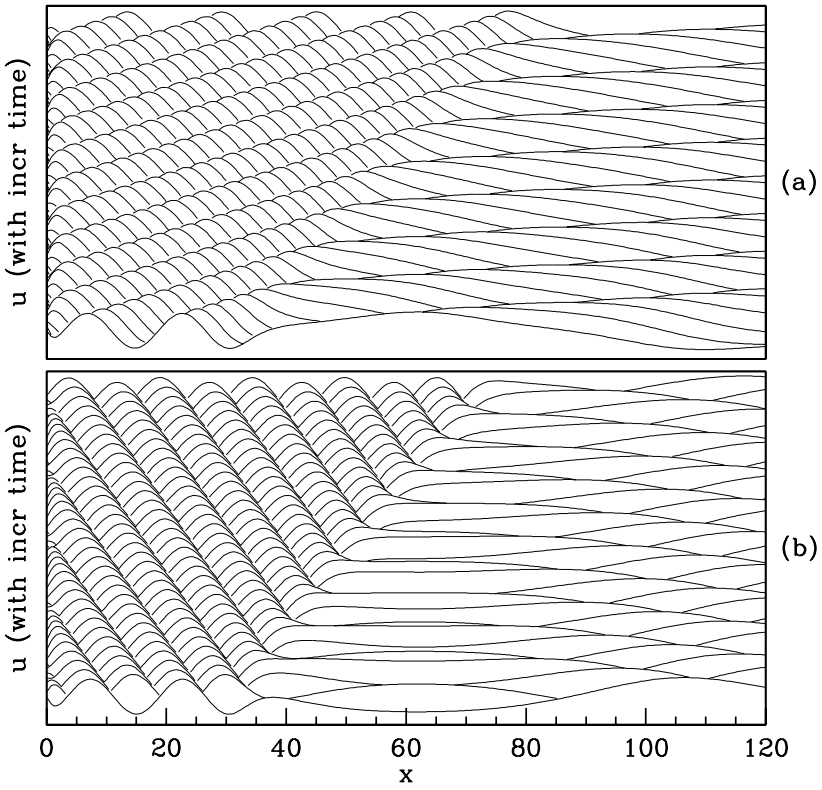


FIG. 1. *Solutions of* (2.1) *with boundary conditions* $u = v = 0$ *at* $x = 0$, *and* $u_x = v_x = 0$ *at* $x = 400$; *only part of the solution is plotted. A transition front moves across the domain, behind which PTWs develop, moving in the positive* x*-direction in* (a), *and the negative* x*-direction in* (b). *The solutions are space-time plots, with* u *plotted at equally spaced times between* $t = 100$ *and* $t = 200$ *(time increasing up the page). The solutions for* v *are qualitatively similar. Initial conditions* ($t = 0$) *are generated randomly as described in the main text. The parameter values are* $\omega_1 = 1.0$ *and* (a) $\omega_0 = 1.5$, (b) $\omega_0 = -1.3$. *The equations were solved numerically using a semi-implicit Crank–Nicolson method.*

term behavior consists not of PTWs, but of irregular spatiotemporal oscillations (Figure 2). Later in the paper, I will show that this behavior arises through the same basic mechanism and occurs when the PTW that is selected by the boundary conditions has an amplitude below r_{stab}, defined in (2.3), so that the selected PTW is unstable as a solution of the PDEs.

3. Reduction to an ODE system. The solutions shown in Figure 1 are illustrated more clearly by plotting r and θ_x rather than u and v (Figure 3). The solution changes rapidly from the initial conditions, until $r \approx 1$ and $\theta_x \approx 0$ everywhere away from the $x = 0$ boundary, corresponding to spatially homogeneous oscillations in u and v. A transition wave front in r and θ_x then develops, moving in the positive x-direction. Ahead of this front, $r \to 1$ and $\theta_x \to 0$; behind it, r and θ_x have constant values, r_{ptw} and ψ_{ptw} say, corresponding to the PTW. Numerical results indicate that this transition front moves with constant shape and speed, suggesting that one look

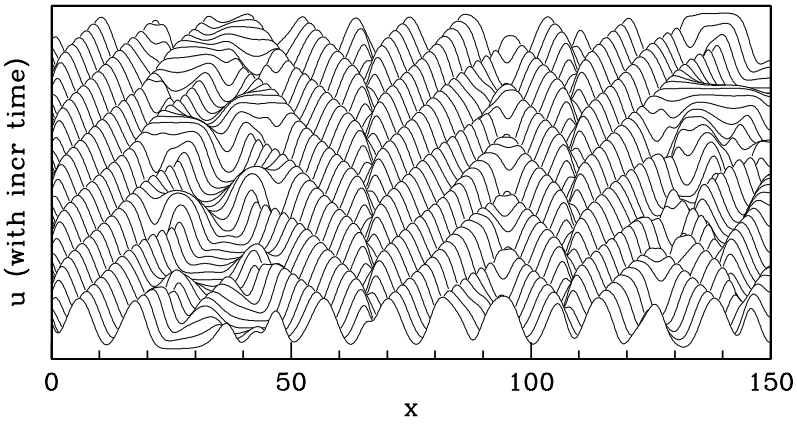


FIG. 2. *Long-term solution of* (2.1) *for which irregular spatiotemporal oscillations develop. The boundary conditions are* $u = v = 0$ *at* $x = 0$*, and* $u_x = v_x = 0$ *at* $x = 500$*; only part of the solution is plotted. Note that a band of PTWs is visible close to the* $x = 0$ *boundary. The solution is a space-time plot, with* u *plotted at equally spaced times between* $t = 1900$ *and* $t = 2000$ *(time increasing up the page). The solution for* v *is qualitatively similar. Initial conditions* ($t = 0$) *are generated randomly as described in the main text. The parameter values are* $\omega_0 = 1.5$ *and* $\omega_1 = 1.65$*. The equations were solved numerically using a semi-implicit Crank–Nicolson method.*

for solutions of (2.1) with the form

$$r(x,t) = \hat{r}(x - st) \quad \text{and} \quad \theta_x(x,t) = \hat{\psi}(x - st) \Rightarrow \theta(x,t) = \int^{z=x-st} \hat{\psi}(z)\, dz + f(t).$$

Here $s > 0$ is the front speed, and $f(t)$ is an arbitrary function of time that enters as a constant of integration. Substituting these solution forms into (2.4) gives

$$\hat{r}'' + s\hat{r}' + \hat{r}(1 - \hat{r}^2 - \hat{\psi}^2) = 0, \tag{3.1a}$$

$$\hat{\psi}' + s\hat{\psi} + \omega_0 - \omega_1\hat{r}^2 + 2\hat{\psi}\hat{r}'/\hat{r} = f'(t). \tag{3.1b}$$

Thus $f'(t)$ must be a constant, independent of t. Moreover, since $\hat{r} \to 1$ and $\hat{\psi} \to 0$ as $x - st \to \infty$, this constant value must be $\omega_0 - \omega_1$. Substituting $\hat{r} = r_{ptw}$ and $\hat{\psi} = \psi_{ptw}$ (values at $x - st = -\infty$) gives solutions for r_{ptw} and ψ_{ptw} in terms of s:

$$r_{ptw} = \sqrt{1 - \frac{s^2}{\omega_1^2}}, \qquad \psi_{ptw} = -\frac{s}{\omega_1}. \tag{3.2}$$

Unfortunately, these formulae cannot be used to obtain the values of r_{ptw} and ψ_{ptw}, since the front speed s is an unknown. However, they do provide one key piece of information: ψ_{ptw} has the sign opposite to that of ω_1, since s must be positive. This will be required in what follows.

Having established the sign of ψ_{ptw}, I now move on to consider the large time form of the solution for r and θ_x. Numerical simulations suggest that this is an equilibrium, which I denote by $r(x,t) = R(x)$ and $\theta_x(x,t) = \Psi(x)$. Hence $\theta = \int^x \Psi(\overline{x})\, d\overline{x} + g(t)$, where $g(t)$ is a constant of integration. Substituting these solution forms into (2.4) implies that $g'(t)$ must be a constant, which it is convenient to take as $\omega_0 - k$, where

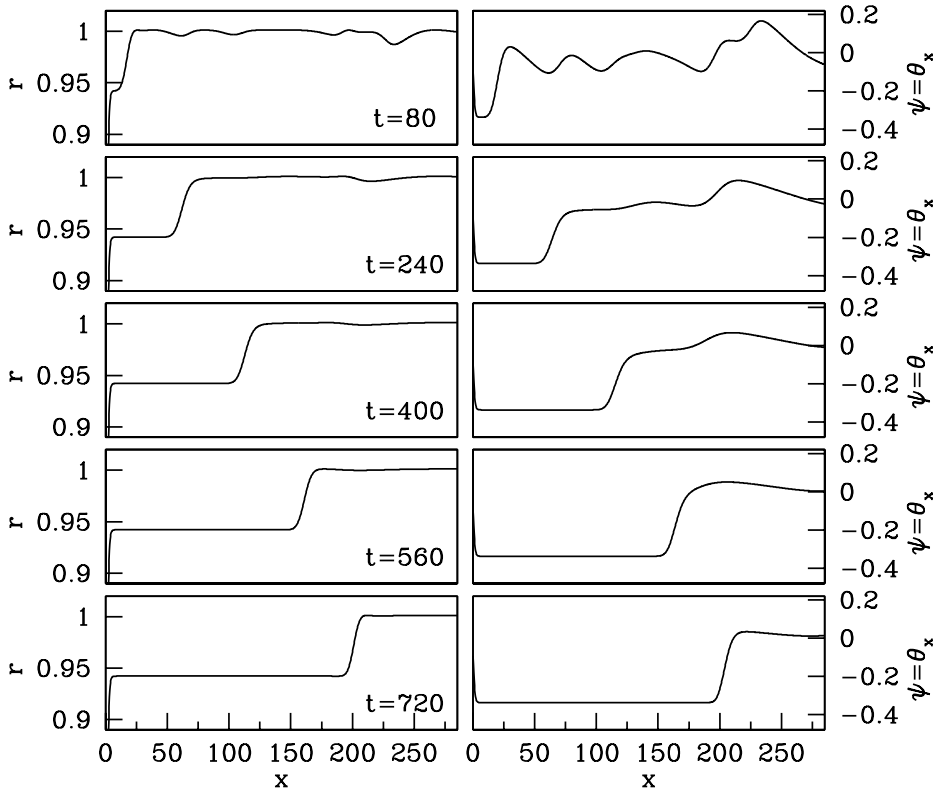


FIG. 3. *Time evolution of the solution of* (2.1) *subject to* $u = v = 0$ *at* $x = 0$. *I solve on* $0 < x < 400$ *with* $u_x = v_x = 0$ *at* $x = 400$, *to approximate a semi-infinite domain. The randomly generated initial condition rapidly evolves to* $u, v \approx 1$. *A transition wave front then develops, moving in the positive* x*-direction. Ahead of the front,* $r = 1$ *and* $\psi = 0$, *while behind it,* r *and* ψ *have values that are constants corresponding to a periodic travelling wave. The parameter values are* $\omega_0 = 0.3$ *and* $\omega_1 = 0.8$. *The equations were solved numerically using a semi-implicit Crank–Nicolson method.*

k is arbitrary and of either sign. The substitution also gives the following equations for R and Ψ:

$$R_{xx} + R(1 - R^2 - \Psi^2) = 0, \tag{3.3a}$$

$$\Psi_x + \frac{2\Psi R_x}{R} + k - \omega_1 R^2 = 0. \tag{3.3b}$$

The boundary condition $u = v = 0$ implies that $R = 0$ at $x = 0$, and I am looking for solutions for which R and Ψ tend to constant values, denoted r_{ptw} and ψ_{ptw}, as $x \to \infty$, with the sign of ψ_{ptw} opposite to that of ω_1. In a solution of this form, the values of r_{ptw} and ψ_{ptw} will be related to k and ω_1 by

$$r_{ptw} = \sqrt{\frac{k}{\omega_1}}, \qquad \psi_{ptw} = -\mathrm{sign}(\omega_1)\sqrt{1 - \frac{k}{\omega_1}}. \tag{3.4}$$

These are given simply by substituting the constant values into (3.3) and using the result that ψ_{ptw} and ω_1 have opposite signs; k is related to the speed s introduced above by $k = \omega_1 - s^2/\omega_1$. Note that k must have the same sign as ω_1.

It is convenient to rescale (3.3) as follows,

$$\phi = R\left(\frac{\omega_1}{k}\right)^{1/2}, \qquad w = R_x\left(\frac{\omega_1 - k}{k}\right)^{1/2} \cdot \frac{\text{sign}(\omega_1)}{k},$$

$$\Gamma = -\frac{\Psi}{R}\left(\frac{k}{\omega_1 - k}\right)^{1/2} \text{sign}(\omega_1), \qquad z = x\left(\frac{\omega_1}{\omega_1 - k}\right)^{1/2} \cdot k \cdot \text{sign}(\omega_1),$$

which gives

$$\phi_z = w, \tag{3.5a}$$

$$w_z = \frac{-\alpha}{k^2}\phi\left[1 - \phi^2 - \alpha\phi^2(\Gamma^2 - 1)\right], \tag{3.5b}$$

$$\Gamma_z = \frac{1 - 3w\Gamma - \phi^2}{\phi}, \tag{3.5c}$$

where $\alpha = 1 - k/\omega_1$, so that $0 \le \alpha \le 1$. In terms of these new variables, the required end conditions are

$$\phi = 0 \text{ at } z = 0 \quad \text{and} \quad \phi = 1,\ w = 0,\ \Gamma = 1 \text{ at } z = \infty. \tag{3.6}$$

Recall that the parameter k in (3.5) is an arbitrary constant of integration, and the initial question to be studied is for which values of k there are solutions of (3.5) satisfying these end conditions.

Numerical investigation of appropriate solutions to (3.5) is easiest if one integrates backwards in z from $(1, 0, 1)$. Straightforward calculation of the eigenvalues at this equilibrium shows that there is a unique stable eigenvector, and one can calculate numerically both trajectories corresponding to this eigenvector. For given values of k and ω_1, there is a solution of (3.5) of the required form if ϕ becomes zero along one of these trajectories. Numerical investigation indicates that this occurs at a large but discrete set of values of k. As illustrated in Figure 4, these values of k are widely separated when $|k|$ is just below $|\omega_1|$, and become closer together as $|k|$ approaches zero. (Recall that the sign of k is determined by that of ω_1.) Figure 5 illustrates how the critical values of k vary with ω_1.

4. Solution for small ω_1. I have been unable to calculate in general the values of k for which (3.5) has a solution of the required form. However for small $|\omega_1|$ the solutions, and thus the critical values of k, can be found using perturbation theory. Here I am exploiting the relative simplicity of (2.1) when $\omega_1 = 0$, a special case which has been used by a number of previous authors (for example, Kopell and Howard (1981)). Figure 6 illustrates the typical form of the solution for the largest few critical values of $|k|$ when $|\omega_1|$ is small. There is a characteristic solution form, with almost periodic oscillations in ϕ, w, and Γ. To calculate the solutions, it is enough to investigate one cycle of these oscillations.

Numerical solutions suggest that when $|\omega_1|$ is small, $|k|$ is also small, with the ratio $A \equiv \alpha/k^2$ being $O(1)$ as $|\omega_1| \to 0$. Then $\alpha^2/k^2 = A^2\omega_1^2 + O(\omega_1^4)$, and thus the equations (3.5) have the form

$$\phi_{zz} = -A\phi(1 - \phi^2) + \epsilon A^2\phi^3(\Gamma^2 - 1), \tag{4.1a}$$

$$\Gamma_z = \frac{1 - 3\phi_z\Gamma - \phi^2}{\phi}, \tag{4.1b}$$

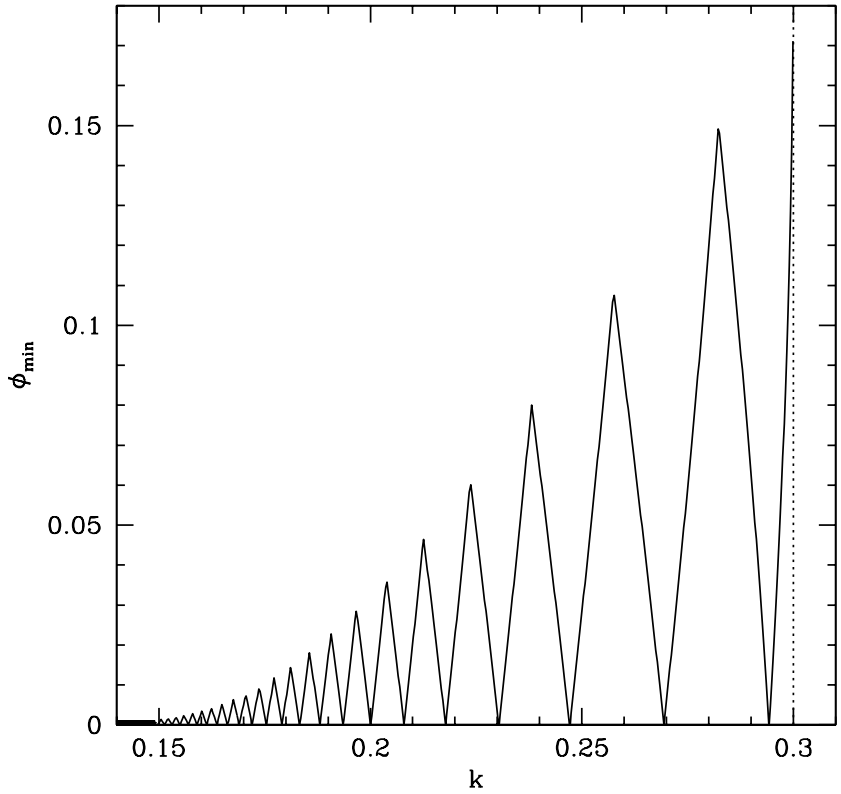


FIG. 4. *A typical plot of the variation with k of ϕ_{min}, the minimum value of ϕ along the two solution trajectories passing through the point $\phi = 1$, $w = 0$, $\Gamma = 1$. The variation has a "zig-zag" form, with ϕ_{min} approaching zero at a series of discrete values of k. The case shown is for $\omega_1 = 0.3$, and the dotted line is $k = \omega_1$; the constant k must lie between 0 and ω_1. The value of ϕ_{min} is calculated by solving* (3.5) *in the negative z-direction starting on the (unique) stable eigenvector at* $(1, 0, 1)$. *The solution is continued until $\phi > 1$, keeping track of the minimum value of ϕ. For each parameter set, this procedure must be followed twice, starting on either side of* $(1, 0, 1)$ *along the stable eigenvector.*

where $\epsilon = \omega_1^2$. The appropriate solution structure for these equations when $\epsilon \ll 1$ is illustrated in Figure 7. I consider one cycle of the solution in three separate regions a–c, with a fourth region a′ corresponding to region a in the next cycle. The boundary between regions a and b is the position at which ϕ has its local maximum, while region c is a thin layer centered on the local minimum of ϕ. No thin transition layer is required between regions a and b, but the location $z = z_1$ of the interface may depend on ϵ, as may the location $z = z_2$ of region c, and these dependencies must be found as part of the solution. The position $z = z_0$ to the left of region a is arbitrary.

In region a, there is no rescaling, and the leading order solutions ϕ_0^a, Γ_0^a satisfy (4.1) with ϵ set to zero. The two equations decouple, giving

$$\frac{d^2\phi_0^a}{dz^2} = -A\phi_0^a\left[1 - \phi_0^{a\,2}\right],$$
$$\frac{d\Gamma_0^a}{dz} = \frac{1 - 3\Gamma_0^a d\phi_0^a/dz - \phi_0^{a\,2}}{\phi_0^a}.$$

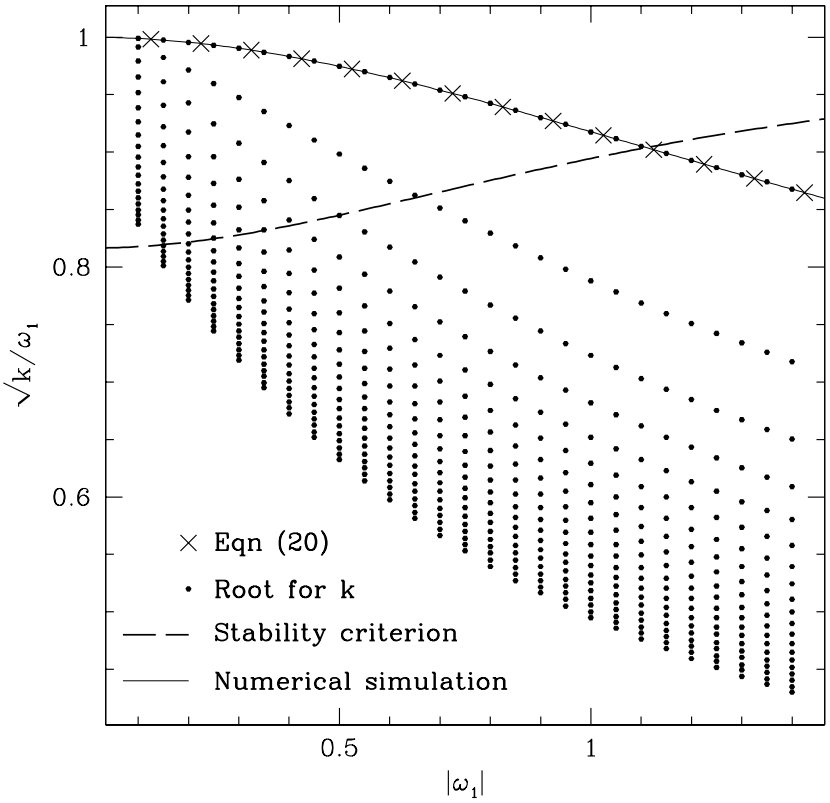


FIG. 5. *A plot of the amplitude* ($\sqrt{k/\omega_1}$) *of possible PTW solutions, corresponding to the critical values of* k *for which* (3.5) *has a solution of the required form. For a series of values of* ω_1, *I plot the amplitude corresponding to the largest* 20 *critical values of* k, *calculated numerically as discussed in the legend to Figure* 4. *Superimposed on the plot are the amplitude of PTWs predicted by numerical simulations of the PDEs* (2.1), *the theoretical prediction* (6.2) *of the PTW amplitude, and the curve determining PTW stability, which is given by* (2.3). *Note that this last curve does not refer to the stability of the solution of* (3.5), (3.6), *but simply to the PTW which this solution approaches as* $z \to \infty$. *Stability of this PTW is clearly a necessary but not sufficient condition for the stability of the solution of* (3.5), (3.6).

Thus

$$\left(\frac{d\phi_0^a}{dz}\right)^2 = \frac{1}{2}A({\phi_0^a}^2 - 1)^2 + C_1,$$

where C_1 is a constant of integration. Numerical solutions suggest that the local maxima in ϕ occur at $\phi = 1 + o(1)$ as $\epsilon \to 0$, and thus $C_1 = 0$. By construction, ϕ_0^a has positive slope, and thus further integration gives

$$\phi_0^a = \tanh\left[(z - z_0)\sqrt{\frac{A}{2}}\right], \tag{4.2a}$$

$$\Gamma_0^a = \sqrt{\frac{2}{9A}} + \frac{k_1}{\tanh^3[(z - z_0)\sqrt{A/2}\,]}. \tag{4.2b}$$

Here I am taking $\phi_0^a = 0$ at $z = z_0$, since numerical solutions suggest that the minima of ϕ are $o(1)$ as $\epsilon \to 0$; k_1 is a constant of integration.

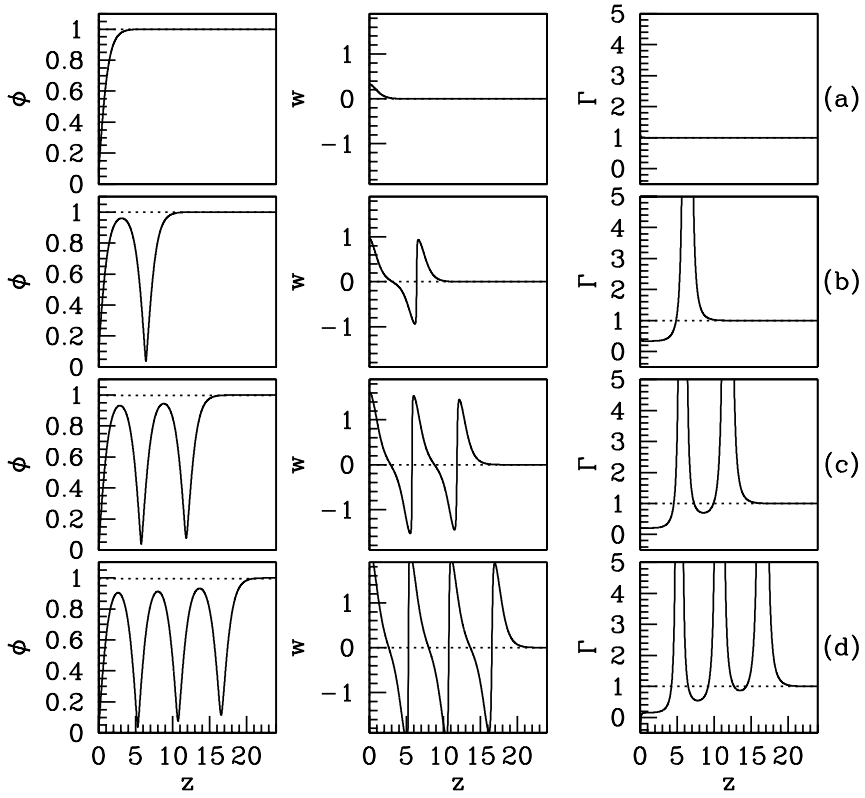


FIG. 6. *Typical solutions of* (3.5) *for small* $|\omega_1|$. *I show the solution for the largest three critical values of* $|k|$, *illustrating the typical form of the solution. For the largest value of* $|k|$, *the solution is monotonic in* ϕ, *and in successive solutions the variables cycle. The parameter values are* $\omega_1 = 0.03$ *and* (a) $k = 0.029994$, (b) $k = 0.029947$, (c) $k = 0.02986$, (d) $k = 0.02973$.

Similarly, in region b

$$\phi_0^b = -\tanh\left[(z - z_2)\sqrt{\frac{A}{2}}\right], \tag{4.3a}$$

$$\Gamma_0^b = -\sqrt{\frac{2}{9A}} + \frac{k_2}{\tanh^3[(z - z_2)\sqrt{A/2}\,]}, \tag{4.3b}$$

where k_2 is a constant of integration. The solutions in regions a and b are linked by conditions at $z = z_1$, namely, that $d\phi/dz = 0$ with ϕ and Γ continuous. Continuity of ϕ requires $z_1 - z_0 = z_2 - z_1 \equiv Z$, say, so that $z_1 = (z_0 + z_2)/2$. The zero derivative for ϕ then implies that $\operatorname{sech}^2\,[Z\sqrt{A/2}\,] = o(1)$, so that $Z \to \infty$ as $\epsilon \to 0$. Thus the widths of regions a and b become infinite as $\epsilon \to 0$. Further details of these widths are not determined at leading order, but higher order solutions (omitted for brevity) show that $Z = O_s(\log \epsilon)$ as $\epsilon \to 0$. Finally, continuity of Γ at $z = z_1$ gives a relationship between k_1 and k_2:

$$k_2 = k_1 + \sqrt{\frac{8}{9A}}\,;$$

here I use the fact that $\tanh(Z\sqrt{A/2}) = 1 + o(1)$ as $\epsilon \to 0$.

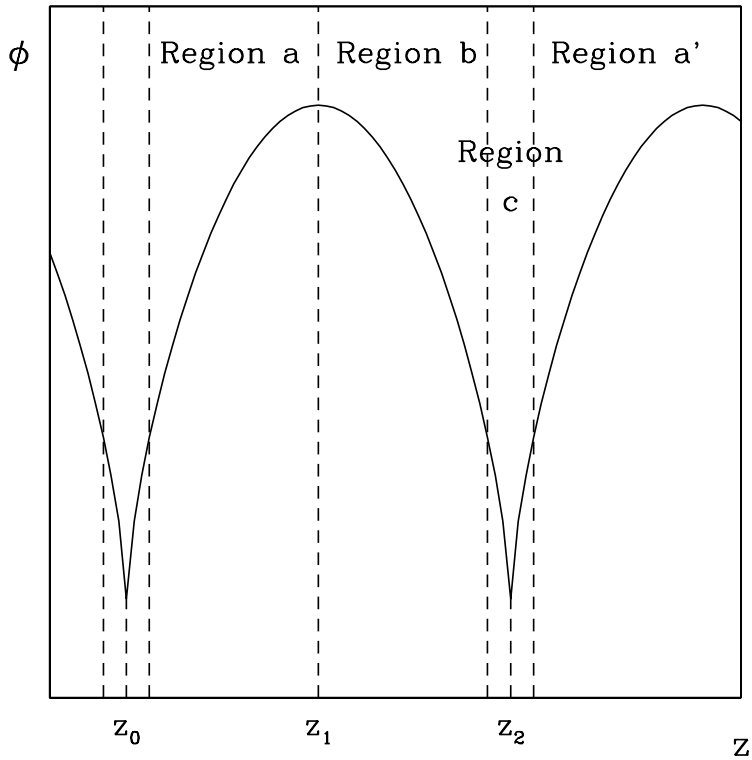


FIG. 7. *A schematic illustration of one cycle of the solution for* ϕ *of* (3.5) *when* ω_1 *is small, illustrating the different regions into which the solution is divided for the perturbation theory calculation.*

In region c, which is centered on the minimum of ϕ, a rescaling of the variables is required. Numerical solutions suggest that ϕ is small and Γ large in this region, with rapid changes in z. Therefore I substitute

$$\tilde{\phi} = \frac{\phi}{\nu_1}, \qquad \tilde{\Gamma} = \Gamma \cdot \nu_2, \qquad \zeta = \frac{z - z_2}{\nu_1}$$

into (3.5), where ν_1 and ν_2 are $o(1)$ as $\epsilon \to 0$; the rescaling of ϕ and z must be the same to allow matching of ϕ in regions b and c. This gives

$$\frac{d^2\tilde{\phi}}{d\zeta^2} = -\nu_1^2 A\tilde{\phi}(1 - \nu_1^2\tilde{\phi}^2) + \epsilon A^2\nu_1^4\tilde{\phi}^3\left(\frac{\tilde{\Gamma}^2}{\nu_2^2} - 1\right),$$
$$\frac{d\tilde{\Gamma}}{d\zeta} = \frac{\nu_2 - 3\tilde{\Gamma}d\tilde{\phi}/d\zeta - \nu_1^2\nu_2\tilde{\phi}^2}{\tilde{\phi}}.$$

Therefore the distinguished limit has $\epsilon^{1/2}\nu_1^2/\nu_2 = 1$, in which case the leading order solutions $\tilde{\phi}_0^c$ and $\tilde{\Gamma}_0^c$ satisfy

$$\frac{d^2\tilde{\phi}_0^c}{d\zeta^2} = A^2\tilde{\phi}_0^{c\,3}\tilde{\Gamma}_0^{c\,2},$$
$$\frac{d\tilde{\Gamma}_0^c}{d\zeta} = \frac{-3\tilde{\Gamma}_0^c d\tilde{\phi}_0^c/d\zeta}{\tilde{\phi}_0^c}.$$

Thus $\tilde{\Gamma}_0^c = k_3/\tilde{\phi}_0^{c\,3}$ and

$$\left(\frac{d\tilde{\phi}_0^c}{d\zeta}\right)^2 = k_4 - \frac{A^2 k_3^2}{\tilde{\phi}_0^{c\,2}}. \tag{4.4}$$

Here k_3 and k_4 are constants of integration, which are determined by matching the solution to that in region b. As $\zeta \to \pm\infty$, (4.4) implies that $\tilde{\phi}_0^c = \pm k_4^{1/2}\zeta + o(\zeta)$, so that $\tilde{\Gamma}_0^c = \pm k_3/(k_4^{3/2}\zeta^3) + o(\zeta^{-3})$. In comparison, as $z \to z_2^-$, $\phi_0^b \sim -(z - z_2)\sqrt{A/2}$ and $\Gamma_0^b \sim (2/A)^{3/2} k_2 (z - z_2)^{-3}$. Therefore, matching requires $k_4 = A/2$, $k_2 = k_3$, and $\nu_2 = \nu_1^3 \Rightarrow \nu_1 = \epsilon^{1/2}$, $\nu_2 = \epsilon^{3/2}$.

The final step in the leading order solution is to determine behavior in region a′. The solution here is the same as in region a, but with new constants of integration:

$$\phi_0^{a'} = \tanh\left[(z - z_0')\sqrt{\frac{A}{2}}\right], \tag{4.5a}$$

$$\Gamma_0^{a'} = \sqrt{\frac{2}{9A}} + \frac{k_1'}{\tanh^3[(z - z_0)\sqrt{A/2}\,]}. \tag{4.5b}$$

Matching this solution with that in region c is directly analogous to the matching of solutions in regions b and c, and requires $k_1' = k_3$. Hence $k_1' = k_1 + \sqrt{8/9A}$.

Consider now a solution of the form illustrated in Figure 6, and with N local maxima in ϕ before ϕ approaches 1 asymptotically. Let k_1^i be the constant of integration k_1 in the leading order solution in region a before the ith maximum ($i = 1, \dots, N-1$) or in the part of the solution in which ϕ approaches 1 asymptotically ($i = N$). Then I have shown that $k_1^i = k_1^{i-1} + \sqrt{8/9A}$. Now I require Γ finite at $z = 0$, and $\Gamma \to 1$ as $z \to \infty$. Thus $k_1^1 = 0$ and $\sqrt{2/9A} + k_1^N = 1 + O(\epsilon^{1/2})$; the correction is based on the next order term in the expansion for Γ in region a (omitted for brevity). Therefore $\sqrt{2/9A} + (N-1)\sqrt{8/9A} = 1$, which can be rearranged to give $A = \frac{8}{9}(N - \frac{1}{2})^2$.

This calculation shows that there are a discrete but infinite set of values of A for which (4.1) has a solution of the required form. These correspond to the values of k plotted in Figure 4. To make this correspondence precise, recall that $A = \alpha/k^2$, with $\alpha = 1 - k/\omega_1$ and $\omega_1 = \epsilon^{1/2}$, so that $k = \epsilon^{1/2} - A\epsilon^{3/2} + O(\epsilon^{5/2})$. Therefore, at least for sufficiently small ω_1, there are solutions of (3.5) for an infinite set of values of k, given by $k = \omega_1 - \frac{8}{9}(N - \frac{1}{2})^2\omega_1^3 + O(\omega_1^5)$ ($N = 1, 2, \dots$). The solution with index N has $N - 1$ local maxima and minima in ϕ. The values of ϕ at these extrema depend on ϵ: the minima have a height that is $O(\epsilon^{1/2}) = O(\omega_1)$, and the leading order correction to ϕ in regions a and b is $O(\epsilon)$, implying that the maxima have a height that is $1 - O(\epsilon) = 1 - O(\omega_1^2)$.

5. Solutions for general ω_1. In the plot of the critical values of k in Figure 4, I superimpose a plot of the amplitude of the PTW that develops in numerical solutions of (2.1) subject to $u = v = 0$ at $x = 0$ on $0 < x < \infty$. In every case, this corresponds to the critical value of k with largest absolute value. This is despite the fact that each of the other critical values of k corresponds to a possible long term solution of (2.1). In this section, I will discuss in more detail the structure of the ϕ-w-Γ phase plane, to give further insight into the equation forms at different critical values of k.

I begin by investigating the behavior of (3.5) near $\phi = 0$, which is a singularity. As $\phi \to 0$, simple inspection of (3.5c) shows that $|d\Gamma/dz| \to \infty$ away from the curve $w\Gamma = 1/3$. Behavior near this curve requires more careful investigation, and I look

for trajectories of the form $w = w_0 + \tilde{w}(\phi)$, $\Gamma = w_0/3 + \tilde{\Gamma}(\phi)$, where w_0 is an arbitrary constant and $\tilde{w}$ and $\tilde{\Gamma}$ are $o(1)$ as $\phi \to 0$. Substituting these solution forms into (3.5) shows that to leading order

$$\tilde{w}(\phi) = \frac{-\alpha}{2k^2 w_0}\phi^2, \qquad \tilde{\Gamma}(\phi) = \frac{1}{10w_0}\left(\frac{\alpha}{w_0^2 k^2} - 2\right)\phi^2. \tag{5.1}$$

Therefore, despite the singularity of (3.5) at $\phi = 0$, there is a family of nonsingular trajectories which cross the $\phi = 0$ plane through the curve $\Gamma w = 1/3$. Such a curve is sometimes known as a "hole in a singular barrier" (Perumpanani et al. (1999); Pettet, McElwain, and Norbury (2000)). Of the trajectories crossing $\phi = 0$ through this curve, only those crossing at positive values of w and Γ are of interest, and taken together, these make up a surface in ϕ-w-Γ phase space, which I denote by $\mathcal{S}(k)$. There is a trajectory of the required form for a given value of k if and only if this surface $\mathcal{S}(k)$ contains the point $\phi = 1$, $w = 0$, $\Gamma = 1$.

The surface $\mathcal{S}(k)$ has a very complex form, especially for small values of k, making visualization in three dimensions very difficult. I have found it most instructive to plot a cross section of the surface, and a natural cross section is the $w = 0$ plane, which is illustrated in Figure 8 for the largest three critical values of k when $\omega_1 = 1$. Note that, in each case, the intersection includes the point $\Gamma = \phi = 1$.

In Figure 8(a) the trajectories making up $\mathcal{S}(k)$ intersect the $w_z < 0$ portion of the $w = 0$ plane only once. This implies that the corresponding solution of (3.5), (3.6) is monotonic in ϕ, and numerical results suggest that this is also true for other ω_1, when k is at its largest critical value. Conversely, for the other critical values of k, numerical solutions suggest that the trajectory passing through $\phi = \Gamma = 1$, $w = 0$ does so only after previously crossing the $w = 0$ plane. Based on this, I hypothesize that for given ω_1 there is only one solution of (3.5), (3.6) that is monotonic in ϕ, namely, that corresponding to the largest initial value of k. Further, I hypothesize that any solutions of (3.5) with nonmonotonic ϕ are unstable as solutions of (2.1). A number of results of the form "nonmonotonicity implies instability" are known for scalar reaction-diffusion equations (Hagan (1981b), Henry (1981)), and numerical simulations using the solutions of (3.5), (3.6) with small perturbations as initial conditions for (2.1) suggest that a corresponding result applies in this case. Taken together, these hypotheses provide an explanation for the solution of (2.1) always corresponding to the solution of (3.5), (3.6) with the largest critical value of k.

Although I cannot prove these hypotheses, I will present a sketch proof of the first one. Proof that a solution that is monotonic in ϕ exists for some value of k is obtained by direct construction. Numerical solutions of (2.1) with $u = v = 0$ at $x = 0$ suggest that the ratio of Ψ and R is constant in the observed solution. Based on this, I look for a solution of (3.5) in which $\Gamma \equiv 1$. The equations (3.5) then give

$$\phi_z = w, \tag{5.2a}$$

$$w_z = -\left(\frac{\alpha}{k^2}\right)\phi(1 - \phi^2), \tag{5.2b}$$

$$3w + \phi^2 = 1. \tag{5.2c}$$

Combining (5.2a) and (5.2b), and requiring $\phi = 1$ when $w = 0$, gives

$$w^2 = \frac{\alpha}{2k^2}(1 - \phi^2)^2.$$

This is consistent with (5.2c) if and only if $k^2 = 9\alpha/2$. (Recall that the sign of k must

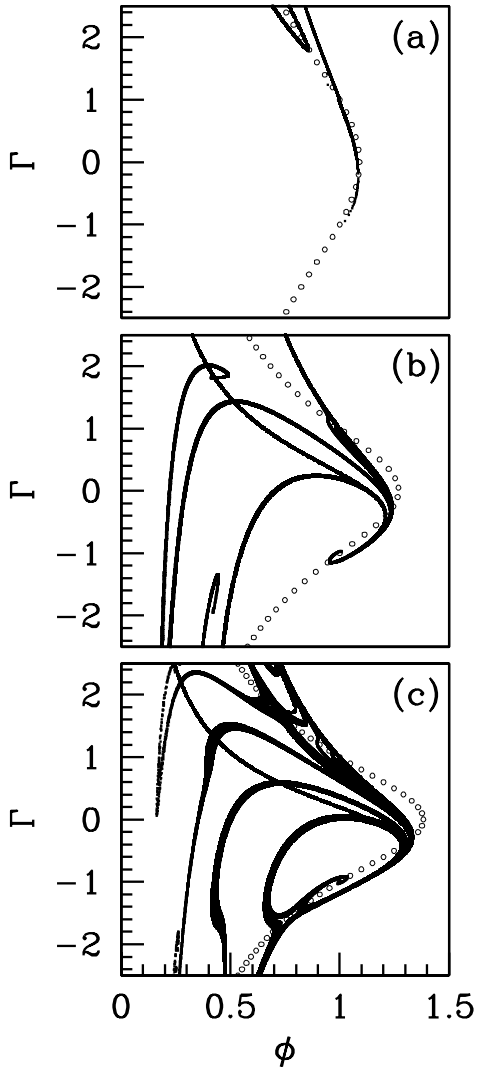


FIG. 8. *An illustration of the intersection between the surface $\mathcal{S}(k)$ and the $w = 0$ plane for the largest three critical values of k when $\omega_1 = 1$. The small circles represent the curve $\phi^2\left[1-\alpha+\alpha\Gamma^2\right] = 1$; to the right of this curve, $w_z > 0$, so that the trajectories are crossing from negative to positive w, and to the left of the curve, $w_z < 0$. I have solved* (3.5) *numerically for values of w_0 increasing from* 0.01 *in increments of 10^{-6}; initial conditions are $w = w_0 + \tilde{w}$, $\Gamma = \Gamma_0 + \tilde{\Gamma}$ with $\tilde{w}$ and $\tilde{\Gamma}$ given by* (5.1)*, and $\phi = 0.001$. In each of these solutions, I record and plot each point at which the $w = 0$ plane is crossed. The values of k are* (a) $k = 0.842$, (b) $k = 0.621$, (c) $k = 0.523$.

be the same as that of ω_1.) Recalling that $\alpha = 1 - k/\omega_1$, this implies

$$k = k^* \equiv \frac{-9 + \sqrt{81 + 72\omega_1^2}}{4\omega_1}.$$

Note that the solution trajectory corresponding to this value of k is monotonic in w as well as ϕ, and thus lies within $\hat{\mathcal{S}}(k)$, the subset of $\mathcal{S}(k)$ formed by the portion of the trajectories starting on $\phi = 0$, $w\Gamma = 1/3$, until they leave the region $\phi > 0$, $w > 0$, $\phi^2[1 + \alpha(\Gamma^2 - 1)] < 1$; this last condition is simply $w_z < 0$.

To study uniqueness, it is convenient to consider varying k with α, rather than ω_1, fixed. Suppose that there is a value of k, k_1 say, not equal to k^*, for which there is a trajectory that is monotonic in ϕ and connects $\phi = 0$, $w\Gamma = 1/3$ with $w = 0$, $\phi = \Gamma = 1$. To be specific, I assume that $k_1 < k^*$, though a corresponding argument is valid if $k_1 > k^*$. The solution trajectories for $k = k_1$ and k^* are contained in $\mathcal{S}(k_1)$ and $\hat{\mathcal{S}}(k^*)$, respectively; note that in general the trajectories for k^* and k_1 will cross $\phi = 0$ at different points along $w\Gamma = 1/3$. Straightforward examination of the eigenvalues and eigenvectors of (3.5) at $w = 0$, $\phi = \Gamma = 1$ implies that $\mathcal{S}(k_1)$ lies above $\hat{\mathcal{S}}(k^*)$ close to this point, in the sense that w is greater on $\mathcal{S}(k_1)$ than on $\hat{\mathcal{S}}(k^*)$. Moreover eliminating w_0 in (5.1) implies that $\hat{\mathcal{S}}(k^*)$ lies above $\mathcal{S}(k_1)$ for sufficiently small ϕ. This suggests that the surfaces $\mathcal{S}(k_1)$ and $\hat{\mathcal{S}}(k^*)$ intersect, which is impossible since (3.5b) implies that w_z increases with k in the region $\phi^2[1 + \alpha(\Gamma^2 - 1)] < 1$. It remains possible that $\mathcal{S}(k_1)$ and $\hat{\mathcal{S}}(k^*)$ wind around one another; I have been unable to rule this out, although numerical solutions suggest that it does not happen.

6. The form of the observed periodic travelling wave. The sketch proof in the previous section is constructive, in the sense that it explicitly determines a formula for k^*. The corresponding solution of (3.5) can easily be determined from (5.2c) as $\phi = \tanh(z/3)$. Converting back to the original variables gives

$$R(x) = r_{ptw} \tanh\left(\frac{x}{\sqrt{2}}\right), \qquad \Psi(x) = \psi_{ptw} \tanh\left(\frac{x}{\sqrt{2}}\right), \tag{6.1}$$

where

$$r_{ptw} = \left\{\frac{1}{2}\left[1 + \sqrt{1 + \frac{8}{9}\omega_1^2}\right]\right\}^{-1/2}, \qquad \psi_{ptw} = -\text{sign}(\omega_1)\left\{\frac{\sqrt{1 + \frac{8}{9}\omega_1^2} - 1}{\sqrt{1 + \frac{8}{9}\omega_1^2} + 1}\right\}^{1/2}. \tag{6.2}$$

The solution given by (6.1), (6.2) is an excellent match with the long term behavior predicted by numerical simulations of (2.1) subject to $u = v = 0$ at $x = 0$, as illustrated in Figure 4. Moreover, (6.2) enables direct determination of the properties of PTWs generated by Dirichlet boundary conditions. For example, substitution of (6.2) into (2.3) gives the condition for PTW stability as

$$8\omega_1^6 + 16\omega_1^4 - 10\omega_1^2 - 27 < 0 \quad \Longleftrightarrow \quad |\omega_1| < 1.110468\ldots. \tag{6.3}$$

A detailed numerical study shows that cases such as that shown in Figure 3, in which irregular oscillations develop, correspond exactly to values of ω_1 above this critical value. Similarly the direction of the PTWs can be determined—this depends on ω_0 as well as ω_1. The PTW solution for u and v, given in (2.2), moves in the positive x-direction if and only if $(\omega_0 - \omega_1 r^2)$ and θ_x have opposite signs. Therefore the PTW given by (6.2) moves in the positive x-direction if and only if

$$\psi_{ptw} \cdot (\omega_0 - \omega_1 r_{ptw}^2) < 0 \Leftrightarrow \omega_0 \text{ and } \omega_1 \text{ have the same sign} \tag{6.4a}$$

$$\text{and } |\omega_0| > \frac{2|\omega_1|}{1 + \sqrt{1 + \frac{8}{9}\omega_1^2}}. \tag{6.4b}$$

Conditions (6.3) and (6.4) are illustrated graphically in Figure 9.

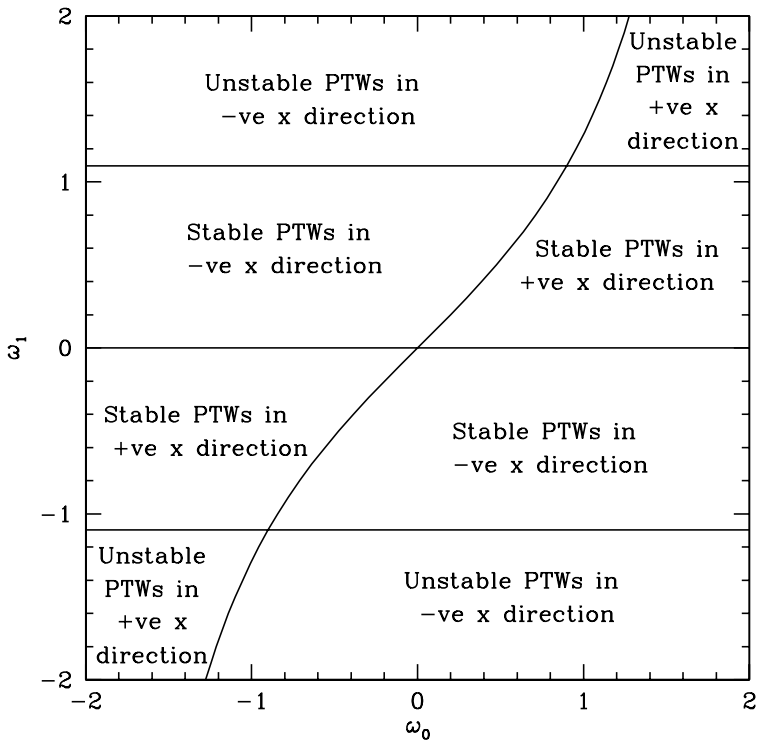


FIG. 9. *An illustration of conditions* (6.3) *and* (6.4) *for the stability and direction of the PTW solution generated in the region* $x > 0$ *by the boundary condition* $u = v = 0$ *at* $x = 0$.

7. Discussion. Using a combination of analysis and numerical simulation, I have shown that when the λ-ω system (2.1) is solved on a finite domain subject to zero Dirichlet boundary conditions, PTWs develop. I have shown that there is a discrete family of possible wave amplitudes for which solutions exist, but my results suggest that in only one of these cases does the amplitude vary monotonically in space. I hypothesize that this family is the only stable solution, implying that the boundary conditions select a unique PTW amplitude that is independent of initial conditions. A formula for this amplitude is given in (6.2).

An obvious extension of the work presented in this paper is to consider behavior in two spatial dimensions. I have done a limited program of numerical simulations of the λ-ω system (2.1) in two dimensions, and a typical result is illustrated in Figure 10. For this figure, equations (2.1) were solved on an approximately circular but irregular domain, with the boundary condition $u = v = 0$. A solution of "target pattern" form develops, moving inwards from the boundary; an animation of this solution can be seen at www.ma.hw.ac.uk/∼jas/supplements/dirichlet/. This solution is a natural two-dimensional extension of the one-dimensional results I have been discussing, in which planar PTW solutions are modulated by the curvature of the domain. A natural topic for future analytical work is the spatial scale over which the curvature of the wave fronts varies.

The mathematical study of PTWs has been given a significant boost recently by the identification by ecologists of PTWs in cyclic populations. This empirical work is slow, requiring spatiotemporal field data gathered over many years. (Pop-

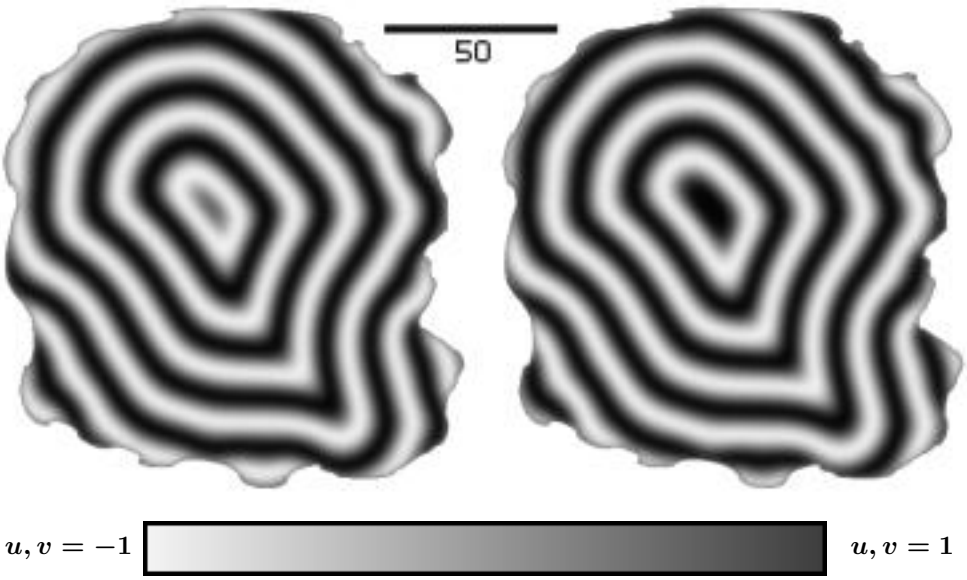


FIG. 10. *Numerical simulation of the λ-ω system* (2.1) *in two space dimensions on an irregular domain with boundary condition* $u = v = 0$ *on the perimeter. The solutions for* u *(left) and* v *(right) are shown at time* $t = 400$*; a "target pattern"-type solution moves inwards from the boundary towards the center of the domain. The initial conditions (at* $t = 0$*) are generated randomly, and the parameters are* $\omega_0 = \omega_1 = 1$*. The size of the spatial domain is indicated by the scale bar, which is* 50 *space units long. An animation corresponding to this figure can be seen at* `http://www.ma.hw.ac.uk/~jas/supplements/dirichlet/`*. The equations were solved numerically using an alternating direction semi-implicit method.*

ulation cycles typically have a period of between 4 and 10 years.) The analysis of the data then depends on spatiotemporal statistical methods developed only recently (Bjørnstad, Ims, and Lambin (1999)). For these reasons, it is too early to assess how widespread PTWs are in real populations. However, there is now very strong evidence for the existence of a PTW in cyclic field vole populations in the Kielder forest, on the Scotland–England border (Lambin et al. (1998); MacKinnon et al. (2001)), and more limited evidence for PTWs in various other populations, including red grouse in Northeast Scotland (Moss, Elston, and Watson (2000)), water voles in Eastern France (Giraudoux et al. (1997)), and larch budmoths in the European Alps (Bjørnstad et al. (2002)). The major question raised by these ecological studies is, what are the cause(s) of the PTWs? One possibility is that the PTWs are generated by the invasion of a prey population by predators (Sherratt, Lewis, and Fowler (1995); Sherratt et al. (2000)). However, once a whole domain has been invaded, PDE models predict that the waves will gradually die out (with zero flux boundary conditions; see Kay and Sherratt (1999)). Thus this mechanism requires a recent invasion, which does not apply in most cases. In contrast, the mechanism studied in this paper is consistent with conditions found in many real ecological systems. A Dirichlet boundary condition (with population density equal to zero) is appropriate when the domain is surrounded by a region of different habitat in which the population cannot survive—for example, an area of woodland surrounded by open fields. In a companion paper (Sherratt et al. (2002)), coworkers and I show that numerical simulations of both PDE and spatially discrete predator-prey models predict the generation of periodic waves by Dirichlet

boundary conditions. Extension of the analytical results in the present paper to realistic predator-prey models is a major challenge for future work, which would enable systematic prediction of the occurrence of periodic waves in real populations.

Acknowledgments. This work arose from discussions with Tom Sherratt and Xavier Lambin, to whom I am very grateful. In particular, the original idea for periodic wave generation by Dirichlet boundary conditions in ecological systems is due to Tom Sherratt. I also thank Vadim Biktashev and Jack Carr for helpful discussions.

REFERENCES

P. Ashwin, M. V. Bartuccelli, T. J. Bridges, and S. A. Gourley (2002), *Travelling fronts for the KPP equation with spatio-temporal delay*, Z. Angew. Math. Phys., 53, pp. 103–122.

J. F. G. Auchmuty and G. Nicolis (1976), *Bifurcation analysis of reaction-diffusion equations.* III. *Chemical oscillations*, Bull. Math. Biol., 38, pp. 325–350.

O. N. Bjørnstad, R. A. Ims, and X. Lambin (1999), *Spatial population dynamics: Analyzing patterns and processes of population synchrony*, Trends in Ecology and Evolution, 14, pp. 427–432.

O. N. Bjørnstad, M. Peltonen, A. M. Liebhold, W. Baltensweiler (2002), *Waves of larch budmoth outbreaks in the European Alps*, Science, 298, pp. 1020–1023.

G. B. Ermentrout (1981), *Stable small amplitude solutions in reaction-diffusion systems*, Quart. Appl. Math., 39, pp. 61–86.

G. B. Ermentrout, X. Chen, and Z. Chen (1997), *Transition fronts and localised structures in bistable reaction-diffusion equations*, Phys. D, 108, pp. 147–167.

G. B. Ermentrout and N. Kopell (1984), *Frequency plateaus in a chain of weakly coupled oscillators*, I, SIAM J. Math. Anal., 15, pp. 215–237.

G. B. Ermentrout and N. Kopell (1986), *Symmetry and phase-locking in chains of weakly coupled oscillators*, Commun. Pure Appl. Anal., 49, pp. 623–660.

P. Giraudoux, P. Delattre, M. Habert, J. P. Quere, S. Deblay, R. Defaut, R. Duhamel, M. F. Moissenet, D. Salvi, and D. Truchetet (1997), *Population dynamics of fossorial water vole (Arvicola terrestris scherman): A land use and landscape perspective*, Agr. Ecosyst. Environ., 66, pp. 47–60.

P. S. Hagan (1981a), *Target patterns in reaction-diffusion systems*, Adv. Appl. Math., 2, pp. 400–416.

P. S. Hagan (1981b), *The instability of non-monotonic wave solutions of parabolic equations*, Stud. Appl. Math., 64, pp. 57–88.

D. Henry (1981), *Geometric Theory of Semilinear Parabolic Equations*, Springer-Verlag, Berlin.

T. Kapitula (1994), *On the nonlinear stability of plane waves for the Ginzburg–Landau equation*, Comm. Pure Appl. Math., 47, pp. 831–841.

A. L. Kay and J. A. Sherratt (1999), *On the persistence of spatiotemporal oscillations generated by invasion*, IMA J. Appl. Math., 63, pp. 199–216.

A. L. Kay and J. A. Sherratt (2000), *Spatial noise stabilizes periodic wave patterns in oscillatory systems on finite domains*, SIAM J. Appl. Math., 61, pp. 1013–1041.

N. Kopell (1981), *Target pattern solutions to reaction-diffusion equations in the presence of impurities*, Adv. Appl. Math., 2, pp. 389–399.

N. Kopell and L. N. Howard (1973), *Plane wave solutions to reaction-diffusion equations*, Stud. Appl. Math., 52, pp. 291–328.

N. Kopell and L. N. Howard (1981), *Target patterns and horseshoes from a perturbed central force problem: Some temporally periodic solutions to reaction-diffusion problems*, Stud. Appl. Math., 64, pp. 1–56.

N. Kopell, G. B. Ermentrout, and T. L. Williams (1991), *On chains of oscillators forced at one end*, SIAM J. Appl. Math., 51, pp. 1397–1417.

X. Lambin, D. A. Elston, S. J. Petty, and J. L. MacKinnon (1998), *Spatial asynchrony and periodic travelling waves in cyclic populations of field voles*, Proc. Roy. Soc. London Ser. B, 265, pp. 1491–1496.

J. L. MacKinnon, X. Lambin, D. A. Elston, C. J. Thomas, T. N. Sherratt, and S. J. Petty (2001), *Scale invariant spatio-temporal patterns of field vole density*, J. Animal Ecology, 70, pp. 101–111.

K. Maginu (1981), *Stability of periodic travelling wave solutions with large spatial periods in reaction-diffusion systems*, J. Differential Equations, 39, pp. 73–99.

R. Moss, D. A. Elston, and A. Watson (2000), *Spatial asynchrony and demographic travelling waves during red grouse population cycles*, Ecology, 81, pp. 981–989.

H. Nagashima (1991), *Target patterns and pacemakers in a reaction-diffusion system*, J. Phys. Soc. Japan, 60, pp. 2797–2799.

A. J. Perumpanani, J. Norbury, J. A. Sherratt, and H. M. Byrne (1999), *A two parameter family of travelling waves with a singular barrier arising from the modelling of matrix mediated malignant invasion*, Phys. D, 126, pp. 145–159.

S. V. Petrovskii and H. Malchow (1999), *A minimal model of pattern formation in a prey predator system*, Math. Comput. Modelling, 29, pp. 49–63.

S. V. Petrovskii and H. Malchow (2000), *Critical phenomena in plankton communities: KISS model revisited*, Nonlinear Anal. Real World Appl., 1, pp. 37–51.

G. J. Pettet, D. L. McElwain, and J. Norbury (2000), *Lotka–Volterra equations with chemotaxis: Walls, barriers and travelling waves*, IMA J. Math. Appl. Med. Biol., 17, pp. 395–413.

L. Ren and G. B. Ermentrout (1998), *Monotonicity of phaselocked solutions in chains and arrays of nearest-neighbor coupled oscillators*, SIAM J. Math. Anal., 29, pp. 208–234.

J. L. Romero, M. L. Gandarias, and E. Medina (2000), *Symmetries, periodic plane waves and blow-up of lambda-omega systems*, Phys. D, 147, pp. 259–272.

S. K. Scott, B. R. Johnson, A. F. Taylor, and M. R. Tinsley (2000), *Complex chemical reactions—A review*, Chem. Eng. Sci., 55, pp. 209–215.

J. A. Sherratt (1994a), *On the evolution of periodic plane waves in reaction-diffusion systems of λ–ω type*, SIAM J. Appl. Math., 54, pp. 1374–1385.

J. A. Sherratt (1994b), *Irregular wakes in reaction-diffusion waves*, Phys. D, 70, pp. 370–384.

J. A. Sherratt, M. A. Lewis, and A. C. Fowler (1995), *Ecological chaos in the wake of invasion*, Proc. Natl. Acad. Sci. USA, 92, pp. 2524–2528.

J. A. Sherratt, X. Lambin, C. J. Thomas, and T. N. Sherratt (2002), *Generation of periodic waves by landscape features in cyclic predator-prey systems*, Proc. Roy. Soc. London Ser. B, 269, pp. 327–334.

T. N. Sherratt, X. Lambin, S. J. Petty, J. L. MacKinnon, C. F. Coles, and C. J. Thomas (2000), *Application of coupled oscillator models to understand extensive synchrony domains and travelling waves in populations of the field vole in Kielder forest, UK*, J. Appl. Ecol., 37, pp. 148–158.

J. Sneyd and J. Sherratt (1997), *On the propagation of calcium waves in an inhomogeneous medium*, SIAM J. Appl. Math., 57, pp. 73–94.

A. T. Winfree (2001), *The Geometry of Biological Time*, Springer-Verlag, New York.

SIAM J. APPL. MATH.
Vol. 63, No. 5, pp. 1539–1560

SOLITARY WAVES IN LAYERED NONLINEAR MEDIA*

RANDALL J. LEVEQUE† AND DARRYL H. YONG‡

Abstract. We study longitudinal elastic strain waves in a one-dimensional periodically layered medium, alternating between two materials with different densities and stress-strain relations. If the impedances are different, dispersive effects are seen due to reflection at the interfaces. When the stress-strain relations are nonlinear, the combination of dispersion and nonlinearity leads to the appearance of solitary waves that interact like solitons. We study the scaling properties of these solitary waves and derive a homogenized system of equations that includes dispersive terms. We show that pseudospectral solutions to these equations agree well with direct solutions of the hyperbolic conservation laws in the layered medium using a high-resolution finite volume method. For particular parameters we also show how the layered medium can be related to the Toda lattice, which has discrete soliton solutions.

Key words. nonlinear elasticity, solitons, layered media, homogenization, Toda lattice

AMS subject classifications. 74J35, 74Q10, 35B27, 35L65, 37K60

DOI. 10.1137/S0036139902408151

1. Introduction.

Consider a heterogeneous medium composed of alternating layers of two different materials labeled A and B. The layers have widths $\delta_A = \alpha\delta$ and $\delta_B = (1-\alpha)\delta$, repeating periodically with period δ. The densities of the two materials are ρ_A and ρ_B, respectively, and their response to compression or expansion is characterized by the stress-strain relations $\sigma_A(\epsilon)$ and $\sigma_B(\epsilon)$. Then compressional waves propagating in the direction of layering are modeled by the one-dimensional hyperbolic system of conservation laws

$$\epsilon_t(x,t) - u_x(x,t) = 0, \tag{1.1a}$$

$$(\rho(x)u(x,t))_t - \sigma(\epsilon(x,t),x)_x = 0, \tag{1.1b}$$

where $\epsilon(x,t)$ is the strain and $u(x,t)$ the velocity. For the layered medium we have

$$(\rho(x),\sigma(\epsilon,x)) = \begin{cases} (\rho_A,\sigma_A(\epsilon)) & \text{if } j\delta < x < (j+\alpha)\delta \text{ for some integer } j, \\ (\rho_B,\sigma_B(\epsilon)) & \text{otherwise.} \end{cases} \tag{1.2}$$

For sufficiently small strains, the response can be modeled by linear constitutive relations

$$\sigma_A(\epsilon) = K_A\epsilon, \qquad \sigma_B(\epsilon) = K_B\epsilon, \tag{1.3}$$

where the bulk moduli K_A and K_B of both materials are constants. Waves having long wavelength relative to the layer width can be modeled by a homogenized linear PDE that has the form of a wave equation with small dispersive term. The effective wave speed and dispersion coefficient can be calculated from the above parameters describing the layered medium. This linear case is reviewed in section 2.

*Received by the editors May 22, 2002; accepted for publication (in revised form) October 16, 2002; published electronically June 12, 2003. This work was supported in part by DOE grant DE-FG03-00ER25292 and NSF grants DMS-9803442 and DMS-0106511.
http://www.siam.org/journals/siap/63-5/40815.html
†Department of Applied Mathematics, University of Washington, Box 352420, Seattle, WA 98195-2420 (rjl@amath.washington.edu).
‡Department of Applied and Computational Mathematics, California Institute of Technology, 1200 E. California Blvd., MC 217-50, Pasadena, CA 91125 (dyong@caltech.edu).

More interesting behavior is observed when the constitutive relations are nonlinear in each layer. In this case a long-wavelength pulse breaks up into a series of solitary waves that are each only a few layers wide. This is not a complete surprise since we expect that a nonlinear wave equation with a dispersive term (again arising from the layering) may give rise to soliton-like solutions, and the classic soliton equations such as the KdV equation also exhibit this type of behavior. However, the waves appearing in the layered medium are harder to characterize than classical solitary waves. The wave shape is constantly modulating as it passes through the layers, and thus it cannot be expressed in the form of a fixed-shape wave propagating at constant speed. These waves appear to interact as solitons, essentially passing through one another with at most a shift in phase, but it is not yet clear to what extent they are truly solitons in the technical sense.

To the best of our knowledge, these waves were first observed computationally in [4], where a high-resolution finite volume method is presented that calculates accurate approximations to these waves. This method is also described in [1]. Here we use this method to further explore the nature of these waves.

Solitary waves in nonlinear elastic bars have been observed and studied in the past; see, for example, [5] and the references therein. In this case the material is homogeneous but the finite cross-sectional area gives rise to reflections at the traction-free boundaries and hence dispersion.

In section 4 we show that a particular choice of the layered medium can be directly related to the Toda lattice. This may be significant since the Toda lattice is completely integrable and has discrete soliton solutions that can be compared directly to the solitary waves we observe in the corresponding layered medium.

In section 5 we present a homogenized set of equations for the nonlinear layered medium, and show that solutions to this system agree well with solutions to the layered medium equations. These equations contain dispersive terms and more complicated nonlinearities.

2. Waves in linear media. For a homogeneous linear medium with constant material parameters ρ and K, the governing equations (1.1) are simply

$$\epsilon_t - u_x = 0, \tag{2.1a}$$

$$\rho u_t - K\epsilon_x = 0 \tag{2.1b}$$

or $q_t + Aq_x = 0$, where

$$q = \begin{bmatrix} \epsilon \\ \rho u \end{bmatrix}, \qquad A = -\begin{bmatrix} 0 & 1/\rho \\ K & 0 \end{bmatrix}. \tag{2.2}$$

The eigenvalues of A are $\lambda^1 = -c$ and $\lambda^2 = +c$, where $c = \sqrt{K/\rho}$ is the speed of sound in the material. Purely leftgoing or rightgoing waves have $q(x,t)$ proportional to the corresponding eigenvector r^1 or r^2, respectively, given by

$$r^1 = \begin{bmatrix} 1 \\ Z \end{bmatrix} \quad \text{and} \quad r^2 = \begin{bmatrix} 1 \\ -Z \end{bmatrix}.$$

Here $Z = \rho c = \sqrt{K\rho}$ is the *impedance* of the material.

In a layered linear medium, wave propagation can be more complicated. If the layers are impedance-matched, $Z_A = Z_B$, then a rightgoing wave (for example) has the same characteristic form in both layers, and the wave simply has a different speed

in each layer. It will distort as it speeds up and slows down, but remains entirely rightgoing and moves with an effective velocity

$$\bar{c} = \left(\frac{\alpha}{c_A} + \frac{(1-\alpha)}{c_B} \right)^{-1}, \tag{2.3}$$

as is easily verified by computing the time required to cross two adjacent layers.

If $Z_A \neq Z_B$, on the other hand, then waves are partially reflected at each interface. It is impossible to have a purely rightgoing wave in such a composite material. However, a wave with long wavelength relative to the scale of the layers (i.e., wavelength $\gg \delta$) can appear to be essentially rightgoing, and propagates at an effective velocity of

$$\bar{c} = \sqrt{\hat{K}/\bar{\rho}}, \tag{2.4}$$

where

$$\bar{\rho} = \langle \rho \rangle = \alpha \rho_A + (1-\alpha)\rho_B \tag{2.5}$$

is the average density and

$$\hat{K} = \left\langle K^{-1} \right\rangle^{-1} = \left(\frac{\alpha}{K_A} + \frac{(1-\alpha)}{K_B} \right)^{-1} \tag{2.6}$$

is the harmonic average of the bulk moduli. The velocity (2.4) reduces to (2.3) if $Z_A = Z_B$, but more generally (2.3) does not hold. In particular, if $c_A = c_B \equiv c$ but $Z_A \neq Z_B$, then $\bar{c} < c$, so that even though all waves propagate with speed c, the effective velocity observed will be smaller. This is because the wave is constantly reflecting at each interface, and thus the energy propagates more slowly than the local wave speed.

The layered linear medium can be modeled by a homogenized equation with a dispersion relation of the form

$$\omega = \bar{c}\xi + d\xi^3 + \cdots, \tag{2.7}$$

where ξ is the spatial wavenumber and ω the temporal frequency. The effective speed $\bar{c}$ and the dispersion coefficient d were derived by Santosa and Symes [6] using Bloch wave expansions. These can also be determined from the more general homogenized equations derived in section 5 for the nonlinear case.

Figures 1 and 2 show a comparison of waves propagating in a homogeneous medium and two different layered media. In each case the initial data are $q(x,0) \equiv 0$ for $x \geq 0$, and a wave is generated by motion of the left boundary,

$$u(0,t) = \begin{cases} \bar{u}(1 + \cos(\pi(t-10)/10)) & \text{if } 0 \leq t \leq 20, \\ 0 & \text{if } t > 20. \end{cases} \tag{2.8}$$

The left edge is pulled outward for $0 < t < 20$, generating a strain wave that propagates to the right. Since the equations are linear, the magnitude of the disturbance scales out and we take $\bar{u} = 1$, but for the nonlinear case the magnitude will be important.

In the figures, the solution is shown at time $t = 40$, and four different quantities are displayed in each case: the strain $\epsilon(x,t)$, the corresponding stress $\sigma(\epsilon(x,t),x)$, the

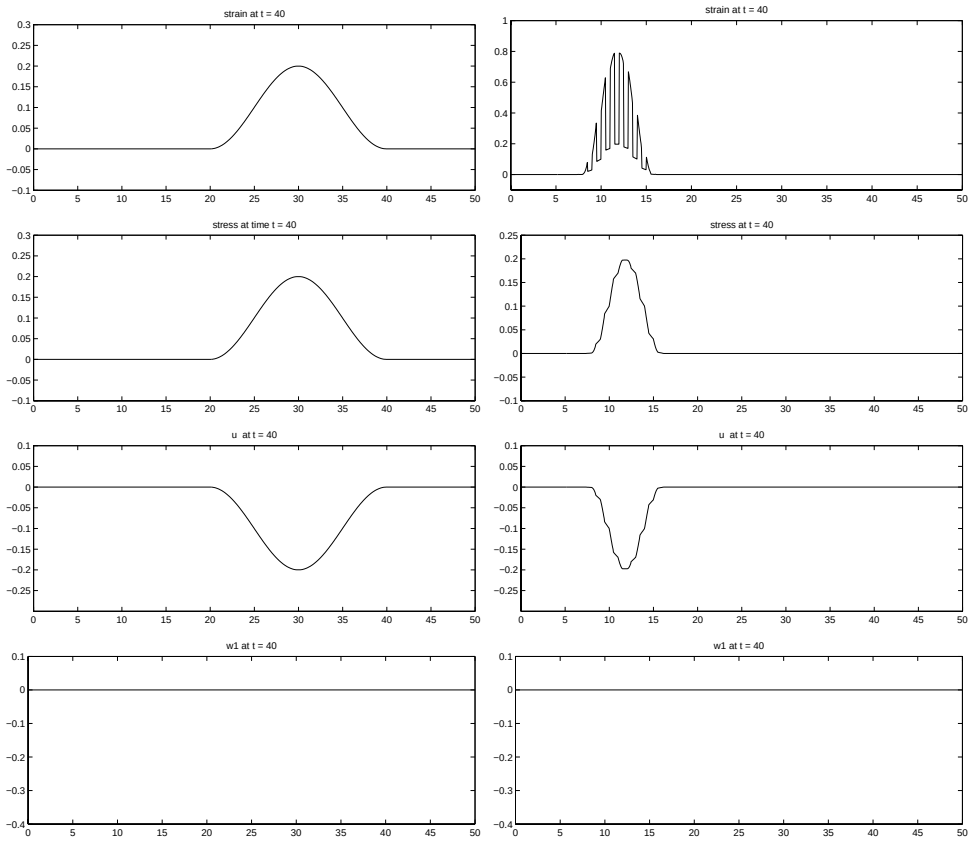


FIG. 1. *The left column shows a strain wave propagating in a homogeneous medium with wave speed $c = 1$. The right column shows a strain wave propagating in a layered medium with constant impedance. In each case four quantities are shown: strain, stress, velocity, and the characteristic variable $w^1(x,t)$, at time $t = 40$.*

velocity $u(x,t)$, and $w^1(x,t)$, where the characteristic variables w^1 and w^2 are defined by

$$w^1(x,t) = \frac{1}{2Z(x)}(Z(x)\epsilon(x,t) + \rho(x)u(x,t)), \tag{2.9a}$$

$$w^2(x,t) = \frac{1}{2Z(x)}(Z(x)\epsilon(x,t) - \rho(x)u(x,t)). \tag{2.9b}$$

These satisfy $w(x,t) = R^{-1}(x)q(x,t)$, where $R(x)$ is the matrix of right eigenvectors of the coefficient matrix $A(x)$ given in (2.2):

$$R(x) = \begin{bmatrix} 1 & 1 \\ Z(x) & -Z(x) \end{bmatrix}, \qquad R^{-1}(x) = \frac{1}{2Z(x)} \begin{bmatrix} 1 & Z(x) \\ 1 & -Z(x) \end{bmatrix}. \tag{2.10}$$

The quantities w^1 and w^2 give the magnitude of leftgoing and rightgoing waves, respectively.

The left column of Figure 1 shows results for a homogeneous medium with $\rho \equiv 1$, $K \equiv 1$. The wave moves at velocity $c = 1$ and so lies between $x = 20$ and $x = 40$ at

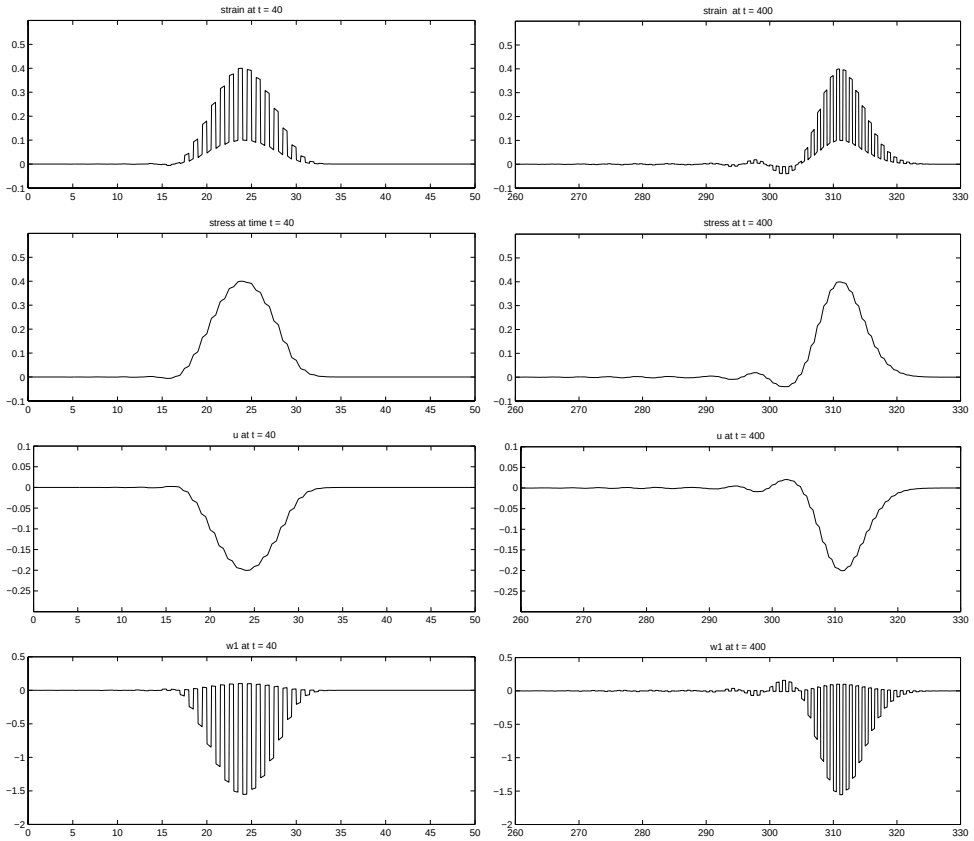


FIG. 2. *A strain wave propagating in a layered medium with impedance mismatch. The left column shows the solution at time $t = 40$, the right column at time $t = 400$. In each case four quantities are shown: strain, stress, velocity, and the characteristic variable $w^1(x,t)$.*

time $t = 40$, with peak at $x = 30$. Note that $w^1(x,t) \equiv 0$ since the wave is purely rightgoing.

The right column of Figure 1 shows a layered medium in which $\delta = 1$ and

$$\begin{array}{lllll} \delta_A = 0.5, & \rho_A = 4, & K_A = 0.25, & c_A = 0.25, & Z_A = 1, \\ \delta_B = 0.5, & \rho_B = 1, & K_B = 1, & c_B = 1, & Z_B = 1. \end{array} \tag{2.11}$$

In this case $Z_A = Z_B$, and so again the wave is purely rightgoing ($w^1(x,t) \equiv 0$) and propagates with velocity $\bar{c} = 2/5$ from (2.3). The peak of the disturbance is now observed at $x = 30\bar{c} = 12$. Note that in this case the strain $\epsilon(x,t)$ is discontinuous at each layer interface. The B layers are more easily stretched than the A layers since $K_B > K_A$. The stress $\sigma(\epsilon(x,t),x)$ must be continuous, however, since the force acting on each side of the interface must be equal. This condition can be used to find the jump conditions on ϵ. Similarly, the velocity $u(x,t)$ must be continuous everywhere, but the momentum $\rho(x)u(x,t)$ will be discontinuous at the interfaces where $\rho(x)$ has a jump discontinuity.

The left column of Figure 2 shows a layered medium in which

$$\begin{array}{lllll} \delta_A = 0.5, & \rho_A = 4, & K_A = 4, & c_A = 1, & Z_A = 4, \\ \delta_B = 0.5, & \rho_B = 1, & K_B = 1, & c_B = 1, & Z_B = 1. \end{array} \tag{2.12}$$

In this case $Z_A \neq Z_B$, and we have

$$\bar{\rho} = \frac{5}{2}, \qquad \hat{K} = \frac{8}{5}, \qquad \bar{c} = \sqrt{\frac{\hat{K}}{\bar{\rho}}} = \frac{4}{5}. \tag{2.13}$$

The peak of the disturbance is now located at $x = 30\bar{c} = 24$. Note that in this case $w^1(x,t)$ is not identically zero, showing that the wave has a significant leftgoing component, although the envelope of w^1 propagates to the right at the effective velocity $\bar{c}$.

The right column of Figure 2 shows this same wave at a much later time, $t = 400$, at which point the dispersive effect of the layered medium is apparent. The leading edge of the wave is still at approximately $\bar{c}t = 320$, as expected, but trailing oscillations have appeared behind the wave due to the dispersion.

3. Waves in nonlinear media. We now consider the case of a nonlinear stress-strain relation $\sigma(\epsilon, x)$. In particular, we consider the exponential relation

$$\sigma(\epsilon, x) = e^{K(x)\epsilon} - 1, \tag{3.1}$$

which will be related to the Toda lattice in section 4, and the simpler quadratic relation

$$\sigma(\epsilon, x) = K(x)\epsilon + \beta K^2(x)\epsilon^2, \tag{3.2}$$

which approximates the exponential relation if $\beta = 1/2$, and reduces to the linear case if $\beta = 0$. For both of these nonlinear constitutive relations $\sigma_\epsilon(\epsilon, x) = K(x) + O(\epsilon) \to K(x)$ as $\epsilon \to 0$, and so for very small amplitude waves the linear theory of the last section applies with bulk modulus $K(x)$.

The system of conservation laws (1.1) can now be written as

$$q_t + f(q, x)_x = 0 \tag{3.3}$$

or, for smooth solutions, as

$$q_t + f_q(q, x)q_x = -f_x(q, x). \tag{3.4}$$

Note that we distinguish between $f(q, x)_x$, the total derivative of $f(q(x,t), x)$ with respect to x, and $f_x(q, x)$, the partial derivative of $f(q, x)$ with respect to the second variable. The Jacobian matrix for the system (3.4) is

$$f_q(q, x) = \begin{bmatrix} 0 & 1/\rho(x) \\ \sigma_\epsilon(\epsilon, x) & 0 \end{bmatrix}, \tag{3.5}$$

with eigenvalues $\lambda^1 = -c$ and $\lambda^2 = +c$, where the sound speed c now depends on the strain as well as x,

$$c(\epsilon, x) = \sqrt{\frac{\sigma_\epsilon(\epsilon, x)}{\rho(x)}}.$$

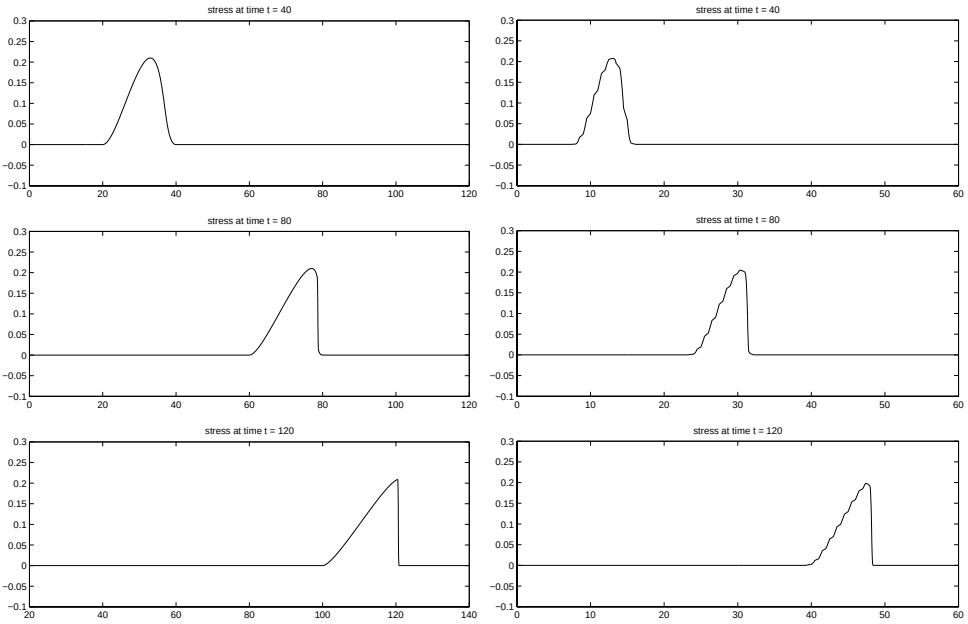


FIG. 3. *The left column shows a wave propagating in a homogeneous nonlinear medium. Only the stress is shown at three different times, illustrating shock formation. The right column shows propagation in a layered nonlinear medium with constant linearized impedance.*

The corresponding right eigenvectors of f_q are

$$r^1 = \begin{bmatrix} 1 \\ Z(\epsilon, x) \end{bmatrix}, \qquad r^2 = \begin{bmatrix} 1 \\ -Z(\epsilon, x) \end{bmatrix}, \tag{3.6}$$

where the impedance $Z(\epsilon, x) = \sqrt{\rho(x)\sigma_\epsilon(\epsilon, x)}$ now also depends on ϵ. For either constitutive model (3.1) or (3.2) we have $c(\epsilon, x) \to \sqrt{K(x)/\rho(x)}$ and $Z(\epsilon, x) \to \sqrt{\rho(x)K(x)}$ as $\epsilon \to 0$, which are the linearized sound speed and impedance.

Figures 3 and 4 show computed results for three cases analogous to those shown for the linear problem in the previous section, but now using the exponential relation (3.1) and $\bar{u} = 0.2$ in the boundary data (2.8). Here we display only one quantity, the stress $\sigma(\epsilon(x, t), x)$, but plot the solution at several different times to show the evolution.

The left column of Figure 3 shows nonlinear propagation in a homogeneous medium, with $\rho \equiv 1$, $K \equiv 1$. The cosine-shaped wave generated by the boundary condition (2.8) steepens into a shock followed by a rarefaction wave, as expected from standard nonlinear conservation law theory.

The right column of Figure 3 shows a layered medium with

$$\begin{aligned} \delta_A &= 0.5, \qquad \rho_A = 4, \qquad K_A = 0.25, \\ \delta_B &= 0.5, \qquad \rho_B = 1, \qquad K_B = 1. \end{aligned} \tag{3.7}$$

In this case the linearized impedances are matched since $\rho_A K_A = \rho_B K_B$, and again the rightgoing wave appears to steepen into a shock followed by a rarefaction wave. However, the full nonlinear impedance does not remain matched, and some reflection occurs at interfaces. This shock wave is not a sharp discontinuity but remains smeared

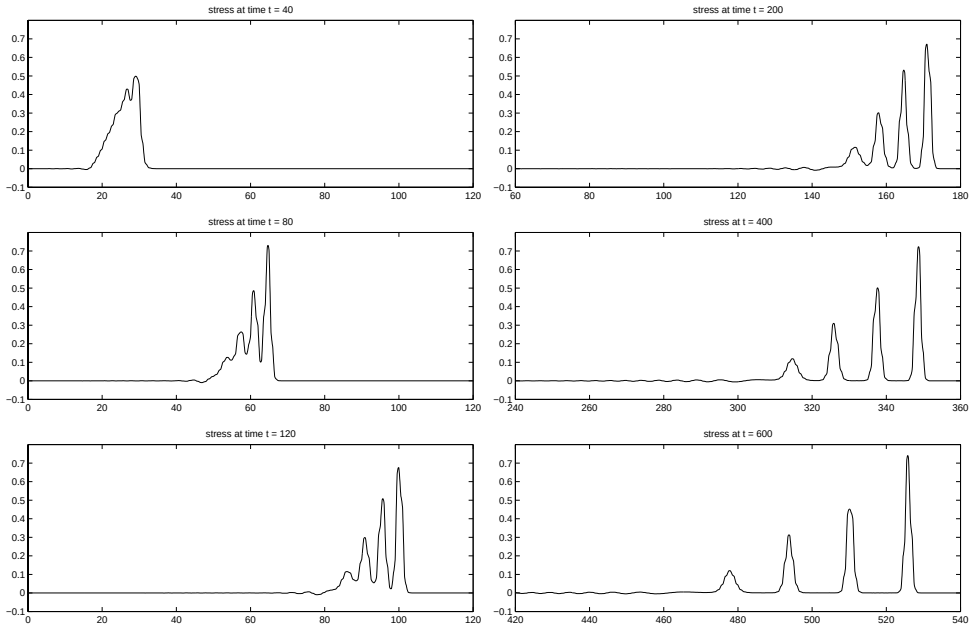


FIG. 4. *Propagation in a layered nonlinear medium with impedance mismatch at the interfaces. The left column shows the stress at three times, and the right column at three later times, illustrating breakup into a train of solitary waves.*

over a few layers. In this case the layering gives rise to an effective viscous equation. (In this case C_{13} is nonzero in (5.17) below, while the other C coefficients are all zero.)

Figure 4 shows a more interesting case that is the primary object of study in this paper. Here the material parameters are

$$\begin{aligned} \delta_A &= 0.5, \qquad \rho_A = 4, \qquad K_A = 4, \\ \delta_B &= 0.5, \qquad \rho_B = 1, \qquad K_B = 1. \end{aligned} \tag{3.8}$$

Now the linearized impedance is not matched, and large-scale reflections at each interface lead to dispersive behavior. This dispersion, coupled with the nonlinearity, results in the existence of solitary waves. In Figure 4 the same boundary motion (2.8) is applied as before. The resulting pulse, which is long compared to the layer width, initially starts to steepen as if a shock were forming. But then oscillations develop and the pulse ultimately breaks up into a train of solitary waves. Similar behavior is seen with nonlinear dispersive equations such as the KdV equation that are known to have soliton solutions.

Figure 5 shows both the stress and the strain in the first two solitary waves at time $t = 600$. Since it is not clear whether these solitary waves are formally solitons, we will refer to them as *stegotons* for shorthand, coming from the Greek root "stego-," meaning roof or ridge, and suggested by the rough resemblance of these strain waves in a layered medium to the back of a stegosaurus.

Note that each stegoton has a width of about ten layers, a fact that is independent of the periodicity δ used. If δ is made smaller, then the stegotons scale with δ but remain about ten layers wide. This is expected from the form of the equations. If $\tilde{q}(x,t)$ is a solution to (3.3) for $\delta = 1$, then $q(x,t) = \tilde{q}(x/\delta, t/\delta)$ is a solution for arbitrary δ (with α fixed). The width in layers does vary slightly with amplitude, as

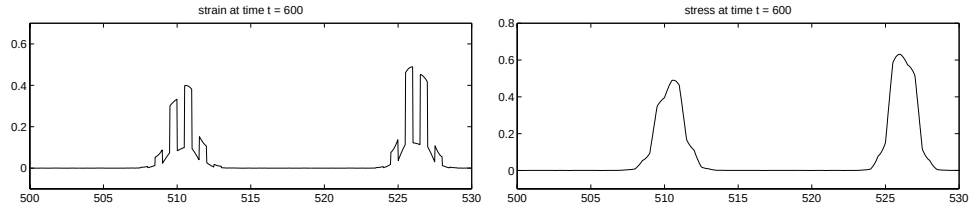


FIG. 5. *Close-up view of the first two solitary waves from Figure* 4 *at time* $t = 600$*, showing both the strain and the stress.*

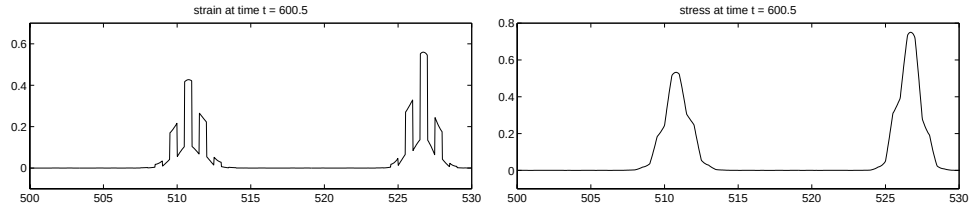


FIG. 6. *Close-up view of the two solitary waves from Figure* 5 *at a slightly later time* $t = 600.5$.

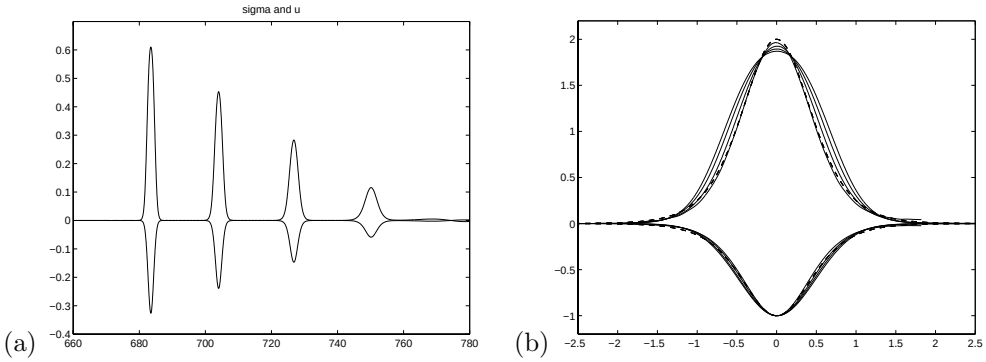


FIG. 7. (a) *A time trace of the stress* $\sigma(x_0, t) \geq 0$ *and velocity* $u(x_0, t) \leq 0$ *at a fixed point* x_0 *as four stegotons pass by.* (b) *The same four waves replotted as functions of* τ *with the scaling described in the text. The dashed lines are* $-\operatorname{sech}^2(b\tau)$ *and* $2\operatorname{sech}^2(b\tau)$.

does the speed of the solitary wave. The taller stegotons are thinner and travel faster than shorter ones, similar to the behavior of KdV solitons, for example.

Stegotons do not translate with a fixed wave shape, as illustrated in Figure 6, where the two stegotons from Figure 5 are shown at a slightly later time, $t = 600.5$. For this reason it is difficult to carefully investigate the scaling properties of stegotons by studying their appearance as functions of x for fixed t. On the other hand, if we pick a physical location x and observe the solution as a function of t as a stegoton passes by, then all components of the solution vary smoothly in time. By observing the waves in this manner it is easy to determine the scaling relation between amplitude, speed, and width. Figure 7(a) shows the stress $\sigma(x_0, t)$ as a function of t at $x_0 = 600.25$, a location that is in the center of an A-layer. This shows four distinct solitary waves passing by, followed by some trailing noise. Here we have also plotted the velocity $u(x_0, t)$, which is negative in each wave.

Figure 7(b) shows plots of the four leading solitary waves from Figure 7(a) after

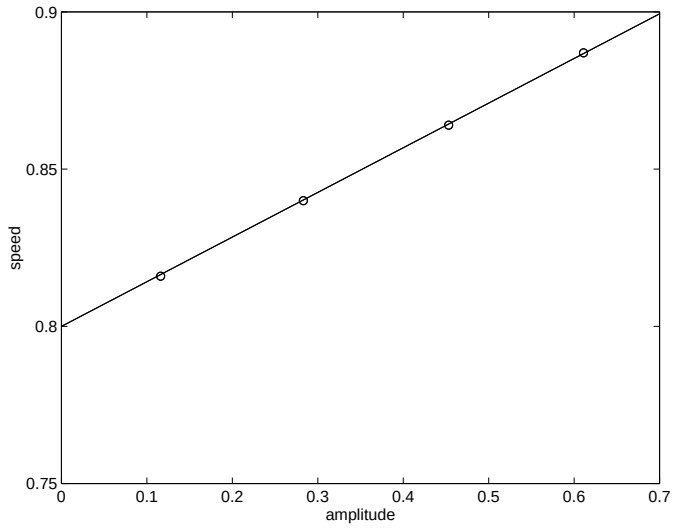


FIG. 8. *The observed speed of a solitary wave plotted as a function of amplitude, for the four waves shown in Figure* 7.

shifting them to a common location, rescaling each by its amplitude a as measured from the velocity, $a = \max|u|$, and rescaling the width by $\sqrt{a}$. Hence we plot

$$\frac{1}{a}\sigma(x_0,\tau) \qquad \text{and} \qquad \frac{1}{a}u(x_0,\tau) \tag{3.9}$$

as functions of τ, where $\tau = \sqrt{a}(t-t_m)$ and t_m is the time at which the velocity reaches its peak value $-a$. The velocity plots lie nearly on top of one another, suggesting that at $x = x_0$ a stegoton of amplitude a has velocity of the form

$$u(x_0,t) = a\,U\left(\sqrt{a}(t-t_m)\right) \tag{3.10}$$

for some function $U(\tau)$. The dashed lines in Figure 7(b) show the functions $-\operatorname{sech}^2(b\tau)$ and $2\operatorname{sech}^2(b\tau)$ for $b = 1.7$. This shows that the stegoton has roughly, though not exactly, the sech^2 shape seen for many solitons.

The stress does not scale quite as nicely as the velocity, and in particular the stress is not simply a scalar multiple of the velocity. It is nearly so, however, for the stress curve that appears tallest in Figure 7(b). This actually corresponds to the shortest stegoton in the original time trace of Figure 7(a). For this small amplitude wave we observe $\sigma \approx -2u$. This is consistent with the fact that a linearized homogeneous medium with impedance Z would have $\sigma = -Zu$, and the layered medium we are using has an effective impedance in the linearized case that is $(\hat{K}\bar{\rho})^{1/2} = 2$.

By observing when the peak appears in the time history at different points x (each in the center of an A-layer), we can estimate the velocity of each wave. The observed velocity is plotted against the amplitude in Figure 8 for each of these four stegotons. They lie almost exactly on the line

$$v = 0.8 + 0.142a. \tag{3.11}$$

Recall that $\bar{c} = 0.8$ is the effective velocity of the linearized medium from (2.13), which is the velocity we expect to observe for very small amplitude waves. The velocity appears to increase linearly with amplitude for nonlinear waves.

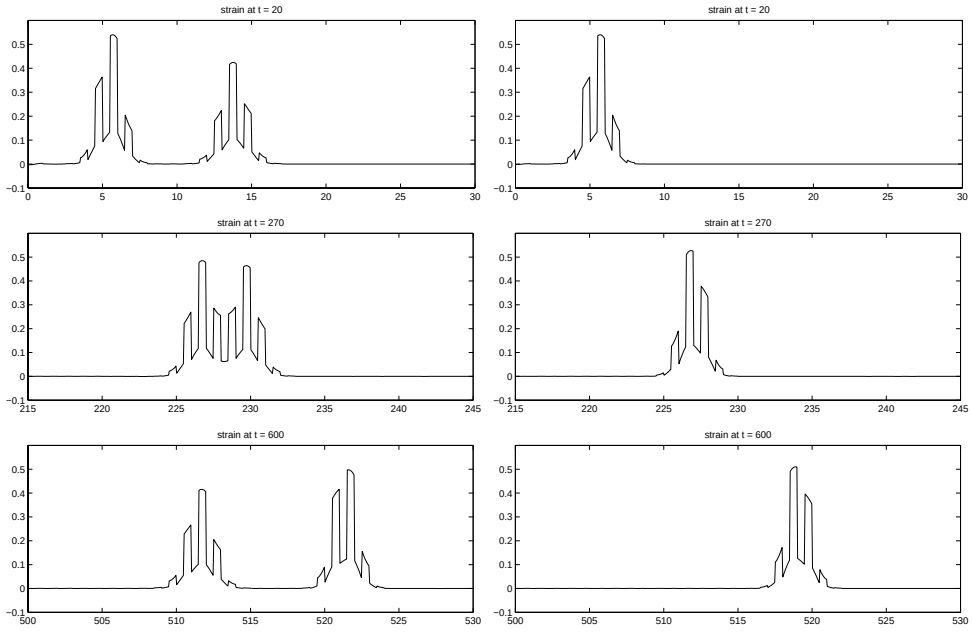


FIG. 9. *The collision of two stegotons at three different times is shown in the left column. For comparison, the right column shows results at the same times for a single stegoton.*

These results are for the case in which x_0 is in the center of an A-layer. If we fix x_0 at a different spot, then the scaling is the same but the amplitude a and wave shape $U(\tau)$ are different. Indeed, the amplitude varies significantly at different points, even within the same layer. A more complete characterization of these waves is still under development.

By recording $u(x_0, t)$ at a location x_0 that is at the left edge of an A-layer, and then selecting the part of this time history that captures a single wave passing by, we obtain data that can be used to replace (2.8) as boundary data for generating a single stegoton. By rescaling the amplitude and width of this data appropriately using the scaling determined above, we can also generate stegotons of arbitrary amplitude a. We have verified that these also propagate as solitary waves, at least if the amplitude is not too large.

We can also generate a short stegoton at the boundary followed by a taller stegoton that travels faster and eventually overtakes the first. The left column of Figure 9 shows the results of this experiment. We observe that the two waves interact in a manner analogous to classical solitons: the waves appear to exchange identity, and the wave in front grows and accelerates. After the waves separate, each again has the form of a solitary stegoton, though shifted in location from where they would be without interaction. For comparison, the right column of Figure 9 shows the propagation of the larger stegoton alone, without the presence of the smaller one.

4. Relation to the Toda lattice. The Toda lattice is a discrete lattice of particles having mass m connected by nonlinear springs with a restoring force that depends exponentially on the distance stretched. Let $X_j(t)$ be the location of the jth particle at time t, and assume that the unstretched configuration has $X_j = j\Delta x$. The velocity of this particle is denoted by $U_j(t)$. The spring connecting particle j to $j+1$

has strain

$$\epsilon_{j+1/2}(t) = \frac{X_{j+1}(t) - X_j(t)}{\Delta x} - 1 \tag{4.1}$$

and exerts a restoring force $\sigma(\epsilon_{j+1/2}(t))$, where

$$\sigma(\epsilon) = e^{K\epsilon} - 1.$$

Since $X_j'(t) = U_j(t)$, differentiating (4.1) yields

$$\epsilon'_{j+1/2}(t) = \frac{U_{j+1}(t) - U_j(t)}{\Delta x}. \tag{4.2}$$

The Toda lattice is modeled by a system of ODEs consisting of this equation along with the dynamic equation

$$mU_j'(t) = \sigma(\epsilon_{j+1/2}(t)) - \sigma(\epsilon_{j-1/2}(t)). \tag{4.3}$$

If we rewrite $m = \rho\Delta x$, then the system of equations becomes

$$\begin{aligned} \epsilon'_{j+1/2}(t) &= \frac{U_{j+1}(t) - U_j(t)}{\Delta x}, \\ \rho U_j'(t) &= \frac{\sigma(\epsilon_{j+1/2}(t)) - \sigma(\epsilon_{j-1/2}(t))}{\Delta x}. \end{aligned} \tag{4.4}$$

Note that this can be viewed as a finite-difference discretization of the elastic equation (1.1). Centered differences on a staggered grid are used to approximate each x-derivative in (1.1). The classical theory of finite-difference methods thus leads us to expect dispersive behavior, and a "modified equation" analysis of this system would show that the discrete equations can be approximated by a dispersive nonlinear system of PDEs. With this combination of nonlinearity and dispersion, solitary waves can arise. Toda showed that with exponential springs the discrete system is completely integrable and discrete soliton solutions exist; see [7], [8].

To relate the Toda lattice to a layered medium, it may be tempting to introduce two different types of springs with constitutive relations $\sigma_A(\epsilon)$ and $\sigma_B(\epsilon)$ and have these alternate in the discrete lattice. However, this would introduce a second form of dispersion and result in a doubly dispersive system.

Since the original Toda lattice already has soliton solutions, we wish to relate the layered medium directly to the lattice. This is easily done by realizing that the Toda lattice does in fact consist of alternating layers, since particles alternate with springs. The corresponding layered medium is one in which thin "particle layers" with finite mass and infinitesimal compressibility alternate with thicker "spring layers" having infinitesimal mass and finite compressibility, i.e.,

$$\begin{aligned} &\delta_A \ll 1, && \rho_A = O\left(\frac{1}{\delta_A}\right), && K_A \gg 1, \\ &\delta_B = 1 - \delta_A, && \rho_B \ll 1, && K_B = O(1). \end{aligned} \tag{4.5}$$

See Figure 10 for an illustration of this correspondence. (In fact, one observes close correspondence even when δ_A is not small relative to δ_B.)

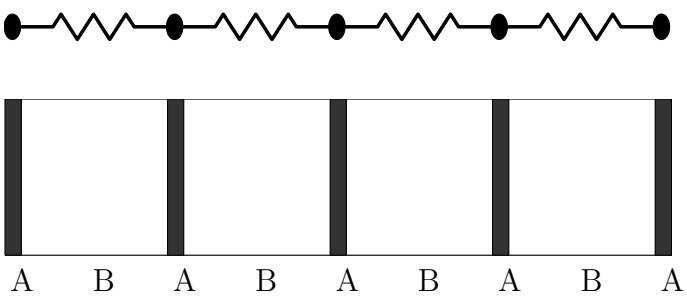


FIG. 10. *The Toda lattice and a roughly equivalent layered medium.*

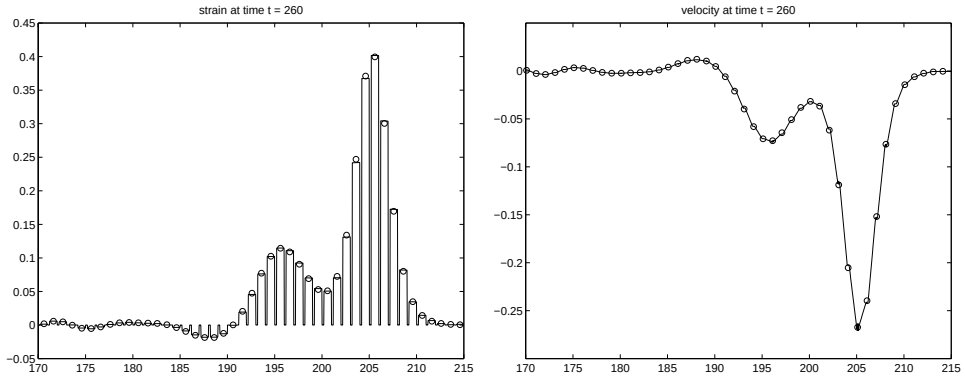


FIG. 11. *Comparison of the solution to the discrete Toda lattice (circles) with the finite-volume solution to the layered medium equations (solid line). The strain and velocity are shown.*

Figure 11 shows a sample calculation obtained by solving the layered media equations using

$$\begin{array}{lll} \delta_A = 0.2, & \rho_A = 5, & K_A = 2000, \\ \delta_B = 0.8, & \rho_B = 0.005, & K_B = 0.5. \end{array} \tag{4.6}$$

In the particle layers, which exhibit little strain, the constitutive relation is taken to be linear with $\sigma_A(\epsilon) = K_A\epsilon$. The wave speed is then constant in these layers, $c_A = 20$, which is convenient numerically since by taking $\Delta t = \Delta x/20$ (so the Courant number is exactly 1) an accurate solution is computed even when there are very few points in these thin layers. The spring layers have the exponential relation $\sigma_B(\epsilon) = e^{K_B\epsilon} - 1$, with a wave speed that approaches $c_B = \sqrt{K_B/\rho_B} = 10$ for small ϵ and remains below 20 so that the calculations are stable.

The parameters (4.6) lead to an effective wave speed for the linearized response (small ϵ) of

$$\bar{c} = \sqrt{\frac{\hat{K}}{\bar{\rho}}} = \sqrt{\frac{0.625}{1.004}} \approx 0.789, \tag{4.7}$$

and solitary waves move with a speed that is less than 1. To have a close connection with the Toda lattice, it is important that the wave speeds c_A and c_B are much larger than this homogenized wave speed. This means that small amplitude waves bounce

back and forth within each layer on a much faster time scale than the observed wave motion, leading to a local equilibration within each layer. As a result, the stress observed in Figure 11 at any time is roughly constant in each thick spring layer, and varies rapidly, but essentially linearly, in the thin particle layers. The velocity u, on the other hand, is roughly constant in each particle layer, since the particle moves as a rigid unit, and varies linearly across a spring layer. Only under these conditions can we hope to model the continuum material by a set of discrete springs and masses, where each spring has a single stress associated with it and is assumed to compress and expand uniformly.

Figure 11 shows that there is in fact good agreement between the layered medium and the Toda lattice in this case. The parameters for the Toda lattice used here are $m = 1.004$ for the particle mass and $K = 0.5$ in the stress-strain relation. The mass is the average density from the layered medium, while the value of K is taken to be K_B, the corresponding parameter from the spring layers. Note that, for the layered medium solution (solid line), the strain is nearly zero in each thin particle layer and nearly constant throughout each spring layer. The circles show the discrete solution computed by solving the ODEs of the Toda lattice. These results are shown at a time when the initial pulse is just breaking up into solitary waves. At later times both solutions break up into similar trains of solitary waves.

Because of the correspondence in Figure 10, it is not surprising that the layered medium exhibits solitary waves in this special case. What is more interesting is the fact that it continues to exhibit solitary wave behavior even for situations that are far from this limit, as was exhibited in section 3. In the next section we derive homogenized equations for the general case that may help to shed some light on this.

5. Homogenized equations. In this section we derive homogenized equations that describe the effective behavior of the layered media (both linear and nonlinear) studied in the previous sections. Since the strain and momentum have discontinuities at each interface and cannot be approximated directly by continuous functions, we start by rewriting the equations in terms of the stress and velocity, which are continuous. Equation (1.1b), $\rho(x)u_t - \sigma_x = 0$, is one equation of this system. We must use (1.1a) to derive an equation for σ_t. To do so, we use

$$\sigma_t = \sigma_\epsilon(\epsilon, x)\epsilon_t = \sigma_\epsilon(\epsilon, x)u_x$$

and will assume that the constitutive relation $\sigma(\epsilon, x)$ is such that we can solve for $\sigma_\epsilon(\epsilon, x)$ as a function of σ and x in the form

$$\sigma_\epsilon(\epsilon, x) = K(x)G(\sigma). \tag{5.1}$$

In particular, for the constitutive equations used in this paper we have

$$\begin{array}{lll} \sigma(\epsilon, x) = K(x)\epsilon & \Longrightarrow & G(\sigma) = 1, \\ \sigma(\epsilon, x) = \exp(K(x)\epsilon) - 1 & \Longrightarrow & G(\sigma) = 1 + \sigma, \\ \sigma(\epsilon, x) = K(x)\epsilon + \beta K^2(x)\epsilon^2 & \Longrightarrow & G(\sigma) = 1 + 2\beta\sigma - 2\beta^2\sigma^2 + O(\sigma^3). \end{array} \tag{5.2}$$

Then the system (1.1) can be rewritten as

$$\begin{aligned} \sigma_t - K(x)G(\sigma)u_x &= 0, \\ \rho(x)u_t - \sigma_x &= 0. \end{aligned} \tag{5.3}$$

This nonlinear system is not in conservation form but is still valid since we are not interested in shock waves. The leading order terms in the homogenized equation are

easy to derive by rewriting (5.3) as

$$\begin{aligned} K^{-1}(x)\sigma_t - G(\sigma)u_x &= 0, \\ \rho(x)u_t - \sigma_x &= 0. \end{aligned} \tag{5.4}$$

The functions $K(x)$ and $\rho(x)$ vary on a much faster scale than σ and u, and thus averaging these equations over a period leads to

$$\begin{aligned} \langle K^{-1}(x)\rangle\sigma_t - G(\sigma)u_x &\approx 0, \\ \langle \rho(x)\rangle u_t - \sigma_x &\approx 0. \end{aligned} \tag{5.5}$$

This gives a homogenized system that again has the form (5.3) but with $\rho(x)$ and $K(x)$ replaced by the average and harmonic average, respectively. This gives the expected effective velocity in the linear case, but is lacking the crucial dispersive terms needed to explain the solitary waves.

To obtain a more accurate description of the homogenized equations, we use a multiple scale homogenization technique, following [3]. We begin by defining a fast spatial variable $x^* = x/\delta$, where δ is the period of the medium, as before, and writing the bulk modulus and density as functions of x^*: $K = K(x^*)$ and $\rho = \rho(x^*)$. (This is a slight abuse of notation.) We now adopt the formalism that $\bar{x} = x$ and x^* are independent variables by assuming $\delta \ll 1$. As a result, spatial derivatives in (5.3) must be transformed according to

$$\frac{\partial}{\partial x} = \frac{\partial}{\partial \bar{x}} + \frac{1}{\delta}\frac{\partial}{\partial x^*}.$$

(We introduce $x = \bar{x}$ in this step to clearly delineate the original spatial scale x and the new multiple scales $\bar{x}$ and x^*, but we will simply use x from now on.) System (5.3) becomes

$$\begin{aligned} \sigma_t - K(x^*)G(\sigma)\left[u_x + \delta^{-1}u_{x^*}\right] &= 0, \\ \rho(x^*)u_t - \left[\sigma_x + \delta^{-1}\sigma_{x^*}\right] &= 0. \end{aligned} \tag{5.6}$$

Using the convention that all underlined quantities are independent of the fast variable x^*, we insert the asymptotic expansions

$$\begin{aligned} u(x,t;\delta) &= \underline{u}^{(0)}(x,t) + \delta u^{(1)}(x,x^*,t) + \delta^2 u^{(2)}(x,x^*,t) + O(\delta^3), \\ \sigma(x,t;\delta) &= \underline{\sigma}^{(0)}(x,t) + \delta\sigma^{(1)}(x,x^*,t) + \delta^2\sigma^{(2)}(x,x^*,t) + O(\delta^3) \end{aligned}$$

into (5.6) and collect terms by their powers of δ. During this process, the function $G(\sigma)$ must be expanded as

$$\begin{aligned} G(\sigma) &= G(\underline{\sigma}^{(0)} + \delta\sigma^{(1)} + \delta^2\sigma^{(2)} + \cdots) \\ &= G(\underline{\sigma}^{(0)}) + G'(\underline{\sigma}^{(0)})(\delta\sigma^{(1)} + \delta^2\sigma^{(2)} + \cdots) \\ &\qquad + \frac{1}{2}G''(\underline{\sigma}^{(0)})(\delta\sigma^{(1)} + \delta^2\sigma^{(2)} + \cdots)^2 + \cdots. \end{aligned}$$

Once the proper choices for scales and asymptotic expansions have been made, obtaining homogenized equations becomes a purely mechanical, although algebraically intensive, procedure. For the system of equations that is proportional to δ^n, we solve for $u_{x^*}^{(n+1)}$ and $\sigma_{x^*}^{(n+1)}$ and identify any terms that are independent of x^*. These terms

must be set to zero; otherwise, when $u^{(n+1)}_{x^*}$ and $\sigma^{(n+1)}_{x^*}$ are integrated with respect to x^*, secular terms will arise. The terms that are set to zero give rise to the homogenized equations.

To identify terms that are independent of x^*, we introduce the following linear operators:

$$\langle a(x^*)\rangle = \int_0^1 a(x^*)dx^*,$$
$$\{a\}(x^*) = a(x^*) - \langle a(x^*)\rangle,$$
$$[\![a]\!](x^*) = \int_s^{x^*} \{a\}(\xi)d\xi, \quad \text{where } s \text{ is chosen such that } \langle [\![a]\!](x^*)\rangle = 0.$$

When using the normal spatial scale, δ is the period of the medium, but when using the fast spatial scale $x^* = x/\delta$, the period of the functions $K(x^*)$ and $\rho(x^*)$ is 1. Therefore, the averaging operator $\langle K(x^*)\rangle$, defined above, gives the average value of $K(x^*)$. (A similar averaging operator can be defined for nonperiodic functions.) The $\{\cdot\}$ operator generates the fluctuating part of a function: the part of the function that has zero average. The $[\![\cdot]\!]$ operator gives the integral of the fluctuating part of a function, where the constant of integration is chosen such that the average of the integral of the fluctuating part is zero. As opposed to $\langle\cdot\rangle$, both $\{\cdot\}$ and $[\![\cdot]\!]$ return functions of x^*. Some useful properties of these operators are derived in [11].

In the case of piecewise constant functions (1.2), $\langle\rho(x^*)\rangle = \bar\rho = \alpha\rho_A + (1-\alpha)\rho_B$,

$$\{\rho\}(x^*) = \begin{cases} (1-\alpha)(\rho_A - \rho_B) & \text{if } j < x^* < (j+\alpha) \text{ for some integer } j, \\ \alpha(\rho_B - \rho_A) & \text{otherwise,} \end{cases}$$

and

$$[\![\rho]\!](x^*) = \begin{cases} (1-\alpha)(\rho_A - \rho_B)\left(x^* - \frac{\alpha}{2}\right) & \text{if } j < x^* < (j+\alpha) \text{ for some integer } j, \\ \alpha(\rho_B - \rho_A)\left(x^* - \frac{1+\alpha}{2}\right) & \text{otherwise.} \end{cases}$$

Now we illustrate how the first few terms of the homogenized equations are derived. The leading order equations are

$$\underline{\sigma}^{(0)}_t - K(x^*)G(\underline{\sigma}^{(0)})(\underline{u}^{(0)}_x + u^{(1)}_{x^*}) = 0, \tag{5.7a}$$
$$\rho(x^*)\underline{u}^{(0)}_t - \underline{\sigma}^{(0)}_x - \sigma^{(1)}_{x^*} = 0. \tag{5.7b}$$

We solve for $u^{(1)}_{x^*}$ and $\sigma^{(1)}_{x^*}$ to obtain

$$u^{(1)}_{x^*} = \frac{\underline{\sigma}^{(0)}_t}{K(x^*)G(\underline{\sigma}^{(0)})} - \underline{u}^{(0)}_x, \tag{5.8a}$$
$$\sigma^{(1)}_{x^*} = \rho(x^*)\underline{u}^{(0)}_t - \underline{\sigma}^{(0)}_x. \tag{5.8b}$$

Before integrating with respect to x^*, we must first remove x^*-independent terms from the right-hand sides of (5.8). To do this, we apply $\langle\cdot\rangle$ to the right-hand sides of (5.8) and set the result to zero:

$$0 = \left\langle K^{-1}\right\rangle \underline{\sigma}^{(0)}_t - G(\underline{\sigma}^{(0)})\underline{u}^{(0)}_x, \tag{5.9a}$$
$$0 = \langle\rho\rangle\, \underline{u}^{(0)}_t - \underline{\sigma}^{(0)}_x. \tag{5.9b}$$

We can then integrate the remaining terms in (5.8) to get

$$u^{(1)} = \left[\!\left[K^{-1}\right]\!\right] \frac{\underline{\sigma}_t^{(0)}}{G(\underline{\sigma}^{(0)})} + \underline{u}^{(1)}, \tag{5.10a}$$

$$\sigma^{(1)} = \left[\!\left[\rho\right]\!\right] \underline{u}_t^{(0)} + \underline{\sigma}^{(1)}, \tag{5.10b}$$

where $\underline{u}^{(1)}$ and $\underline{\sigma}^{(1)}$ are "constants" of integration in terms of x^* but vary with x and t.

The $O(\delta)$ equations are

$$\sigma_t^{(1)} - K(x^*)G(\underline{\sigma}^{(0)})(u_x^{(1)} + u_{x^*}^{(2)}) - K(x^*)G'(\underline{\sigma}^{(0)})\sigma^{(1)}(\underline{u}_x^{(0)} + u_{x^*}^{(1)}) = 0, \tag{5.11a}$$

$$\rho(x^*)u_t^{(1)} - \sigma_x^{(1)} - \sigma_{x^*}^{(2)} = 0. \tag{5.11b}$$

We substitute in (5.7a) and (5.10) and solve for $u_{x^*}^{(2)}$ and $\sigma_{x^*}^{(2)}$ to obtain

$$\begin{aligned} u_{x^*}^{(2)} = {} & \frac{\left[\!\left[\rho\right]\!\right] \underline{u}_{tt}^{(0)} + \underline{\sigma}_t^{(1)}}{K(x^*)G(\underline{\sigma}^{(0)})} - \left[\!\left[K^{-1}\right]\!\right] \frac{G(\underline{\sigma}^{(0)})\underline{\sigma}_{xt}^{(0)} - \underline{\sigma}_t^{(0)}\underline{\sigma}_x^{(0)}G'(\underline{\sigma}^{(0)})}{G(\underline{\sigma}^{(0)})^2} - \underline{u}_x^{(1)} \\ & - \frac{G'(\underline{\sigma}^{(0)})\underline{\sigma}_t^{(0)}}{K(x^*)G(\underline{\sigma}^{(0)})^2} \left[\left[\!\left[\rho\right]\!\right] \underline{u}_t^{(0)} + \underline{\sigma}^{(1)}\right], \end{aligned} \tag{5.12a}$$

$$\sigma_{x^*}^{(2)} = \rho(x^*) \left[\left[\!\left[K^{-1}\right]\!\right] \frac{G(\underline{\sigma}^{(0)})\underline{\sigma}_{tt}^{(0)} - (\underline{\sigma}_t^{(0)})^2 G'(\underline{\sigma}^{(0)})}{G(\underline{\sigma}^{(0)})^2} + \underline{u}_t^{(1)} \right] - \left[\!\left[\rho\right]\!\right] \underline{u}_{xt}^{(0)} - \underline{\sigma}_x^{(1)}. \tag{5.12b}$$

We remove x^*-independent terms by setting

$$\begin{aligned} 0 = {} & \left\langle \left[\!\left[\rho\right]\!\right] K^{-1} \right\rangle \underline{u}_{tt}^{(0)} + \left\langle K^{-1} \right\rangle \underline{\sigma}_t^{(1)} - G(\underline{\sigma}^{(0)})\underline{u}_x^{(1)} - \left\langle \left[\!\left[\rho\right]\!\right] K^{-1} \right\rangle \frac{\underline{u}_t^{(0)}\underline{\sigma}_t^{(0)}G'(\underline{\sigma}^{(0)})}{G(\underline{\sigma}^{(0)})} \\ & - \left\langle K^{-1} \right\rangle \frac{\underline{\sigma}^{(1)}\underline{\sigma}_t^{(0)}G'(\underline{\sigma}^{(0)})}{G(\underline{\sigma}^{(0)})}, \end{aligned} \tag{5.13a}$$

$$0 = \left\langle \rho \left[\!\left[K^{-1}\right]\!\right] \right\rangle \frac{G(\underline{\sigma}^{(0)})\underline{\sigma}_{tt}^{(0)} - (\underline{\sigma}_t^{(0)})^2 G'(\underline{\sigma}^{(0)})}{G(\underline{\sigma}^{(0)})^2} + \langle \rho \rangle \, \underline{u}_t^{(1)} - \underline{\sigma}_x^{(1)}. \tag{5.13b}$$

To save space, we will not perform the final step in the analysis of the $O(\delta)$ system of equations, which involves integrating remaining terms in (5.12).

The most important part of the analysis above is the removal of all terms that are independent of x^*, since these terms lead to the homogenized equations. To see this, let us define

$$\begin{aligned} \underline{u}(x,t) &= \langle u(x,x^*,t) \rangle = \underline{u}^{(0)}(x,t) + \delta \underline{u}^{(1)}(x,t) + O(\delta^2), \\ \underline{\sigma}(x,t) &= \langle \sigma(x,x^*,t) \rangle = \underline{\sigma}^{(0)}(x,t) + \delta \underline{\sigma}^{(1)}(x,t) + O(\delta^2) \end{aligned}$$

and add up the homogenized equations (5.9) and (5.13) to get

$$0 = \left\langle K^{-1} \right\rangle \underline{\sigma}_t - G(\underline{\sigma}^{(0)})\underline{u}_x + \delta \left\langle \left[\!\left[\rho\right]\!\right] K^{-1} \right\rangle \underline{u}_{tt} - \delta \left\langle \left[\!\left[\rho\right]\!\right] K^{-1} \right\rangle \frac{\underline{u}_t \underline{\sigma}_t G'(\underline{\sigma}^{(0)})}{G(\underline{\sigma}^{(0)})}$$

$$- \delta \left\langle K^{-1} \right\rangle \frac{\underline{\sigma}^{(1)} \underline{\sigma}_t G'(\underline{\sigma}^{(0)})}{G(\underline{\sigma}^{(0)})} + O(\delta^2), \tag{5.14a}$$

$$0 = \langle \rho \rangle \, \underline{u}_t - \underline{\sigma}_x + \delta \left\langle \rho \left[\!\left[K^{-1} \right]\!\right] \right\rangle \frac{G(\underline{\sigma}^{(0)})\underline{\sigma}^{(0)}_{tt} - (\underline{\sigma}^{(0)}_t)^2 G'(\underline{\sigma}^{(0)})}{G(\underline{\sigma}^{(0)})^2} + O(\delta^2). \tag{5.14b}$$

To make these equations look more like the original system of equations (3.3), we replace all t-derivatives (except for the first terms in (5.14)) with x-derivatives by substituting the equations into themselves. For example,

$$\begin{aligned} \frac{G(\underline{\sigma}^{(0)})\underline{\sigma}^{(0)}_{tt} - (\underline{\sigma}^{(0)}_t)^2 G'(\underline{\sigma}^{(0)})}{G(\underline{\sigma}^{(0)})^2} &= \frac{\partial}{\partial t}\left[\frac{\underline{\sigma}^{(0)}_t}{G(\underline{\sigma}^{(0)})}\right] = \frac{\partial}{\partial t}\left[\frac{\underline{u}^{(0)}_x}{\langle K^{-1} \rangle}\right] + O(\delta) \\ &= \frac{\underline{\sigma}^{(0)}_{xx}}{\langle \rho \rangle \langle K^{-1} \rangle} + O(\delta) = \frac{\underline{\sigma}_{xx}}{\langle \rho \rangle \langle K^{-1} \rangle} + O(\delta). \end{aligned}$$

Similar substitutions eventually produce

$$0 = \left\langle K^{-1} \right\rangle \underline{\sigma}_t - G(\underline{\sigma}^{(0)})\underline{u}_x - \delta \underline{\sigma}^{(1)} \underline{u}_x G'(\underline{\sigma}^{(0)}) + \delta \frac{\left\langle \left[\!\left[\rho\right]\!\right] K^{-1} \right\rangle}{\langle \rho \rangle \langle K^{-1} \rangle} G(\underline{\sigma}^{(0)}) \underline{u}_{xx} + O(\delta^2), \tag{5.15a}$$

$$0 = \langle \rho \rangle \, \underline{u}_t - \underline{\sigma}_x + \delta \frac{\left\langle \rho \left[\!\left[K^{-1} \right]\!\right] \right\rangle}{\langle \rho \rangle \langle K^{-1} \rangle} \underline{\sigma}_{xx} + O(\delta^2). \tag{5.15b}$$

Furthermore, we recognize the first few terms of the expansion of $G(\underline{\sigma})$ in (5.15a), and thus this can be simplified to

$$0 = \left\langle K^{-1} \right\rangle \underline{\sigma}_t - G(\underline{\sigma})\underline{u}_x + \delta \frac{\left\langle \left[\!\left[\rho\right]\!\right] K^{-1} \right\rangle}{\langle \rho \rangle \langle K^{-1} \rangle} G(\underline{\sigma}) \underline{u}_{xx} + O(\delta^2), \tag{5.16a}$$

$$0 = \langle \rho \rangle \, \underline{u}_t - \underline{\sigma}_x + \delta \frac{\left\langle \rho \left[\!\left[K^{-1} \right]\!\right] \right\rangle}{\langle \rho \rangle \langle K^{-1} \rangle} \underline{\sigma}_{xx} + O(\delta^2). \tag{5.16b}$$

The equations above represent the homogenized versions of (3.3), with terms up to $O(\delta)$ included. However, $\left\langle \rho \left[\!\left[K^{-1} \right]\!\right] \right\rangle = 0$ for piecewise constant functions $\rho(x)$ and $K(x)$, so we have to compute more terms of this equation to see any interesting behavior. Hence the terms involving second derivatives vanish in this case, and dispersive effects at the next order will dominate. (For other choices of $\rho(x)$ and $K(x)$ that are rapidly varying but not piecewise constant, the second-derivative terms may not drop out. Numerical experiments show different behavior in this case, but we have not yet investigated this in detail.)

As the algebra involved increases exponentially with each order of δ, we have employed Mathematica to perform the calculations. The homogenized equations in-

cluding $O(\delta^2)$ terms are found to be

$$0 = \left\langle K^{-1} \right\rangle \underline{\sigma}_t - G(\underline{\sigma})\underline{u}_x + \delta C_{11} G(\underline{\sigma})\underline{u}_{xx} + \delta^2 C_{12} G(\underline{\sigma})\underline{u}_{xxx} + \delta^2 C_{13}\left(G'(\underline{\sigma})\underline{\sigma}_x \underline{u}_{xx} + \frac{1}{2} G''(\underline{\sigma})\underline{\sigma}_x^2 \underline{u}_x \right) + O(\delta^3), \tag{5.17a}$$

$$0 = \langle \rho \rangle \, \underline{u}_t - \underline{\sigma}_x + \delta C_{21} \underline{\sigma}_{xx} + \delta^2 C_{22} \underline{\sigma}_{xxx} + O(\delta^3), \tag{5.17b}$$

where

$$\begin{aligned}
C_{11} &= \frac{\left\langle [\![\rho]\!] K^{-1} \right\rangle}{\langle \rho \rangle \left\langle K^{-1} \right\rangle}, \\
C_{21} &= \frac{\left\langle \rho [\![K^{-1}]\!] \right\rangle}{\langle \rho \rangle \left\langle K^{-1} \right\rangle}, \\
C_{12} &= \frac{\left\langle [\![K^{-1}]\!] [\![\rho]\!] \right\rangle}{\left\langle K^{-1} \right\rangle \langle \rho \rangle} - \frac{\left\langle K^{-1} [\![\rho]\!]^2 \right\rangle}{\left\langle K^{-1} \right\rangle \langle \rho \rangle^2}, \\
C_{22} &= \frac{\left\langle [\![K^{-1}]\!] [\![\rho]\!] \right\rangle}{\left\langle K^{-1} \right\rangle \langle \rho \rangle} - \frac{\left\langle [\![K^{-1}]\!]^2 \rho \right\rangle}{\left\langle K^{-1} \right\rangle^2 \langle \rho \rangle}, \\
C_{13} &= 2\frac{\left\langle [\![K^{-1}]\!] \rho \right\rangle^2}{\left\langle K^{-1} \right\rangle^2 \langle \rho \rangle^2} + 2\frac{\left\langle [\![K^{-1}]\!] [\![\rho]\!] \right\rangle}{\left\langle K^{-1} \right\rangle \langle \rho \rangle} - 2\frac{\left\langle [\![K^{-1}]\!]^2 \rho \right\rangle}{\left\langle K^{-1} \right\rangle^2 \langle \rho \rangle} - \frac{\left\langle K^{-1} [\![\rho]\!]^2 \right\rangle}{\left\langle K^{-1} \right\rangle \langle \rho \rangle^2}.
\end{aligned} \tag{5.18}$$

In general it can be shown that $C_{11} = -C_{12}$; see [11]. For the special case of piecewise constant material parameters considered in this paper, we find that

$$\begin{aligned}
C_{11} &= C_{21} = 0, \\
C_{12} &= -\frac{1}{12}\alpha^2(1-\alpha)^2 \frac{(\rho_A - \rho_B)(Z_A^2 - Z_B^2)}{K_A K_B \left\langle K^{-1} \right\rangle \langle \rho \rangle^2}, \\
C_{22} &= -\frac{1}{12}\alpha^2(1-\alpha)^2 \frac{(K_A - K_B)(Z_A^2 - Z_B^2)}{K_A^2 K_B^2 \left\langle K^{-1} \right\rangle^2 \langle \rho \rangle}, \\
C_{13} &= -\frac{1}{12}\alpha^2(1-\alpha)^2 \frac{\langle \rho \rangle^2 (K_A - K_B)^2 + (Z_A^2 - Z_B^2)}{K_A^2 K_B^2 \left\langle K^{-1} \right\rangle^2 \langle \rho \rangle^2}.
\end{aligned} \tag{5.19}$$

Here $Z_A = \sqrt{\rho_A K_A}$ and $Z_B = \sqrt{\rho_B K_B}$ are the linearized impedances. Note that if $Z_A = Z_B$, then $C_{12} = C_{22} = 0$. Also note that for the linear case $G(\sigma) = 1$, the factor multiplying C_{13} vanishes in the homogenized equations (5.17).

Additional terms in the homogenized equations are too complicated to present in their general form. For the specific case used in section 3 to produce Figures 4–9, we have calculated the homogenized equations including $O(\delta^4)$ terms. We take $\alpha = 1/2$ and use the exponential stress-strain relationship $G(\sigma) = \sigma + 1$ and the piecewise constant material parameters $\rho_A = k_A = 4$, $\rho_B = k_B = 1$. The homogenized

equations are then

$$u_t = \frac{2\sigma_x}{5} + \frac{3\delta^2\sigma_{xxx}}{500} + \delta^4\left(\frac{3\sigma_{xxx}\sigma_x^2}{15625(\sigma+1)^2} - \frac{72u_{xx}^2\sigma_x}{15625(\sigma+1)} - \frac{12\sigma_{xxxx}\sigma_x}{15625(\sigma+1)}\right.$$

(5.20a)

$$\left. - \frac{96u_{xx}u_{xxx}}{15625} - \frac{12\sigma_{xx}\sigma_{xxx}}{15625(\sigma+1)} - \frac{357\sigma_{xxxxx}}{1000000}\right) + O(\delta^6),$$

$$\sigma_t = \frac{8(\sigma+1)u_x}{5} + \delta^2\left(\frac{3(\sigma+1)u_{xxx}}{125} + \frac{3u_{xx}\sigma_x}{50}\right)$$

$$+ \delta^4\left(\frac{48u_xu_{xx}^2}{15625} - \frac{48\sigma_x\sigma_{xx}u_{xx}}{15625(\sigma+1)} - \frac{4761\sigma_{xxx}u_{xx}}{500000} - \frac{72u_{xxx}\sigma_x^2}{15625(\sigma+1)}\right.$$

(5.20b)

$$\left. - \frac{357(\sigma+1)u_{xxxxx}}{250000} - \frac{3543u_{xxxx}\sigma_x}{500000} - \frac{3891u_{xxx}\sigma_{xx}}{500000}\right) + O(\delta^6).$$

(The underlines have been dropped for clarity.) Note again that for this case the u_{xx} and σ_{xx} terms drop out, and we can see the appearance of dispersive terms like u_{xxx} and σ_{xxx} in δ^2 terms. Also notice that (5.20) has the property that it is invariant under the reflection $u(x,t) \to -u(-x,t)$ and $\sigma(x,t) \to \sigma(-x,t)$, a property inherited from the original equations (3.3).

Furthermore, we see that each term proportional to δ^n always has a total of $n+1$ spatial derivatives, which corroborates the observation made earlier that if $\tilde{u}(x,t)$ and $\tilde{\sigma}(x,t)$ are solutions to (5.20) for $\delta = 1$, then $u(x,t) = \tilde{u}(x/\delta, t/\delta)$ and $\sigma(x,t) = \tilde{\sigma}(x/\delta, t/\delta)$ are solutions for arbitrary δ. This property implies that the value of δ does not have to be small and that these homogenized equations are valid for any δ. This suggests that the "higher-order" terms can be dropped, not because δ is small, but because they involve small coefficients coming from higher-order averages of the rapidly varying coefficients, as is already apparent in (5.20).

In Figure 12 we compare the solution to the homogenized equations (5.20) with $\delta = 1$, dropping the $O(\delta^6)$ terms, with the numerical solution to the original equations (3.3). The homogenized equations were solved using a pseudospectral numerical method that computes these smooth solutions with high accuracy (using techniques discussed in [2] and [9] for similar equations). In this test we take initial data consisting of a nonzero strain in the middle of the computational domain, with zero velocity, which results in two outgoing pulses that each break up into a train of stegotons. At later times we show only the leftgoing wave train. We use this initial data with periodic boundary conditions for ease in using the pseudospectral method, and also because it is not clear how to impose appropriate boundary conditions of the type used earlier for the higher-order homogenized equations. The agreement is good, especially when one considers that the waves have traveled over 400 units by $t = 500$. If only the terms up to $O(\delta^3)$ are retained, solitary waves are still observed, but the agreement is not as good.

Solving the homogenized equations with boundary data corresponding directly to the layered medium tests done earlier would require the derivation of appropriate boundary conditions for these higher-order equations. An example of how this can be done in a weakly nonlinear case is shown in [10] and [11], but this has not yet been carried through for the homogenized equations presented here.

6. Conclusions. We have studied elastic waves in nonlinear layered media by computing numerical solutions to a first-order hyperbolic system of conservation laws

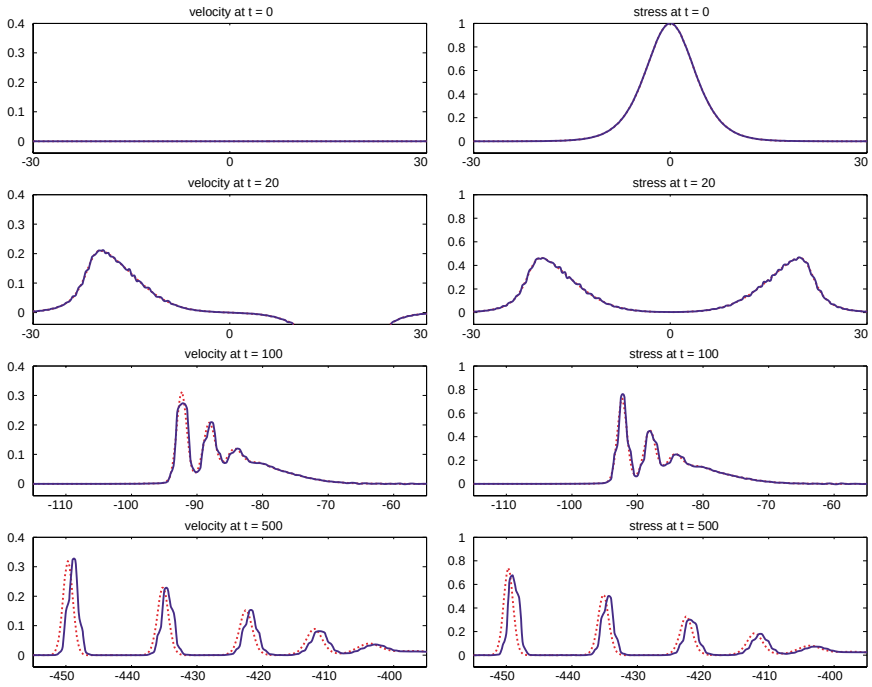


FIG. 12. *Comparison of a pseudospectral solution (dashed curve) to the homogenized equations* (5.20) *with the finite-volume solution (solid curve) to the layered medium equations* (3.3).

with a spatially varying flux function, using the method described in [1] and [4]. The layering leads to dispersive behavior if the layers are not impedance-matched, which in turn gives rise to the appearance of solitary waves that appear to interact in the manner of classical solitons. There is an approximate one-parameter family of such waves, whose velocity varies linearly with amplitude. By studying the solution as a function of time at a fixed point x_0, we obtain some indication of the scaling properties of these waves as the amplitude is varied.

We also showed that, for a special limiting choice of parameters, the layered medium can be modeled directly by the Toda lattice. Since the Toda lattice is known to have exact soliton solutions, it is not surprising that similar behavior is observed in the layered medium in this case. It is more surprising that solitary waves are observed in cases far from this limit.

A set of nonlinear homogenized equations has been derived that contains dispersive terms. Numerical solution of these equations yields results that agree well with the direct solution of the original hyperbolic system. We hope that further study of these equations may provide more insight into the nature of these solitary waves.

We have studied in detail only one particular choice of material parameters in the piecewise constant case with an exponential stress-strain relation. Preliminary experiments with different choices show a rich variety of other interesting behavior that should be explored further.

Acknowledgment. The authors would like to thank Arnold D. Kim for advice on pseudospectral numerical methods.

REFERENCES

[1] D. S. Bale, R. J. LeVeque, S. Mitran, and J. A. Rossmanith, *A wave propagation method for conservation laws and balance laws with spatially varying flux functions*, SIAM J. Sci. Comput., 24 (2002), pp. 955–978.

[2] T. B. Benjamin, J. L. Bona, and J. J. Mahony, *Model equations for long waves in nonlinear dispersive systems*, Philos. Trans. Roy. Soc. London Ser. A, 272 (1972), pp. 47–78.

[3] J. Kevorkian and D. L. Bosley, *Multiple-scale homogenization for weakly nonlinear conservation laws with rapid spatial fluctuations*, Stud. Appl. Math., 101 (1998), pp. 127–183.

[4] R. J. LeVeque, *Finite volume methods for nonlinear elasticity in heterogeneous media*, Internat. J. Numer. Methods Fluids, 40 (2002), pp. 93–104.

[5] A. M. Samsonov, *Strain Solitons in Solids and How to Construct Them*, Chapman & Hall/CRC, Boca Raton, FL, 2001.

[6] F. Santosa and W. W. Symes, *A dispersive effective medium for wave propagation in periodic composites*, SIAM J. Appl. Math., 51 (1991), pp. 984–1005.

[7] M. Toda, *Nonlinear Waves and Solitons*, Kluwer Academic Publishers, Boston, 1989.

[8] M. Toda, *Theory of Nonlinear Lattices*, Springer-Verlag, New York, Berlin, 1989.

[9] L. N. Trefethen, *Spectral Methods in* MATLAB, Software Environ. Tools. 10, SIAM, Philadelphia, 2000.

[10] D. H. Yong, *Solving Boundary-Value Problems for Systems of Hyperbolic Conservation Laws with Rapidly Varying Coefficients*, Ph.D. dissertation, University of Washington, Seattle, 2000.

[11] D. H. Yong and J. Kevorkian, *Solving boundary-value problems for systems of hyperbolic conservation laws with rapidly varying coefficients*, Stud. Appl. Math., 108 (2002), pp. 259–303.

SIAM J. APPL. MATH.
Vol. 63, No. 5, pp. 1561–1584

FORMAL ASYMPTOTIC MODELS OF VEHICULAR TRAFFIC. MODEL CLOSURES*

ALEXANDROS SOPASAKIS†

Abstract. Formal closed models for vehicular traffic flow are obtained based on the novel equilibrium solution of the Prigogine–Herman equation. To that effect, Hilbert and Chapman–Enskog asymptotic series expansions are employed, obtaining the Euler and Navier–Stokes equivalent equations for traffic flow.

Key words. traffic flow models, Chapman–Enskog, Hilbert asymptotic expansions, nonlinear kinetic equations

AMS subject classifications. 82D99, 76R50, 41A60, 35K57

DOI. 10.1137/S0036139902403020

1. Introduction. We show that hydrodynamic (continuum) models of traffic can be obtained via asymptotic expansions about equilibrium solutions of the Prigogine–Herman kinetic equation. We therefore focus our efforts on the interval of vehicular concentrations c, which are less than or equal to the critical concentration c_{crit}. Here c_{crit} is the concentration for which the equilibrium solutions of the Prigogine–Herman equation bifurcate from a one-parameter family of solutions to a two-parameter family [25]. It is interesting to note that for many reasonable data (i.e., f_0, P, T) in the Prigogine–Herman equation the mathematical value of the critical concentration closely coincides with the value of the concentration at which traffic flow is observed to become "unstable" [7]. The term "unstable" is used by traffic engineers to describe the onset of traffic breakdown (flow is no longer smooth and is prone to sudden fluctuations). Mathematically this unstable regime is comprised of the concentrations c above some constant critical concentration defined via $c_{crit} = \frac{1}{Tw(1-P)}$, where w denotes the desired speed of drivers (T, P are as in (3.5), (3.6), respectively).

Macroscopic (continuum) models of traffic flow typically take the form of partial differential equations that have as their dependent variables quantities that can be identified with low-order polynomial moments (in speed) of the distribution function that is the dependent variable in a kinetic equation. One natural application of kinetic equations is therefore the rational development of continuum approximations for traffic flow. One possible approach is via asymptotic expansions, similar to those of Chapman and Enskog (e.g., Chapman and Cowling [6]) or Hilbert [12] in the kinetic theory of gases, about some underlying "equilibrium solution" as the low-order approximation. Because equilibrium solutions of nonlinear kinetic equations of vehicular traffic are not easy to obtain [9, 16], approaches other than asymptotic expansions of the type mentioned above have primarily been used to date. For example, Helbing [10] assumes a Gaussian type of equilibrium solution for the Paveri-Fontana [28] kinetic equation and expands asymptotically about it. Wegener and Klar [34] obtain macro-

*Received by the editors February 22, 2002; accepted for publication (in revised form) January 28, 2003; published electronically June 12, 2003. This research was partially supported by the E.U. TMR program, contract ERB FMRX-CT97-0157.

http://www.siam.org/journals/siap/63-5/40302.html

†Department of Mathematics, University of California at Berkeley, Berkeley, CA 94720. Also affiliated with the Department of Mathematics, Georgia Institute of Technology, Atlanta, GA. Current address: Courant Institute, 251 Mercer Street, New York University, New York, NY 10012 (sopasak@cims.nyu.edu).

scopic approximations in the classical manner of introducing *ad hoc* approximations to "close" the system of equations obtained from the first few polynomial moments of a particular kinetic equation.

Cercignani [4] uses an implementation of the method as proposed by Chapman [5] and Enskog [8] to obtain Chapman–Enskog approximations of different orders for the Boltzmann equation. Nelson and Sopasakis [26] have carried out similar calculations using Chapman–Enskog-type expansions, obtaining zeroth- and first-order model equations of traffic flow.

In this work we employ Hilbert expansions, obtaining zeroth- and first-order models for the low concentration regime. Further we obtain a higher-order "Burnett equivalent" model, which is valid for the low concentration regime, using Chapman–Enskog approximations up to second order.

We start by presenting in review the Prigogine–Herman equation and a brief overview of similar models in section 2. In section 3 we describe and explain how to obtain the equilibrium solutions for the Prigogine–Herman equation under different cases of the passing probability P and relaxation time T. Then in section 4 we show how to use those equilibrium solutions and expand asymptotically around them, thus obtaining traffic flow models. Expansions of Chapman–Enskog type are presented in subsection 4.1. Those include the zero-order or Euler-like expansion, the first-order or Navier–Stokes-like expansion, and the second-order or Burnett-like expansion. Equivalent expansions of Hilbert type are presented in subsection 4.3 for the zeroth and first orders only. Last, we briefly present a preliminary numerical simulation for only the zeroth- and first-order approximations and for a very simple traffic incident in section 5. For the complete numerical investigations, we refer to the sequel of this paper [32].

2. Nonlinear models of vehicular traffic. The kinetic model of Prigogine and Herman is

$$\frac{\partial f(x,v,t)}{\partial t} + v\frac{\partial f(x,v,t)}{\partial x} = -\frac{(f(x,v,t) - f_0(x,v,t))}{T} + c(x,t)(\bar{v}-v)(1-P)f(x,v,t). \tag{2.1}$$

Here the various symbols have the following meanings:

– The zeroth-order moment of $f(x,v,t)$, $c(x,t) = \int_0^\infty f(x,v,t)\,dv$ is vehicular *concentration* (or *density*);

– The first-order moment, $\bar{v}(x,t) = \frac{1}{c(x,t)}\int_0^\infty v f(x,v,t)\,dv$, is mean *speed*;

– P is the passing probability, which is assumed known (in this work depending explicitly on c);

– T is the relaxation time, which is assumed known (in this work depending explicitly on c);

– $f_0(x,v,t)$ is the density function for the desired speed of vehicles, which is assumed to be known a priori;

– $f(x,v,t)$ is the *distribution function* of vehicles in space and speed. Thus $f(x,v,t)\,dx\,dv$ is the expected number of vehicles at time t that have position between x and $x+dx$ and speed between v and $v+dv$. This is the unknown function, which presumably is determined by the Prigogine–Herman equation, along with suitable boundary and initial conditions.

In the text to follow, all notation will be suppressed (when understood) in regard to dependence on space (x), time (t), and velocity (v) variables, unless necessary for clarity. All distribution functions to be encountered depend on x, v, and t.

A vehicle, in practice, changes speed due to three types of vehicular interactions: "slowing down," "speeding up," and "passing." In the classical Prigogine–Herman kinetic model of vehicular traffic, the first of these types is treated by a fundamental quadratic interaction term of the Boltzmann type,

$$c(x,t)(\bar{v}-v)(1-P)f(x,v,t) \equiv (1-P)\left(f(v)\int_0^\infty f(v')v'\,dv' - f(v)v\int_0^\infty f(v')\,dv'\right),$$

based on the assumption that the speed of the vehicle slowing down is uncorrelated with that of the vehicle immediately ahead ("lead vehicle") that necessitates the slowing down. (This is reasonable since we assume a one-lane highway where passing is always possible and can occur without change in speed by use of another lane which is designated for passing alone.) By contrast, the other two types of interactions, for which this "zero correlation" assumption is not true, are treated through a phenomenological relaxation term,

$$-\frac{(f(x,v,t)-f_0(x,v,t))}{T}.$$

In keeping with the traditions of the Boltzmann equation of the kinetic theory of gases, both accelerations and decelerations are assumed to occur instantaneously. This imposes some limitations on the distance and time scales for validity of the resulting kinetic equation, and therefore also on any macroscopic (continuum) equation derived therefrom. Limitations of the latter type do not appear to have been extensively discussed in the traffic flow literature; they are discussed in the work of Nelson, Bui, and Sopasakis [27]. One should be very careful in using the Prigogine–Herman equation to attempt to resolve traffic phenomena on time scales comparable to or smaller than the time for a vehicle to accelerate. Also important in the derivation of the equation is the assumption that vehicles are treated as being "point particles," which means that they have zero length.

Paveri-Fontana adds [28] a new dimension to the Prigogine–Herman model by treating desired speeds of drivers as an additional independent variable w,

$$\begin{aligned}\left(\frac{\partial}{\partial t}+v\frac{\partial}{\partial x}\right)g(x,v,t;w)+\frac{\partial}{\partial v}\left(\frac{w-v}{T}g(x,v,t;w)\right)\\ = f(x,v,t)\int_v^\infty (1-P)(v'-v)g(x,v',t;w)\,dv'\\ -g(x,v,t;w)\int_0^v (1-P)(v-v')f(x,v',t)\,dv',\end{aligned}$$

where

$$f(x,v,t)=\int_0^\infty g(x,v,t;w)\,dw.$$

Here $g = g(x,v,w;t)$ is the density function, so that $g(x,v,w,t)\,dx\,dv\,dw$ is the expected number of vehicles at time t that have position between x and $x+dx$, speed between v and $v+dv$, and desired velocity between w and $w+dw$. P is probability of passing, T is relaxation time, and v is velocity. Recently Hoogendoorn and Bovy [13] generalize the Paveri-Fontana theory and develop a traffic model with multiple user classes and thus multiple desired speeds.

Nelson [24] attempts to treat both "slowing down" and "speeding up" interactions, as quadratic interaction terms of the Boltzmann type,

$$\frac{\partial f}{\partial t} + v\frac{\partial f}{\partial x} = \left(\frac{\delta f}{\delta t}\right)_+ + \left(\frac{\delta f}{\delta t}\right)_- .$$

Here $(\frac{\delta f}{\delta t})_+$ and $(\frac{\delta f}{\delta t})_-$ denote the rate of change of f due, respectively, to speeding up and slowing down interactions, and suitable expressions for these were obtained by introducing a certain "mechanical model" describing the behavior of drivers, and a certain correlation model approximating the passing distribution of vehicles in terms of f.

Klar and Wegener [16] built on the ideas of Nelson and derived a kinetic model of vehicular traffic which also takes into account vehicle length:

$$\frac{\partial f}{\partial t} + v\frac{\partial f}{\partial x} = \left(\frac{\delta f}{\delta t}\right)_g + \left(\frac{\delta f}{\delta t}\right)_l,$$

$$\text{where } \left(\frac{\delta f}{\delta t}\right)_g + \left(\frac{\delta f}{\delta t}\right)_l = \sum_{i=1}^{N}\left(\frac{\delta f}{\delta t}\right)_g^i + \sum_{i=1}^{N}\left(\frac{\delta f}{\delta t}\right)_l^i.$$

Here $(\frac{\delta f}{\delta t})_g$ and $(\frac{\delta f}{\delta t})_l$ denote the gain and loss terms, respectively, due to the ith threshold. A threshold here generates a change in velocity. Consider two cars, car 1 with position x_1 and speed v_1, and car 2 at position x_2 and speed v_2. Car 1 is assumed to change velocity only in response to its leading vehicle, car 2. If car 1 is faster than the leading vehicle, car 2, and the headway to car 2 becomes smaller than a certain threshold, the driver will either slow down or pass the leading vehicle. This use of thresholds allows cars to be considered as also having lengths.

3. Equilibrium solutions and behavior at high concentrations. This section will serve to introduce all the necessary background information regarding equilibrium solutions, as a prelude to the asymptotic expansions to follow in the next section. This leads to novel equilibrium solutions of the Prigogine–Herman kinetic equation (based on some mild assumptions to be found in [25]) that comprise a one-parameter family in the "stable" low-concentration regime, $0 < c < c_{crit}$ for a certain $c_{crit} > 0$, but a two-parameter family on the complementary unstable concentration range, $c_{crit} < c < c_{jam}$. Here c_{jam} is defined to be the concentration of vehicles for which the flow becomes zero.

We introduce the following notations:

$$v_0(c) = \frac{1}{Tc(1-P)}, \quad \zeta = \zeta(c) = \bar{v} - v_0(c). \tag{3.1}$$

The equilibrium solutions of (2.1) can then be expressed as

$$f_{eq} = f_{eq}(v) = \frac{v_0}{v-\zeta} f_0 + \alpha c \delta(v-\zeta), \tag{3.2}$$

where δ is the Dirac delta function, α a parameter to be determined later, and all dependence on x and t has been notationally suppressed.

We assume that the desired speed distribution function (f_0), the passing probability (P), and the relaxation time (T) are known a priori (with P and T even taken to depend on concentration (c)), according to the usual assumptions of the

Prigogine–Herman kinetic theory. Under these assumptions the equilibrium solution (3.2) depends upon three parameters, namely c, $\bar{v}$, and α. (For some of the following considerations it is convenient to replace one of these parameters, usually α, by ζ.) However, the basic definitions of the distribution function and the associated concentration requires satisfaction of the *normalization condition* (for the zeroth moment),

$$c = \int_0^\infty f_{eq}(v)\, dv, \tag{3.3}$$

and the *mean-speed condition* (for the first moment),

$$c\bar{v} = \int_0^\infty v f_{eq}(v)\, dv. \tag{3.4}$$

Upon substituting the expression (3.2) into these conditions, there results a system of two equations that must be satisfied by these three parameters. We will work with more realistic forms of passing probability and relaxation time, as also assumed in Chapter 4 of Prigogine and Herman [29],

$$P = 1 - \frac{c}{c_{jam}} \tag{3.5}$$

and

$$T = \tau \frac{c/c_{jam}}{1 - c/c_{jam}}. \tag{3.6}$$

We shall frequently find it convenient to work with the "reduced" desired speed distribution, $\varphi_0(v; c)$, as defined by $\int c\varphi_0(v; c)\, dv = \int f_0(v)\, dv$. The main result of this section can now be summarized by the following theorem.

THEOREM 3.1. *In terms of the reduced desired speed distribution, the possible equilibrium solutions* (3.2) *can be written as*

$$f_{eq}(v; c, \alpha) = cv_0(c)\frac{\varphi_0(v; c)}{v - \zeta} + c\alpha\delta(v - \zeta) \tag{3.7}$$

$$\text{for the cases of} \begin{cases} 0 < c \le c_{crit} & \text{and } \alpha = 0, \text{ or} \\ c_{crit} < c < c_{jam} & \text{and } \alpha_{min} \le \alpha \le \alpha_{max}, \end{cases} \tag{3.8}$$

where $\zeta = \bar{\zeta}(\alpha, c) = w - \frac{v_o(c)}{1-\alpha}$, c_{jam} *is the jam concentration of vehicles,* $v_0(c)$ *is defined in* (3.1), *and* c_{crit} *is the root of* $c := \frac{1}{T(1-P)} \int_{w_-}^{w^+} \frac{\varphi_0(v;c)}{v-\zeta}\, dv$. *Further,* $\alpha_{min} = \max\{0, 1 - v_0(c)F(w_-; c)\}$ *and* $\alpha_{max} = 1 - v_0(c)F(0; c)$, *where* $F(\zeta; c) = \int_{w-}^{w+} \frac{\varphi_0(v;c)}{v-\zeta}\, dv$.

(The exact same theorem will hold if we take constant values of P and T.) For the details regarding the proof of this theorem, we refer the reader to [25]. To obtain this result it is necessary to invoke the fact that the traffic-theoretic interpretation of $f_{eq}(v)$ requires that it have support on the interval 0 to w and assumes only nonnegative values there. The equilibrium solutions comprise a two-parameter family (e.g., c and α), for $c > c_{crit}$. The delta function component can be interpreted as "platoons" of vehicles traveling at some "synchronized" speed, but now the platoon speed can vary from $\bar{\zeta}(\alpha_{max}, c) = 0$ to $\bar{\zeta}(\alpha_{min}, c) > 0$, rather than being fixed at 0 as in the classical results of Prigogine and Herman [29, Chapter 4] and Prigogine, Herman, and Anderson [30].

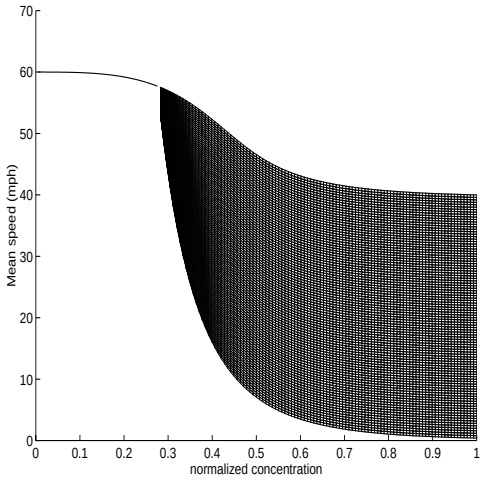


FIG. 1. *Dependence of the mean-speed curve/continuum on the concentration for a uniform desired speed distribution from* 45 *to* 90 *mph* ($c_{jam} = 225$ *vpm,* $\tau = .002$ *hours).*

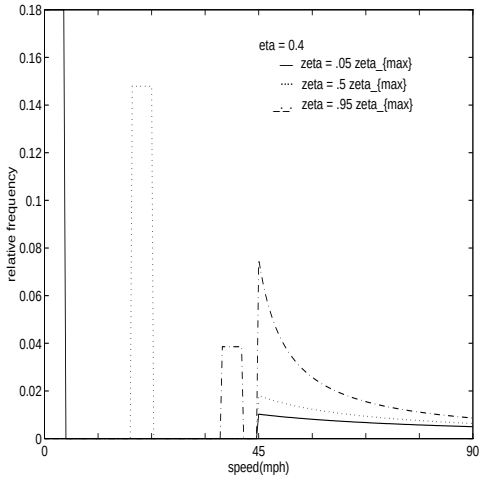


FIG. 2. *Variation of the equilibrium distribution with speed of the collective flow, for normalized concentration* = .4 *and a uniform desired speed distribution from* 45 *to* 90 *mph* ($c_{jam} = 225$ *vpm,* $\tau = .002$ *hours).*

Figure 1 shows the mean-speed curve/continuum as a function of concentration, for a hypothetical situation in which the distribution of desired speeds is uniform from 45 mph to 90 mph, $c_{jam} = 225$ vehicles per mile (vpm), and $\tau = .002$ hours. A graph of this distribution function can be found in Figure 2. (These values are reasonably representative of those obtained observationally by Edie, Herman, and Lam [7].) For this instance the bifurcation from a curve to a continuum occurs at slightly above $c_{jam}/3$ ($c_{crit} \approx 0.29 \cdot c_{jam} = 65$ vpm). This behavior of a two-parameter family of solutions for concentrations above c_{crit} is the reason that the Prigogine–Herman equation is so attractive and preferable to other alternative models of traffic flow. This model equation, which not only allows for an explicitly calculated equilibrium solution to be found, also predicts (see [33]) recently only experimentally observed

concepts such as "synchronized flow" (see Kerner [15]) and "stop and go" traffic.

4. Formal asymptotic solutions of the Prigogine–Herman kinetic equation. Here we use the equilibrium solutions previously obtained for the Prigogine–Herman kinetic equation and asymptotically expand them in a small parameter ϵ (using the Chapman–Enskog [5, 8] or Hilbert [12] method) to derive hydrodynamic equivalent equations for traffic flow for the stable ($0 < c < c_{crit}$) traffic regime. The purpose of these expansions will be to produce hydrodynamic equivalent approximations of traffic flow models.

4.1. The Chapman–Enskog asymptotic expansion. In this section the zeroth-, first-, and second-order Chapman–Enskog asymptotic expansions are obtained. Numerical results under appropriate conservation-preserving and entropy-satisfying methods follow in a sequel to this paper [32].

We find a formal asymptotic solution, for the Prigogine–Herman equation, of the form

$$f = \sum_{n=0}^{\infty} \epsilon^n f^{(n)}. \tag{4.1}$$

To accomplish this we rewrite the Prigogine–Herman equation with the addition of an artificial "small" parameter ϵ on the left-hand side,

$$\epsilon \left(\frac{\partial f}{\partial t} + v \frac{\partial f}{\partial x} \right) = Qf, \tag{4.2}$$

where the operator Qf is defined as

$$Qf = -\frac{f - \varphi_0(v) \int_0^{w^+} f(v')\, dv'}{T} + (1 - P) f(v) \int_0^{w^+} (v' - v) f(v')\, dv'. \tag{4.3}$$

The addition of the parameter ϵ concerns scaling issues of Qf. For the Chapman–Enskog expansion the time derivative is also expanded in powers of ϵ, as

$$\frac{\partial f}{\partial t} = \sum_{n=0}^{\infty} \epsilon^n \frac{\partial^{(n)} f}{\partial t} = \sum_{n=0}^{\infty} \epsilon^n \left(\sum_{k=0}^{n} \frac{\partial^{(k)} f^{(n-k)}}{\partial t} \right). \tag{4.4}$$

In the last equality the expression (4.1) has been used. The $\frac{\partial^{(k)} f^{(n-k)}}{\partial t}$ are assumed to be unknowns to be determined so that existence of terms of the solution (4.1) is ensured when (4.1) is substituted into (4.2).

Although f is expanded in powers of ϵ, the concentration c, according to the Chapman–Enskog method, is not. Thus it is required that all higher-order contributions to c be zero:

$$c^{(n)} = \int_0^{w^+} f^{(n)}\, dv \equiv 0, \quad n = 1, 2, \ldots, \quad \text{and} \tag{4.5}$$

$$c^{(0)} = \int_0^{w^+} f^{(0)}\, dv \equiv c(x, t), \tag{4.6}$$

$$\text{where} \qquad c = \sum_{n=0}^{\infty} \epsilon^n c^{(n)}. \tag{4.7}$$

Another important point which is a result of the previous remark is that even though T and P vary with c, c itself is *not* expanded in a power series in ϵ as will be the case for the Hilbert expansion in the next subsection. Therefore T and P are also *not* expanded in ϵ. This is a very important point and will be used later in the way we define operators and prove results. For a Chapman–Enskog expansion, T depends on $c(= c^{(0)})$ from

$$T(f^{(0)}) \equiv T(c; f^{(0)}) = \tau \frac{\int f^{(0)}(v)\, dv}{c_{jam} - \int f^{(0)}(v)\, dv}. \tag{4.8}$$

To prove the main result in this section we first define the following auxiliary operator,

$$\begin{aligned}\mathbf{L}(g)h(v) := &-\frac{h(v) - \varphi_0(v)\int_0^{w^+} h(v')\, dv'}{T(g)} + (1 - P(g))h(v)\int_0^{w^+} (v' - v)g(v')\, dv' \\ &+ (1 - P(g))g(v)\int_0^{w^+} (v' - v)h(v')dv',\end{aligned} \tag{4.9}$$

and prove the following simple lemma.

LEMMA 4.1. *The null space of* $\mathbf{L}^*(f^{(0)})$ *consists of the constants.*

Proof. The adjoint of $\mathbf{L}(f^{(0)})$ is

$$\begin{aligned}\mathbf{L}^*(f^{(0)})h(v) = &(1 - P)c^{(0)}(\zeta(c^{(0)}) - v)[h(v)(1 - \alpha) + \alpha(h(\zeta))] \\ &+ \frac{v - \zeta(c^{(0)})}{T}\int_0^{w^+} \frac{\varphi_0(v')}{v' - \zeta(c^{(0)})} h(v')\, dv',\end{aligned} \tag{4.10}$$

and we therefore need to solve $\mathbf{L}^*(f^{(0)})h(v) = 0$. So we find functions $h(v)$ such that

$$(1 - P)c^{(0)}T[h(v)(1 - \alpha) + \alpha(h(\zeta))] = \int_0^{w^+} \frac{\varphi_0(v)}{v - \zeta(c^{(0)})} h(v)\, dv, \tag{4.11}$$

implying that $h(v)$ is constant. ∎

LEMMA 4.2 (necessary and sufficient condition). *There exists a solution of* (4.2) *with* (4.1) *if and only if for every order of the expansion* n *the following equations hold:*

$$\frac{\partial^{(n-1)}}{\partial t} c = -\frac{\partial q^{(n-1)}}{\partial x} \qquad \text{for } n = 1, 2, \ldots. \tag{4.12}$$

Proof. We substitute expansions (4.1), (4.4), (4.7) into (4.2). By the Fredholm alternative, (4.2) will have a solution if and only if the integral over $0 < v < w^+$ of the right-hand side is zero, or otherwise, if

$$\int_0^{w^+} \sum_{k=0}^{n-1} \frac{\partial^{(k)} f^{(n-k-1)}}{\partial t} + v\frac{\partial f^{(n-1)}}{\partial x}\, dv = 0. \tag{4.13}$$

The result now follows from (4.5), (4.6), and the definition of

$$q^{(n)} = \int_0^{w^+} v f^{(n)}(v)\, dv \qquad \text{for } n = 0, 1, \ldots. \tag{4.14}$$

∎

Note that (4.12) is nothing more than the Lighthill–Whitham–Richards [22, 31] equations, which, according to this lemma, must hold at every level of the expansion n in order for the problem to have a solution.

THEOREM 4.3. *The Chapman–Enskog expansion for the Prigogine–Herman equation at any order of ϵ has the form*

$$\frac{\partial c}{\partial t} + \sum_{n=0}^{\infty} \epsilon^n \frac{\partial q^{(n)}}{\partial x} = 0, \tag{4.15}$$

where $q^{(n)}(x,t;c) = \int_0^{w^+} v f^{(n)}(x,v,t;c)\, dv$.

For the proof of this theorem, but also the two corollaries that follow, we direct the interested reader to [26].

COROLLARY 4.4. *The zeroth-order approximation (ϵ^0) is the Lighthill–Whitham–Richards (LWR) model,*

$$\frac{\partial c}{\partial t} + \frac{\partial q}{\partial x} = 0, \tag{4.16}$$

where

$$q = q^{(0)} = Q_0(c) := \frac{1}{T(1-P)} + c\zeta, \tag{4.17}$$

with $\zeta(c)$ the root of $F_0(\zeta) = Tc(1-P)(1-\alpha)$, where $F_0(\zeta) := \int_{w^-}^{w^+} \frac{\varphi_0}{v-\zeta}\, dv$.

COROLLARY 4.5. *The first-order Chapman–Enskog approximation is given by*

$$\begin{aligned} \frac{\partial c}{\partial t} + \frac{\partial q}{\partial x} &= \frac{\partial c}{\partial t} + \frac{\partial \left[q_{CE}^{(0)} + q_{CE}^{(1)} \right]}{\partial x} \\ &= \frac{\partial c}{\partial t} + Q_0'(c) \frac{\partial c}{\partial x} - \frac{\partial}{\partial x} \left(\mathcal{D}(c) \frac{\partial c}{\partial x} \right) = 0. \end{aligned} \tag{4.18}$$

Here $Q_0' = \frac{dQ_0}{dc}$, where Q_0 is as in Corollary 4.4, and the "diffusion coefficient" is defined by

$$\mathcal{D}(c) = \frac{T}{F_1(\zeta(c))} \left(cT(1-P) \frac{F_2(\zeta(c))}{F_1(\zeta(c))^2} - 1 \right) \geq 0, \tag{4.19}$$

where $F_n(\zeta(c)) \equiv \int_0^{w^+} \frac{\varphi_0(v)}{(v-\zeta)^{n+1}} dv$.

4.1.1. Second-order (or Burnett-like) expansion.

COROLLARY 4.6. *The Burnett (order 2) approximation to the Prigogine–Herman equation is given by*

$$\begin{aligned} \frac{\partial c}{\partial t} + \frac{\partial q}{\partial x} &= \frac{\partial c}{\partial t} + \frac{\partial \left[q_{CE}^{(0)} + q_{CE}^{(1)} + q_{CE}^{(2)} \right]}{\partial x} \\ &= \frac{\partial c}{\partial t} + Q_0'(c) \frac{\partial c}{\partial x} - \frac{\partial}{\partial c} \left(\mathcal{D}(c) \frac{\partial c}{\partial x} \right) \frac{\partial c}{\partial x} + \frac{\partial}{\partial c} \left(I(c) \left(\frac{\partial c}{\partial x} \right)^2 \right) \frac{\partial c}{\partial x} = 0, \end{aligned} \tag{4.20}$$

where Q_0 and $\mathcal{D}$ are as in (4.17) and (4.19), respectively, and

$$A(c) = \frac{T}{F_1^3} \left(\frac{3F_1' F_2}{F_1} - F_2' \right),$$

$$B(c) = \frac{T}{F_1^3}\left(\frac{F_0^{'}F_2}{F_1} + 2\frac{F_2^{'}F_0}{F_1} - 6\frac{F_0F_1^{'}F_2}{F_1^2} - F_1^{'} - 2\zeta^{'}F_2\right),$$
$$G(c) = \frac{T}{F_1^3}\left(2\frac{F_0}{F_1}(F_1^{'} + \zeta F_2) + 3\frac{F_0}{c}\right),$$
$$H(c) = -3\frac{TF_0}{cF_1^4},$$
$$I(c) = \frac{T}{F_1}(AF_1 + BF_2 + GF_3 + HF_4) - \frac{T\mathcal{D}(c)F_0}{cF_1^3}\left(F_3 - \frac{F_2^2}{F_1}\right),$$

where $F_n(\zeta) := \int_0^{w^+} \frac{\varphi_0(v)}{(v-\zeta)^{n+1}}\, dv$, *and all primes denote derivatives with respect to* c.

Proof. The left-hand side of the Prigogine–Herman equation expanded in powers of a small parameter ϵ has the form

$$\epsilon\left[\frac{\partial f}{\partial t} + v\cdot\frac{\partial f}{\partial x}\right] = \epsilon\Bigg[\left(\frac{\partial^{(0)}}{\partial t} + \epsilon\frac{\partial^{(1)}}{\partial t} + \epsilon^2\frac{\partial^{(2)}}{\partial t} + O(\epsilon^3)\right)(f^{(0)} + \epsilon f^{(1)} + \epsilon^2 f^{(2)} + O(\epsilon^3)) + v\cdot\frac{\partial}{\partial x}(f^{(0)} + \epsilon f^{(1)} + \epsilon^2 f^{(2)} + O(\epsilon^3))\Bigg].$$

Thus the ϵ^2 terms are $\frac{\partial^{(0)}}{\partial t}f^{(1)} + \frac{\partial^{(1)}}{\partial t}f^{(0)} + v\cdot\frac{\partial}{\partial x}f^{(1)}$. Applying Lemma 4.2, there exists a solution if and only if

$$\frac{\partial^{(1)}}{\partial t}c = -\frac{\partial}{\partial x}q^{(1)}. \tag{4.21}$$

This, together with

$$\frac{\partial^{(0)}}{\partial t}c = -\frac{\partial}{\partial x}q^{(0)}, \tag{4.22}$$

which can be similarly derived while showing Corollary 4.5, will be proven useful in the formulations to follow.

Using the conservation law (4.16), or in effect relationship (4.21) and (4.22), we obtain

$$\frac{\partial^{(0)}}{\partial t}f^{(1)} + \frac{\partial^{(1)}}{\partial t}f^{(0)} + v\cdot\frac{\partial}{\partial x}f^{(1)} = \left(v - \frac{\partial q^{(0)}}{\partial c}\right)\frac{\partial f^{(1)}}{\partial c}\frac{\partial c}{\partial x} - \frac{\partial q^{(1)}}{\partial c}\frac{\partial f^{(0)}}{\partial c}\frac{\partial c}{\partial x} := g_2\left(c, \frac{\partial c}{\partial x}\right)\frac{\partial c}{\partial x}.$$

Similarly, the Chapman–Enskog expansion of the right-hand side of the Prigogine–Herman equation at order ϵ^2 provides us with

$$-f^{(2)}(1-P)c(v-\zeta) + (1-P)[f^{(0)}q^{(2)} + f^{(1)}q^{(1)}] := \mathcal{L}(f^{(0)}, f^{(1)})f^{(2)}. \tag{4.23}$$

When we solve the full approximation,

$$g_2\left(c, \frac{\partial c}{\partial x}\right)\frac{\partial c}{\partial x} = \mathcal{L}(f^{(0)}, f^{(1)})f^{(2)} = -f^{(2)}(1-P)c(v-\zeta) + (1-P)(f^{(0)}q^{(2)} + f^{(1)}q^{(1)}),$$

for $f^{(2)}$ we obtain

$$f^{(2)} = \frac{1}{(1-P)c(v-\zeta)}\left((1-P)\left(f^{(0)}q^{(2)} + f^{(1)}q^{(1)}\right) - g_2\frac{\partial c}{\partial x}\right). \tag{4.24}$$

Integrating with respect to v, we obtain the contribution to the flow as

$$0 = \frac{q^{(2)}}{c}\int_0^{w^+}\frac{f^{(0)}}{v-\zeta}\,dv + \frac{q^{(1)}}{c}\int_0^{w^+}\frac{f^{(1)}}{v-\zeta}\,dv - \frac{T}{F_0}\int_0^{w^+}\frac{g_2}{v-\zeta}\,dv\frac{\partial c}{\partial x}. \tag{4.25}$$

Some of the expressions appearing above are now evaluated here. Particularly, we simplify the three terms $\int_0^{w^+}\frac{f^{(0)}}{v-\zeta}\,dv$, $\int_0^{w^+}\frac{f^{(1)}}{v-\zeta}\,dv$, and $\int_0^{w^+}\frac{g_2}{v-\zeta}\,dv$ below. First note that

$$\int_0^{w^+}\frac{f^{(0)}}{v-\zeta}\,dv = c\frac{F_1}{F_0}. \tag{4.26}$$

Similarly,

$$\int_0^{w^+}\frac{f^{(1)}}{v-\zeta}\,dv = \frac{T}{F_1^2}\left(F_3 - \frac{F_2^2}{F_1}\right)\frac{\partial c}{\partial x}. \tag{4.27}$$

Last we calculate

$$g_2(v) = \left(A\frac{\varphi(c)}{(v-\zeta)} + B\frac{\varphi(c)}{(v-\zeta)^2} + G\frac{\varphi(c)}{(v-\zeta)^3} + H\frac{\varphi(c)}{(v-\zeta)^4}\right)\frac{\partial c}{\partial x}, \tag{4.28}$$

where $A = A(c)$, $B = B(c)$, $G = G(c)$, and $H = H(c)$ are as in the statement of the theorem. As a result,

$$\int_0^{w^+}\frac{g_2(v)}{v-\zeta}\,dv = (AF_1 + BF_2 + GF_3 + HF_4)\frac{\partial c}{\partial x}. \tag{4.29}$$

Therefore the second approximation to the flow, $q^{(2)}$, can now be expressed from (4.25) through expressions (4.26), (4.27), and (4.29), as $q^{(2)} = I(c)(\frac{\partial c}{\partial x})^2$, where $I(c)$ is as in the theorem. Thus the second-order hydrodynamic approximation, given the flow $q = q^{(0)} + q^{(1)} + q^{(2)}$, becomes

$$\begin{aligned} 0 &= \frac{\partial c}{\partial t} + \frac{\partial q}{\partial x} \\ &= \frac{\partial c}{\partial t} + Q_0'(c)\frac{\partial c}{\partial x} - \frac{\partial}{\partial c}\left[\mathcal{D}(c)\frac{\partial c}{\partial x} - I(c)\left(\frac{\partial c}{\partial x}\right)^2\right]\frac{\partial c}{\partial x}. \end{aligned}$$ □

Some very simple nondimensional analysis reveals that we have obtained the correct form in this complicated-looking flux function. If we abbreviate quantities such as length (L) and time (t), we have that $\mathcal{D}(c)$ is represented in L^2/t, $\partial c/\partial x$ in $1/L^2$, and $I(c)$ in t^4/L^3. We therefore obtain the extra contribution to the flow $q^{(2)}$ in $1/t$ as expected.

4.2. Linear stability analysis. The question of stability for the Chapman–Enskog expansions just presented is natural, especially in light of the fact that Burnett level expansions for fluid dynamic problems are destabilizing [14]. We therefore now perform a linear stability analysis. A brief outline of this procedure is as follows: We start by linearizing (4.20). We then obtain $\nu_k(\omega)$ the Fourier–Laplace transform for plane (couette) flow of $\hat{c}(x,t)$. (The "hat" on the c represents small perturbations from equilibrium, c_{eq}, in space and time.) As a result, the inverse Laplace transform of $\nu_k(\omega)$ gives contributions in the form

$$Ae^{-i\omega(k)t}, \tag{4.30}$$

where A is the amplitude. Therefore our interest lies in identifying the sign of the imaginary part of ω. It is clear from (4.30) that a positive (negative) value of $\mathrm{Im}(\omega)$ implies instability (stability).

To linearize (4.20) we need to introduce some notation. We define the nonlinear differential operator by

$$F(c) = f(t,x,D^m c) = \frac{\partial c}{\partial t} + \frac{\partial}{\partial x}\left[Q_0(c) - \mathcal{D}(c)\frac{\partial c}{\partial x} + I(c)\left(\frac{\partial c}{\partial x}\right)^2\right]. \tag{4.31}$$

Suppose that $f(t,x,\xi)$ is smooth in its arguments $t \in R^+, x \in \Omega \subset R$, and $\xi = \{\xi_\alpha : |\alpha| \le m\}$, where c can take values on some vector space R and $F : C^\infty(\Omega) \to C^\infty(\Omega)$. We let the equilibrium solution $c_{eq} \in C^m(\omega)$. Then the linearization of F at c_{eq} is $LF(c_{eq}) : C^m(\Omega) \to C(\Omega)$,

$$LF(c_e q)\hat{c} = \frac{\partial}{\partial \epsilon}F(c_{eq} + \epsilon\hat{c})|_{\epsilon=0} = \sum_{|\beta|\le m}\frac{\partial f}{\partial \xi_\beta}(t,x,D^m c_{eq})D^\beta \hat{c}. \tag{4.32}$$

As a result, linearizing (4.20) gives

$$\frac{\partial \hat{c}}{\partial t} + P_1\hat{c} + P_2\frac{\partial \hat{c}}{\partial x} + P_3\frac{\partial^2 \hat{c}}{\partial x^2} = 0, \tag{4.33}$$

where

$$P_1 = Q_0^{''}\frac{\partial c_{eq}}{\partial x} - \mathcal{D}^{''}\left(\frac{\partial c_{eq}}{\partial x}\right)^2 - \mathcal{D}^{'}\frac{\partial^2 c_{eq}}{\partial x^2} + I^{''}\left(\frac{\partial c_{eq}}{\partial x}\right)^3 + 2I^{'}\frac{\partial c_{eq}}{\partial x}\frac{\partial^2 c_{eq}}{\partial x^2},$$

$$P_2 = Q_0^{'} - 2\mathcal{D}^{'}\frac{\partial c_{eq}}{\partial x} + 3I^{'}\left(\frac{\partial c_{eq}}{\partial x}\right)^2 + 2I\frac{\partial^2 c_{eq}}{\partial x^2}, \quad \text{and} \quad P_3 = -\mathcal{D} + 2I\frac{\partial c_{eq}}{\partial x},$$

where all primes denote derivatives with respect to ξ and all coefficients are evaluated at c_{eq}. We now introduce the Fourier–Laplace transform

$$\hat{c}(t;x) = \frac{1}{2\pi}\int_0^\infty e^{i\omega(k)t}\int_{-\infty}^\infty e^{ikx}\nu_k(v;\omega)\,dk\,d\omega. \tag{4.34}$$

Therefore (4.33) gives $\omega(k) = -P_1 i + P_2 k + P_3 k^2 i$, which implies that we must have

$$-P_1 + P_3 k^2 < 0 \quad \text{for stability.} \tag{4.35}$$

In Figure 3 we see, numerically, the contribution to stability due to the advective part alone. Note that in fact we have instability for this case (positive value) but no

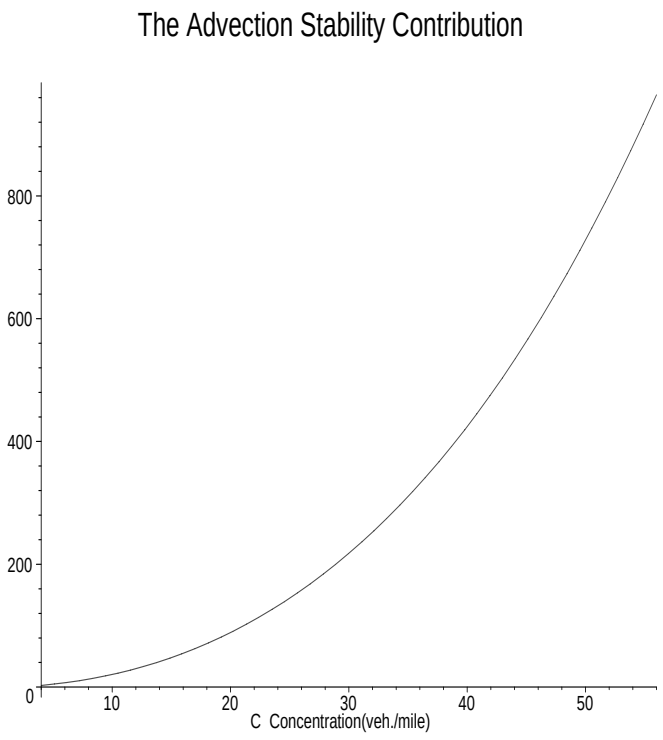


FIG. 3. *The stability contribution term* $-Q_0'' \frac{\partial c_{eq}}{\partial x}$ *due solely to advection, for representative values of* $\frac{\partial c_{eq}}{\partial x}$ *for our traffic release problem. Note that, even though the values are positive, they remain finite for all values of the concentration* c *up to* $c_{crit} \approx 56$*. Also note the exponential increase of instability with concentration.*

blow-up, and thus the amplitude A of (4.30) is finite and controlled for concentrations in the range $0 \leq c \leq c_{crit}$. Also note that, as the concentration increases, stability becomes exponentially difficult to maintain. Similarly, we plot in Figure 4 the stability due to advection and diffusion terms. Here we notice that, due to the inclusion of the diffusion part, we have stability (negative value), as would be expected. At the end of the concentration range we observe small positive (unstable) contributions but still no blow-up. Notice that stability remains, but instability grows logarithmically with concentration. Last, in Figure 5 we plot the (in)stability of the complete system $-P_1 + P_3 k^2$. Based on these three figures, it is clear that the instability is attributed to "Burnett" terms contributions. Therefore, once again we observe, similarly to fluid dynamics problems, that the Burnett expansion is actually destabilizing for all time! In fact we get a blow-up to stability immediately even for very low concentrations. That can also be seen numerically in Figure 6, where we try to show the blow-up in solutions even at times as small as $t = 10^{-13}$. (See section 5.3 regarding details on numerics.)

4.3. The Hilbert asymptotic expansion. The Hilbert [12] expansion of the Prigogine–Herman equation is very similar in form to the Chapman–Enskog expansion considered above.

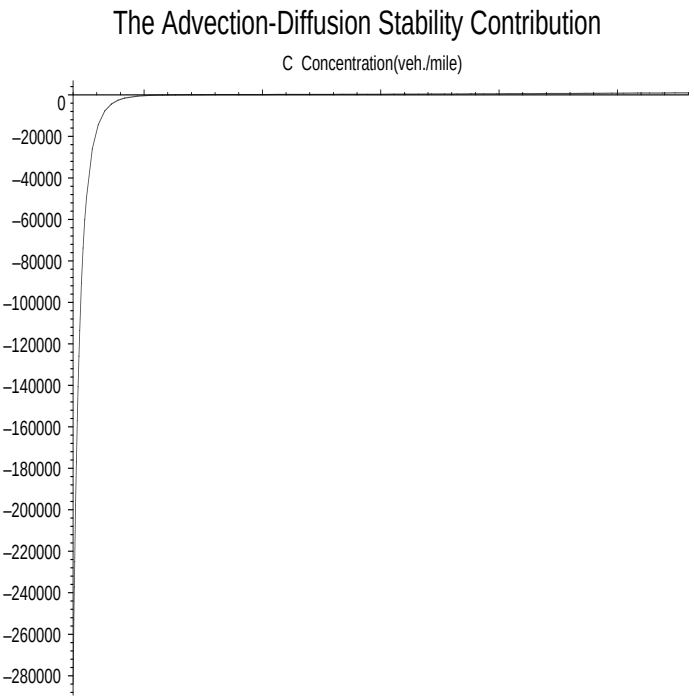


FIG. 4. *The stability contribution term* $-Q_0^{''}\frac{\partial c_{eq}}{\partial x}+\mathcal{D}^{''}(\frac{\partial c_{eq}}{\partial x})^2+\mathcal{D}^{'}\frac{\partial^2 c_{eq}}{\partial x^2}-\mathcal{D}$ *due to advection and diffusion for representative values of* $\frac{\partial c_{eq}}{\partial x}$ *for our traffic release problem. Note that it starts negative and therefore stable while logarithmically increasing but remains finite for all values of the concentration* c *up to* $c_{crit}\approx 56$.

THEOREM 4.7. *The Hilbert hydrodynamic approximation to the concentration* c *for the Prigogine–Herman equation has the form*

$$\frac{\partial c^{(m)}}{\partial t}+\frac{\partial q^{(m)}}{\partial x}=0, \qquad m=0,1,\dots,n, \tag{4.36}$$

where

$$q^{(m)}=Q_m^H(c^{(0)},c^{(1)},\dots,c^{(m)}) \tag{4.37}$$

is known at order n *of the expansion.*

We want to solve this system of $n+1$ partial differential equations for $n+1$ unknowns.

Proof. The proof is similar to the one given for the Chapman–Enskog expansion. However, a major difference is that the expansion of the concentration in powers of the artificially introduced small parameter ϵ will no longer be set to zero for $n\geq 1$. The flow $q^{(m)}$ will be, at least in principle, known in terms of $c^{(0)},c^{(1)},\dots,c^{(m)}$. The macroscopic form of the nth-order Hilbert approximation is the system for $m=0,1,\dots,n$. The corresponding approximation for the density and flow are, respectively, $c=c^{(0)}+c^{(1)}+\cdots+c^{(n)}$ and $q=q^{(0)}+q^{(1)}+\cdots+q^{(n)}$.

We substitute (4.1), (4.7) into (4.2), and for different powers of ϵ we obtain

$$Qf^{(0)}=0 \qquad \text{for } \epsilon^0, \tag{4.38}$$

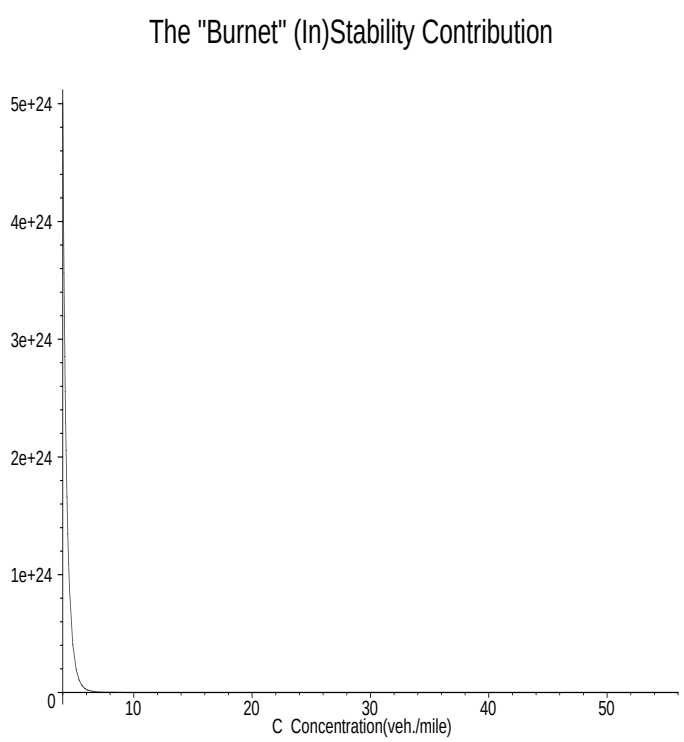


FIG. 5. *The total stability contribution,* $-P_1 + P_3$*, for representative values of* $\frac{\partial c_{eq}}{\partial x}$ *and* $\frac{\partial^2 c_{eq}}{\partial x^2}$ *for our traffic release problem. Note that it is positive and therefore* unstable *for all values of the concentration* c *up to* $c_{crit} \approx 56$*. It is clear from this that the extra "Burnett" approximation determines the stability and dominates the advection diffusion term contributions that we have observed thus far.*

$$\mathbf{L}(f^{(0)})f^{(n)} = \sum_{k=0}^{n-1} \frac{\partial^{(k)} f^{(n-k-1)}}{\partial t} + v\frac{\partial f^{(n-1)}}{\partial x} \quad \text{for } \epsilon^n, \quad \text{where } n = 1, 2, \ldots.$$

Again it is necessary that $\frac{\partial f^{(n-1)}}{\partial t} + v\frac{\partial f^{(n-1)}}{\partial x}$ be orthogonal to the null space of the adjoint, $\mathbf{L}(f^{(0)})^*$. We know that the null space of $\mathbf{L}(f^{(0)})^*$ consists of constants. Therefore the compatibility condition implies again that a solution $f^{(n)}$ exists if and only if $\int_0^{w^+} (\frac{\partial f^{(n-1)}}{\partial t} + v\frac{\partial f^{(n-1)}}{\partial x})\, dv = 0$. This provides the result since it implies the law of conservation of vehicles,

$$\frac{\partial c^{(n)}}{\partial t} + \frac{\partial q^{(n)}}{\partial x} = 0 \qquad \text{for } n = 0, 1, \ldots. \qquad \square \tag{4.39}$$

4.3.1. Zero-order Hilbert expansion.

COROLLARY 4.8. *The zero-order Hilbert asymptotic expansion of the Prigogine–Herman equation has the form*

$$\frac{\partial c^{(0)}}{\partial t} + \frac{\partial q^{(0)}}{\partial c^{(0)}}\frac{\partial c^{(0)}}{\partial x} = 0, \tag{4.40}$$

where

$$q^{(0)}(v) = \int_0^{w^+} v f^{(0)}(v)\, dv = \frac{1}{T(1-P)} + c^{(0)}\zeta(c^{(0)}), \tag{4.41}$$

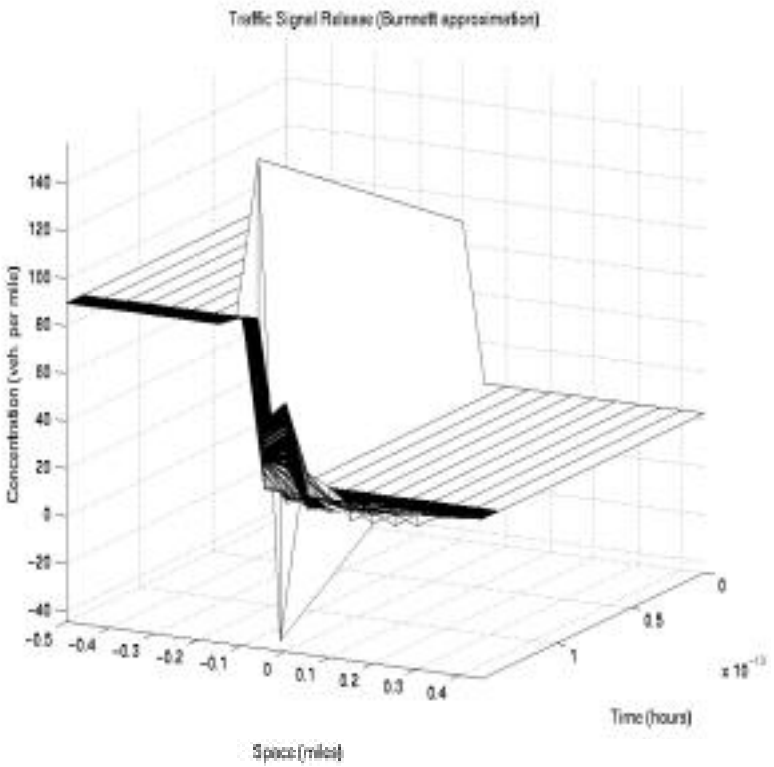


FIG. 6. *Chapman–Enskog solution using Godunov's method for the second-order approximation (Burnett equivalent) equation. The solution blows up (note the nonphysical negative concentrations) even for very small times.*

where T and P will be assumed constant here.

Proof. We solve the right-hand side of (2.1),

$$Q(f^{(0)}) = 0, \tag{4.42}$$

and obtain

$$f^{(0)}(v) = \frac{1}{T(1-P)} \frac{\varphi_0(v)}{v - \zeta(c^{(0)})}. \tag{4.43}$$

Thus,

$$c^{(0)} = \int_0^{w^+} f^{(0)}(v)\, dv = \frac{1}{T(1-P)} \int_0^{w^+} \frac{\varphi_0}{v - \zeta(c^{(0)})}\, dv, \tag{4.44}$$

which gives a simplifying relation we have used many times before in the Chapman–Enskog expansion, with the significant difference that now $c^{(0)} \neq c$,

$$\int_0^{w^+} \frac{\varphi_0}{v - \zeta(c^{(0)})}\, dv = T(1-P)c^{(0)} := F_0(\zeta(c^{(0)})). \tag{4.45}$$

As a result we can write

$$q^{(0)} = \frac{1}{T(1-P)} \int_0^{w^+} v \frac{\varphi_0(v)}{v - \zeta(c^{(0)})}\, dv = \frac{1}{T(1-P)} + c^{(0)}\zeta(c^{(0)}). \tag{4.46}$$

For $f^{(0)}$ to exist, it must satisfy the condition

$$\int_0^{w^+} \frac{\partial f^{(0)}}{\partial t} + v\frac{\partial f^{(0)}}{\partial x}\, dv = 0, \tag{4.47}$$

which gives the zero-order hydrodynamic approximation as

$$\frac{\partial c^{(0)}}{\partial t} + \frac{\partial q^{(0)}}{\partial x} = 0. \qquad \square \tag{4.48}$$

A crucial point that must be emphasized here is that not only will the zero-order Chapman–Enskog expansion and the zero-order Hilbert expansion both give an LWR model, but they give the *same* LWR model! This is not always so obvious in higher-order expansions, as we will soon see.

4.3.2. First-order Hilbert expansion. We now carry out the evaluation of f up to two terms, $f^{(0)}$ and $f^{(1)}$.

COROLLARY 4.9. *The first-order Hilbert approximation to the concentration is given by* (4.40) *and*

$$\frac{\partial c^{(1)}}{\partial t} + \frac{\partial q^{(1)}}{\partial c^{(1)}}\frac{\partial c^{(1)}}{\partial x} + \frac{\partial q^{(1)}}{\partial c^{(0)}}\frac{\partial c^{(0)}}{\partial x} = 0, \tag{4.49}$$

where

$$f^{(1)}(v) = \frac{c^{(1)}}{F_1(\zeta(c^{(0)}))}\frac{\varphi_0(v)}{v - \zeta(c^{(0)})^2}, \tag{4.50}$$

with $c^{(1)} = CF_1(\zeta(c^{(0)}))$ *for* C *an arbitrary constant and*

$$q^{(1)} = c^{(1)}\left\{\frac{F_0(\zeta(c^{(0)}))}{F_1(\zeta(c^{(0)}))} + \zeta(c^{(0)})\right\},$$

where T *and* P *will be assumed constant here.*

Proof. We look for a solution (at first order of the series expansion in ϵ) $f^{(1)}$ of (2.1) which has the form $f_p^{(1)} + f_h^{(1)}$, where $f_p^{(1)}$ is the particular solution and $f_h^{(1)}$ is the homogeneous solution of

$$\mathbf{L}(f^{(0)})f^{(1)} = \frac{\partial f^{(0)}}{\partial t} + v\frac{\partial f^{(0)}}{\partial x}, \tag{4.51}$$

where

$$\begin{aligned}\mathbf{L}(g)h(v) = &-\frac{h(v) - \varphi_0(v)\int_0^{w^+} h(v')\, dv'}{T} + (1-P)h(v)\int_0^{w^+}(v'-v)g(v')\, dv' \\ &+ (1-P)g(v)\int_0^{w^+}(v'-v)h(v')\, dv'.\end{aligned}$$

To obtain the homogeneous solution we start by simplifying:

$$
\begin{aligned}
\mathbf{L}(f^{(0)})f_h^{(1)}(v) = {} & \frac{-f_h^{(1)}}{T} + \frac{\varphi_0(v)}{T}\int_0^{w^+} f_h^{(1)}(v)\,dv + \frac{1}{T}f_h^{(1)}(v)\int_0^{w^+}\frac{v^{'}\varphi_0(v^{'})}{(v^{'}-\zeta(c^{(0)}))}\,dv^{'} \\
& - \frac{vf_h^{(1)}(v)}{T}\int_0^{w^+}\frac{\varphi_0(v^{'})}{(v^{'}-\zeta(c^{(0)}))}\,dv^{'} + \frac{1}{T}\frac{\varphi_0(v)}{(v-\zeta(c^{(0)}))}\int_0^{w^+} v^{'}f_h^{(1)}(v^{'})\,dv^{'} \\
& - \frac{1}{T}\frac{v\varphi_0(v)}{(v-\zeta(c^{(0)}))}\int_0^{w^+} f_h^{(1)}(v^{'})\,dv^{'} \\
= {} & \frac{\zeta(c^{(0)})}{(v-\zeta(c^{(0)}))}\frac{\varphi_0(v)}{T}\int_0^{w^+} f_h^{(1)}(v^{'})\,dv^{'} + \frac{f_h^{(1)}(v)}{T}(\zeta(c^{(0)})-v)Tc(1-P) \\
& + \frac{1}{T}\frac{\varphi_0(v)}{v-\zeta(c^{(0)}}\int_0^{w^+} v^{'}f_h^{(1)}(v^{'})\,dv^{'} \\
= {} & (1-P)c^{(0)}(\zeta(c^{(0)})-v)f_h^{(1)}(v) + \frac{\varphi_0(v)}{T}\int_0^{w^+}\frac{v^{'}-\zeta(c^{(0)})}{v-\zeta(c^{(0)})}f_h^{(1)}(v^{'})\,dv^{'}.
\end{aligned}
$$

We define $K := \int_0^{w^+}(v^{'}-\zeta(c^{(0)})f_h^{(1)}(v^{'})\,dv$, and we solve the equation above for $f_h^{(1)}$:

$$
f_h^{(1)} = \frac{K}{F_0(\zeta(c^{(0)}))}\frac{\varphi_0(v)}{(v-\zeta(c^{(0)}))^2}. \tag{4.52}
$$

A specific description of K can be obtained by simply enforcing the normalization condition (3.3). This gives $K = \frac{c^{(1)}F_0(\zeta(c^{(0)}))}{F_1(\zeta(c^{(0)}))}$. As a result,

$$
f_h^{(1)} = \frac{c^{(1)}}{F_1(\zeta(c^{(0)}))}\frac{\varphi_0(v)}{(v-\zeta(c^{(0)}))^2}. \tag{4.53}
$$

It is convenient to take $f_p^1(v)$ as the solution that satisfies $\int_0^{w^+} f_p^{(1)}(v)\,dv = 0$. That is, the concentration contribution comes only from the homogeneous solution of the equation. Therefore from $\int_0^{w^+}\frac{\partial f^{(0)}}{\partial t} + v\frac{\partial f^{(0)}}{\partial x}\,dv = 0$ we also get $\int_0^{w^+} vf_p^{(1)}(v)\,dv = 0$. Thus,

$$
f^{(1)}(v) = f_p^{(1)}(v) + \frac{c^{(1)}}{F_1(\zeta(c^{(0)}))}\frac{\varphi_0(v)}{(v-\zeta(c^{(0)}))^2}. \tag{4.54}
$$

Therefore the values of $c^{(1)}$ and $q^{(1)}$ are related by

$$
q^{(1)} = c^{(1)}\left\{\frac{F_0(\zeta(c^{(0)}))}{F_1(\zeta(c^{(0)}))} + \zeta(c^{(0)})\right\}, \tag{4.55}
$$

where it is clear here that $= q^{(1)}(c^{(0)}, c^{(1)})$. Thus we have

$$
\frac{\partial c^{(1)}}{\partial t} + \frac{\partial Q^{(1)}}{\partial x} = 0, \tag{4.56}
$$

where $Q^{(1)} = q^{(1)}(c^{(0)}, c^{(1)})$; in effect the first-order model is really a system of two equations in two unknowns. More specifically, it is a system of equations (4.40) and (4.49) with $c^{(0)}$ and $c^{(1)}$ as the unknowns. This is the central difference between the Hilbert and Chapman–Enskog expansions. The complete Hilbert first-order approximation for this order of the expansion is

$$
\begin{aligned}
c^{[1]} &= c^{(0)} + c^{(1)}, \\
q^{[1]} &= q^{(0)} + q^{(1)}. \qquad \square
\end{aligned}
$$

Similarly the nth-order model will be a system of n equations in n unknowns.

5. Brief simulations. We now give some preliminary numerical simulations of solutions to our model equations. We only do this here for the zeroth- and first-order Chapman–Enskog expansion approximations since this section is meant to give only a flavor of the type of solutions we can obtain. For a complete investigation, which will include the second-order approximation and analysis under different traffic situations and different numerical techniques, we refer to the sequel [32] to this paper, which is to follow. For the examples that follow here we will leave most of the numerical details to [32].

5.1. Entropy. It is necessary [21] for uniqueness models, such as the LWR model [22, 31], to require that their solutions satisfy more than just the conservation equation and associated initial/boundary conditions. Solutions must also satisfy [17, 23] the so-called entropy condition. (See Ansorge [1] for a traffic-theoretic interpretation of the entropy condition.) The entropy condition states [2] (in one of its many versions) that the shock speed s is restricted by

$$q^{'}(c_l,t) < s < q^{'}(c_r,t) \tag{5.1}$$

for $c_l > c_r$, where q is as usual, for traffic flow, a concave function. For the numerical methods that follow we use a numerical flux function q^*, which implements (5.1) as follows:

$$q^*(c_l,c_r) = \begin{cases} \min\limits_{c_l \le c \le c_r} q(c) \text{ for } c_l \le c_r, \\ \max\limits_{c_r \le c \le c_l} q(c) \text{ for } c_r < c_l. \end{cases} \tag{5.2}$$

5.2. Conditions. We construct a traffic flow example for the algorithms in this work. The value of the parameter $\tau := .003$ hours ≈ 11 seconds is used, as obtained from data [7]. The reduced desired speed distribution used, φ_0, corresponds to a uniform distribution of desired speeds from 40 to 80 mph and 0 elsewhere. The corresponding equilibrium solution from (3.7), in the stable regime, is

$$f_{eq}(v;c) = \begin{cases} cv_0(c)\frac{\varphi_0(v;c)}{v-\zeta} \text{ for 40 mph } \le v \le 80 \text{ mph}, \\ 0 \text{ otherwise}, \end{cases} \tag{5.3}$$

where c is the density in vehicles per mile per lane (vpmpl).

The problem considered here is defined by the parameters of the preceding paragraph and the following initial conditions:

$$c(x,0) = c(x,t)|_{t=0} = \begin{cases} 59(\approx c_{crit}) \text{ vpmpl}, & x \le 0, \\ 4 \text{ vpmpl}, & x > 0. \end{cases}$$

This corresponds to release into a relatively vacant region at $t = 0$ of a semi-infinite "platoon" of vehicles extending indefinitely to the left from $x = 0$, and initially at the critical concentration. In that respect, Dirichlet-type boundary conditions are implemented in the schemes at the ends of the spatial interval. In fact, the spatial interval has been chosen so that under the given time of consideration of the development of the wave fronts there will be no interaction with the boundaries. The resulting solution should have the form of an "acceleration wave," which is analogous to a rarefaction wave in gas dynamics [21]. Such flows are often termed "queue discharge" in traffic engineering.

The corresponding exact solution to the LWR model (4.16) is a traveling wave, moving to the left, with density $c_\ell = 59$ vpmpl on the left, and density $c_r = 4$ vpmpl on the right. The wave thus propagates to the right (downstream) at a speed s (also called propagation speed) $(q^{(0)}(c_r) - q^{(0)}(c_\ell))/(c_r - c_\ell) = 57.7$ mph initially, and varies accordingly as the concentration changes after each subsequent approximation.

5.3. Godunov's method for the stable flow. Godunov's scheme is the preferred method [18, 19, 21, 2] for solving hyperbolic conservation laws such as (4.16). Originally the method was devised to solve inviscid Euler equations in one dimension. However, here we also implement Godunov's method [21] for solving the higher-order hydrodynamic approximation equations (as produced by the Chapman–Enskog first-order expansion), even though the method is used mostly for hyperbolic problems. This is feasible, since, for instance, we can reformulate the first-order model (4.18) as

$$\frac{\partial c}{\partial t} + \frac{\partial \left(Q_0(c) - \mathcal{D}(c)\frac{dc}{dx}\right)}{\partial x} = 0, \tag{5.4}$$

where now the flux function is $q = Q_0(c) - \mathcal{D}(c)\frac{dc}{dx}$. We apply Godunov's method to this (5.4) form of the advection-diffusion equation with this more elaborate flux function. Note that, numerically, $q = Q(c_i^n, c_{i+1}^n)$, as seen immediately below. This gives a natural way to define $\frac{dc}{dx} = \frac{c_{i+1}^n - c_i^n}{x_{i+1} - x_i}$ for each time step n.

The Godunov scheme in general is implemented through

$$c_i^{n+1} = c_i^n - \frac{\Delta t_n}{h_i}\left(Q(c_i^n, c_{i+1}^n) - Q(c_{i-1}^n, c_i^n)\right), \tag{5.5}$$

where the c_i^n's are approximations to the concentration c at different spatial (i) and time (n) intervals. Any numerical approximation of the form (5.5) is said to be *conservative*. Here $h_i = x_{i+1/2} - x_{i-1/2}$, and Q is a numerical approximation to the average flow q past the section boundary $x_{i+1/2}$ during the time interval $[t_n, t_{n+1}]$,

$$Q(c_i^n, c_{i+1}^n) \approx \frac{1}{\Delta t_n}\int_{t_n}^{t_{n+1}} q(c(x_{i+1/2}, t))\, dt. \tag{5.6}$$

We use relation (5.2) in evaluating the integral above (see [21, 2]). The Courant, Friedrichs, and Lewy (CFL) condition is used as a way to ensure that solutions do not blow up.

We plot the solution of the zero-order (LWR) model in Figure 7. The results of this diffusive equation, as can be seen in Figure 8, show the expected behavior for the concentration. The diffusive action can be seen in the way that the wave front widens in the space dimension as time advances. This diffusive effect is even more evident in Figure 9, where we plot the final solution of the front wave for the Godunov scheme for the advection-diffusion equation together with the corresponding final solution of the approximate and initial solutions for the LWR equation. For this figure the computational parameters used are $\Delta x = .016$ miles, while Δt varies as specified from the CFL condition.

Godunov's method, which was originally designed to solve hyperbolic problems, seems to produce a solution which looks meaningful (and stays bounded) even though we use it on an advection-dominated diffusion equation. The effect on the time step from the contribution due to the diffusion coefficient D is seen through the CFL condition:

$$\Delta t \le \frac{\Delta x}{\frac{\partial Q_0}{\partial c} - \frac{\partial \mathcal{D}}{\partial c}\frac{dc}{dx}}. \tag{5.7}$$

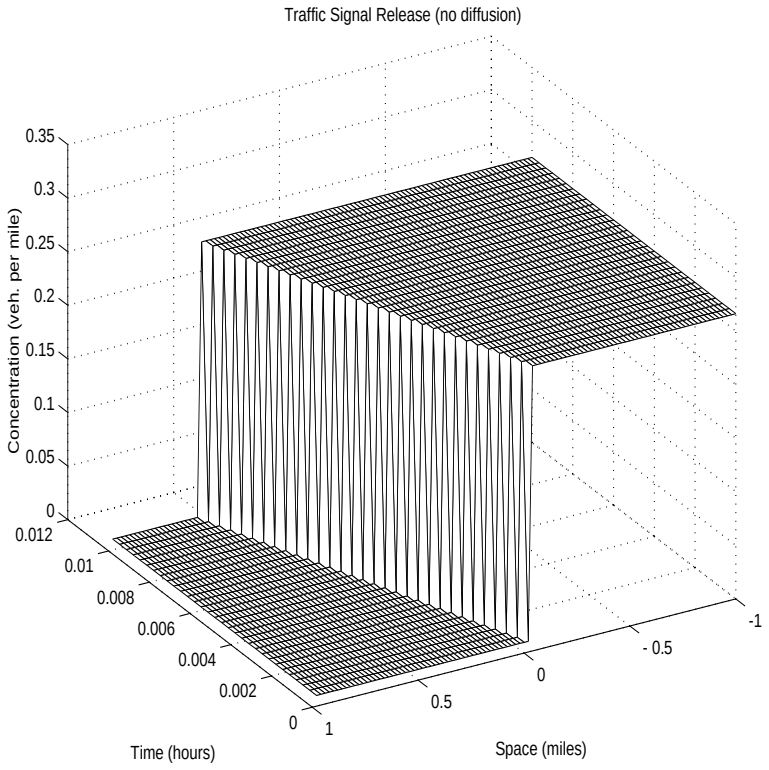


FIG. 7. *Solution using Godunov's method for the advection equation. (Concentration is normalized.)*

Note that the denominator of the quantity above stays positive, since $\frac{dc}{dx} \leq 0$, with a minimum value of $\frac{\partial Q_0}{\partial c}$ for cases in which the concentration is unchanged. In fact, the contributions to the time step due to $\mathcal{D}$ are minimal since, for the traffic parameters involved in the implementation of this example, $0 \leq \frac{\partial \mathcal{D}}{\partial c} < .003$ for the stable concentration regime, and the denominator is therefore dominated by $50 < \frac{\partial Q_0}{\partial c} < 60$. For the extreme cases of $\frac{dc}{dx} = \frac{c_{min} - c_{crit}}{\Delta x}$ the time step suffers a temporary setback, which, however, will be quickly remedied in the next time step from the diffusion of the concentration.

We expect to see that the original shock concentration will eventually "diffuse" and spread in front of and behind the shock formation, so that the concentration in the front of the shock will increase and the concentration behind it will decrease. The solution produced by our method exhibits this type of behavior. Also the scheme seems to behave well in not producing any erroneous results such as infinite or negative concentrations.

6. Conclusions. We have seen how traffic flow models can be obtained from the equilibrium solution of the Prigogine–Herman nonlinear kinetic equation. We rely on a *deterministic* equilibrium solution of the form (3.7), which is suitable for expansion calculations (but also predicts traffic flow for the complete range of possible traffic concentrations, unlike most other such kinetic equations [25, 33]). As a result, we obtain here a number of different models of flow, depending on the order of

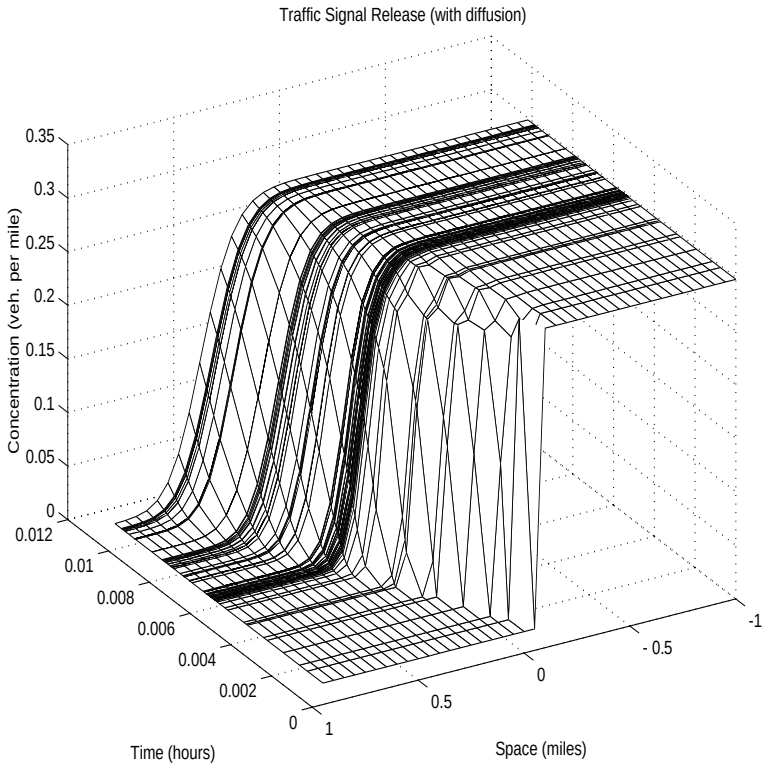


FIG. 8. *Solution using Godunov's method for the advection-diffusion equation. (Concentration is normalized.)*

series expansion carried out. This derivation is based solely on (formal) mathematical principles, which close our equations without any need for ad hoc assumptions.

Chapman–Enskog and Hilbert expansions are presented for zeroth- and first-order approximations. It is rewarding that at the lowest possible approximation level (zeroth) we obtain the well-known LWR model under either type of expansion. For the first-order model we obtain two seemingly different looking equations. In theory, however, the Chapman–Enskog and Hilbert expansions are supposed to produce models which can be shown to be the same. Since numerical results are forthcoming in a sequel to this paper [32], we will leave this question to be partially resolved there. Finally, the second-order Chapman–Enskog expansion is carried out to produce a Burnett equivalent model for traffic flow. Note, however, that, given the stability analysis results of section 4.2, this level of the expansion will not produce a useful model. It is possible that techniques similar to those implemented in [14] will be useful in that respect. The formal mathematical derivation obtained here becomes even more interesting in light of recent work [3] of such higher-order models. In [3] the correlation between car-following models such as [11] and [20] and the Kortweg–de Vries equation is established, and it is shown that the well-known "stop and go" traffic effect can be triggered by a car following equation that has the form of our second-order "Burnett-like" model or its possible successor based on techniques from [14]. In this paper we therefore hint at a mathematical link between empirical models (such as [3, 20]), which shows the relationship of model parameters to physically meaningful variables.

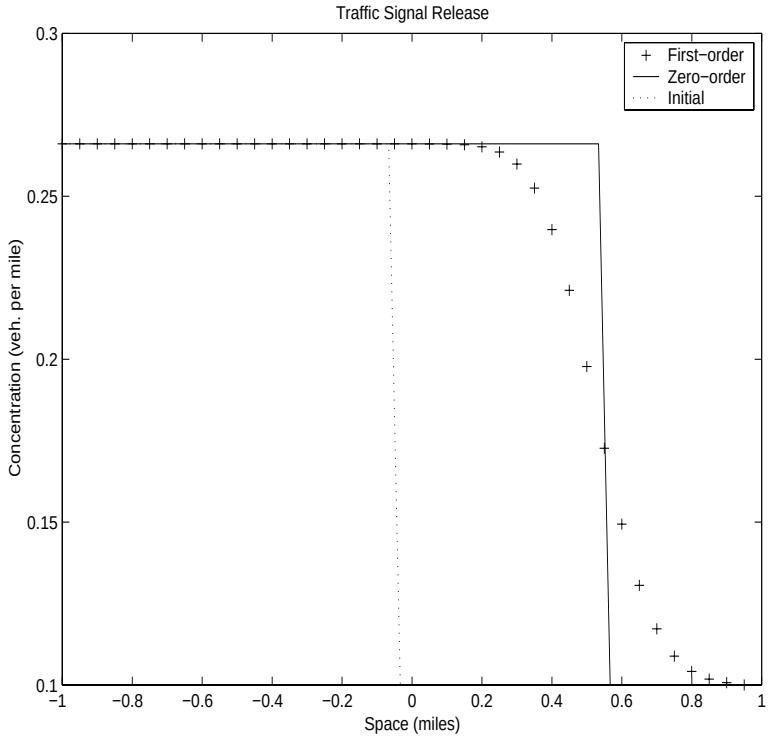


FIG. 9. *Comparison of the first-order and zero-order model solution based on Godunov's scheme.*

REFERENCES

[1] R. ANSORGE, *What does the entropy condition mean in traffic flow theory?*, Transportation Res. Part B, 24 (1990), pp. 133–143.

[2] D. D. BUI, P. NELSON, AND S. NARASIMHAN, *Computational Realizations of the Entropy Condition in Modeling Congested Traffic Flow*, Report No. FHWA/TX-92/1232-7, Texas Transportation Institute, College Station, TX, 1992.

[3] P. BERG AND A. WOODS, *On-ramp simulations and solitary waves of a car-following model*, Phys. Rev. E, 64 (2001), paper 035602.

[4] C. CERCIGNANI, *The Boltzmann Equation and Its Applications*, Springer-Verlag, New York, 1988.

[5] S. CHAPMAN, *On the kinetic theory of a gas. Part* II*—A composite monoatomic gas: Diffusion, viscosity, and thermal conduction*, Philos. Trans. Roy. Soc. London Ser. A, 217 (1918), pp. 115–197.

[6] S. CHAPMAN AND T. G. COWLING, *The Mathematical Theory of Non-uniform Gases; An Account of the Kinetic Theory of Viscosity, Thermal Conduction, and Diffusion in Gases*, Cambridge University Press, Cambridge, UK, 1939.

[7] L. C. EDIE, R. HERMAN, AND T. N. LAM, *Observed multilane speed distribution and the kinetic theory of vehicular traffic*, Transportation Res., 9 (1980), pp. 225–235.

[8] D. ENSKOG, *Kinetische theorie der vorgänge in mäßig verdünnten gasen*, Ph.D. thesis, Uppsala, 1917.

[9] D. HELBING, *Theoretic foundation of macroscopic traffic models*, Phys. A, 219 (1995), pp. 375–390.

[10] D. HELBING, *Gas kinetic derivation of Navier–Stokes-like traffic equations*, Phys. Rev. E, 53 (1996), pp. 2366–2381.

[11] D. HELBING, A. HENNECKE, AND M. TREIBER, *Phase diagram of traffic states in the presence of inhomogeneities*, Phys. Rev. Lett., 82 (1999), pp. 4360–4363.

[12] D. HILBERT, *Grundzüge einer Allgemeinen Theorie der Linearen Integralgleichungen*, Teubner, Vienna, 1924.

[13] S. P. Hoogendoorn and P. H. L Bovy, *Modeling multiple user-class traffic flow*, Transportation Res., 34 (2000), pp. 123–146.
[14] S. Jin and M. Slemrod, *Regularization of the Burnett equations for rapid granular flows via relaxation*, Phys. D, 150 (2001), pp. 207–218.
[15] B. S. Kerner, *The physics of traffic*, Physics World, August (1999), pp. 25–30.
[16] A. Klar and R. Wegener, *Enskog-like kinetic models for multilane vehicular traffic*, J. Statist. Phys., 87 (1997), pp. 91–114.
[17] J. P. Lebacque, *Semimacroscopic Simulation of Urban Traffic*, paper presented at the International 84 Minneapolis summer conference, AMSE, 1984.
[18] J. P. Lebacque, *The Godunov scheme and what it means for first order traffic flow models*, Rapport CERMICS 95-48, ENPC, 1995.
[19] J. P. Lebacque, *The Godunov scheme and what it means for first order traffic flow models*, in Transportation and Traffic Theory, J.-B. Lesort, ed., Pergamon, Oxford, 1996, pp. 647–677.
[20] H. Y. Lee, H. W. Lee, and D. Kim, *Dynamic states of a continuum traffic equation with on-ramp*, Phys. Rev. E, 59 (1999), pp. 5101–5111.
[21] R. J. LeVeque, *Numerical Methods for Conservation Laws*, 2nd ed., Lectures Mathematics, ETH Zürich, Birkhäuser Verlag, Basel, 1992.
[22] M. J. Lighthill and G. B. Whitham, *On kinematic waves* II*—A theory of traffic flow on long crowded roads*, Proc. Roy. Soc. London Ser. A, 229 (1955), pp. 317–345.
[23] P. G. Michalopoulos, D. E. Beskos, and J. K. Lin, *Analysis of interrupted traffic flow by finite difference methods*, Transportation Res., 18B 4/5 (1984), pp. 409–421.
[24] P. Nelson, *A kinetic model of vehicular traffic and its associated bimodal equilibrium solutions*, Transport Theory Statist. Phys., 24 (1995), pp. 383–409.
[25] P. Nelson and A. Sopasakis, *The Prigogine–Herman kinetic model predicts widely scattered traffic flow data at high concentrations*, Transportation Res. Part B, 32 (1998), pp. 589–604.
[26] P. Nelson and A. Sopasakis, *The Chapman–Enskog expansion: A novel approach to hierarchical extension of Lighthill–Whitham models*, in Transportation and Traffic Theory, A. Ceder, ed., Pergamon, Oxford, 1999, pp. 51–79.
[27] P. Nelson, D. D. Bui, and A. Sopasakis, *A novel traffic stream model deriving from a bimodal kinetic equilibrium*, Preprints of the 8th IFAC (meeting of the International Federation of Automatic Control) on Transportation Systems, Chania, Greece, M. Papageorgiou and A. Pouliezos, eds., Pergamon Press, Oxford, UK, 1997, pp. 799–804.
[28] S. L. Paveri-Fontana, *On Boltzmann-like treatments for traffic flow: A critical review of the basic model and an alternative proposal for dilute traffic analysis*, Transportation Res., 9 (1975), pp. 225–235.
[29] I. Prigogine and R. Herman, *Kinetic Theory of Vehicular Traffic*, American Elsevier, New York, 1971.
[30] I. Prigogine, R. Herman, and R. L. Anderson, *On individual and collective flow*, Bull. Acad. R. Belg. Cl. Sci., 48 (1962), pp. 792–804.
[31] P. I. Richards, *Shockwaves on the highway*, Oper. Res., 4 (1956), pp. 42–51.
[32] A. Sopasakis, *Formal Asymptotic Models of Vehicular Traffic. Numerical Investigations*, manuscript.
[33] A. Sopasakis, *Unstable flow theory and modeling*, Math. Comput. Modelling, 35 (2002), pp. 623–641.
[34] R. Wegener and A. Klar, *A kinetic model for vehicular traffic derived from a stochastic microscopic model*, Transport Theory Statist. Phys., 25 (1996), pp. 785–798.

SIAM J. APPL. MATH.
Vol. 63, No. 5, pp. 1585–1603

A PERIODICALLY FORCED WILSON–COWAN SYSTEM*

V. W. NOONBURG†, D. BENARDETE†, AND B. POLLINA†

Abstract. A Wilson–Cowan system, which models the interaction between subpopulations of excitatory and inhibitory neurons, is studied for the case in which the inhibitory neurons are receiving external periodic input. If the feedback within the excitatory population is large enough, the response of the system to large amplitude, low frequency input is determined by the relative values of the excitatory threshold θ_x and the inhibitory-to-excitatory feedback parameter b. Feedback to the inhibitory cells is assumed to be relatively small. In the parameter range considered, the system has two periodic attractors: a high activity state and a low activity state. It is shown that, depending on the parameter values, periodic input can produce two completely different effects; it can either initiate the high activity state or switch it off. If it is assumed that the threshold θ_x increases with increased excitatory activity, there exists a range of b for which periodic input can cause bursting activity in the system.

Key words. Wilson–Cowan system, periodic forcing, bursting, structural stability

AMS subject classifications. 92B20, 34C25, 37G15

DOI. 10.1137/S003613990240814X

1. Introduction.

A Wilson–Cowan system is a system of two first-order differential equations that model the activity in a localized population of neurons. The population is assumed to consist of two homogeneous subpopulations, one containing only excitatory cells and the other only inhibitory cells. The original derivation of the Wilson–Cowan equations appeared in a 1972 paper [14]. As pointed out recently in a book by Whittle [13, p. 122], the approximations used in the derivation are "Draconian in the extreme"; however, as shown by Hoppensteadt and Izhikevich [5, pp. 45–48], the behavior of an autonomous Wilson–Cowan system can be complex and interesting. A recent paper by Eggert and van Hemmen [1] shows that it is possible to design a macroscopic model of neural activity, such as the Wilson–Cowan model, so it will reproduce in a quantitatively exact manner the joint activity of large groups of "integrate-and-fire" neurons. This suggests that new results for the Wilson–Cowan system may apply directly to current models for large populations of single neurons.

In a 1997 paper [12], Tsodyks et al. use a Wilson–Cowan model with biologically derived coefficients to explain paradoxical behavior in the rat hippocampus, where the inhibitory cells (interneurons) are assumed to be receiving external periodic input. The authors of that paper make the assumption that the time-dependent input changes slowly enough so that the system returns to equilibrium at each moment of time. In this paper, we will analytically examine the time-dependent behavior of a Wilson–Cowan system with periodic forcing to the inhibitory cells. Our aim is to determine conditions on the parameters under which periodic forcing of low frequency can significantly alter the behavior of the system. Such results occur when the autonomous system is bistable. In [12] it is assumed that the autonomous system has a single attractor in the interior of the unit square; therefore, most of our results

*Received by the editors May 16, 2002; accepted for publication (in revised form) November 12, 2002; published electronically June 12, 2003.
http://www.siam.org/journals/siap/63-5/40814.html
†Department of Mathematics, University of Hartford, West Hartford, CT 06117 (noonburg@hartford.edu, dbenard@cs.hartford.edu, bpollina@cs.hartford.edu). The research of the first author was supported in part by the Institute for Mathematics and Its Applications, with funds provided by the National Science Foundation.

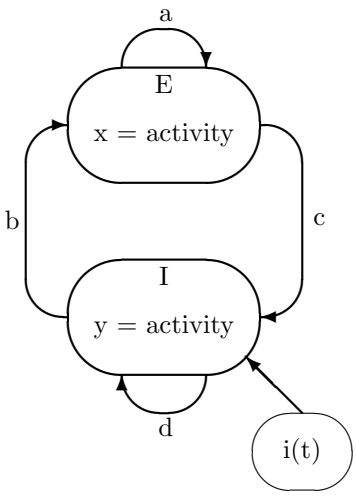


FIG. 1.1. *Feedback in the neural population.*

are complementary to those in [12]. However, in section 3 we point out one region of parameter space where our results do overlap and confirm experimental results described in [12].

The functions $x(t)$ and $y(t)$ in the Wilson–Cowan equations represent, respectively, the fraction of cells in the excitatory subpopulation E and the inhibitory subpopulation I that are active at time t. There are four positive feedback parameters a, b, c, and d which are used to specify the connection strengths both between and within the two subpopulations (see Figure 1.1). For example, the net feedback to E at time t is $ax(t) - by(t)$, and the effect of this on cells in E is determined by a *sigmoidal response function* $\mathbf{S}$.

In our analysis, it is not necessary to know the exact form of the function $\mathbf{S}$ in order to determine the behavior of the system. We require only that $\mathbf{S}$ be a "soft threshold function," as defined in [13, p. 21], that is, have the following properties:

1. $\mathbf{S}(z)$ is at least twice continuously differentiable on $(-\infty, \infty)$.
2. $\mathbf{S}(z)$ increases monotonically from 0 to 1 on $(-\infty, \infty)$.
3. $\mathbf{S}(z)$ has a single point of inflection at $z = 0$.

The third property implies that $\mathbf{S}'(z)$ increases monotonically on $(-\infty, 0)$ and decreases monotonically on $(0, \infty)$.

The Wilson–Cowan system to be studied can then be written in the form

$$\left\{ \begin{array}{rcl} \tau_x x'(t) & = & -x(t) + \mathbf{S}(ax(t) - by(t) - \theta_x), \\ \tau_y y'(t) & = & -y(t) + \mathbf{S}(cx(t) - dy(t) - \theta_y + i(t)), \end{array} \right. \tag{1.1}$$

where $i(t) = \alpha \sin(\omega t)$ is the external periodic input to the inhibitory cells and θ_x and θ_y are the thresholds for the excitatory and inhibitory cells, respectively. Tsodyks et al. [12] define the constants τ_x and τ_y as the time required to bring neurons in the respective populations to firing, as they receive subthreshold excitation. In the original derivation of the equations (see [14]), the time constant τ_x, for example, was introduced by setting $x(t+\tau_x)$ equal to the proportion of cells in E which were sensitive (that is, not in their refractory period) and also receiving threshold excitation at time

t. In our analysis, we assume that the two time constants are equal, and we therefore take $\tau_x = \tau_y = 1$. However, due to the structural stability of the system in the regions of parameter space being considered, it will be made clear, in section 4, that our results can be extended to cases in which these two time constants are not the same.

For the system (1.1) we are able to show, under certain conditions on the parameters, that low frequency periodic input $i(t) = \alpha \sin(\omega t)$, with large enough amplitude α, can be used to *switch* the activity in the network from a high activity state to a low activity state. If the parameters are altered slightly, the input can have the opposite effect; that is, it can force the system into its high activity state. Furthermore, if it is assumed that θ_x increases slightly as the activity level x increases, the input can produce a "bursting" response. The two parameters that are the major determinants of which type of behavior occurs are the excitatory threshold θ_x and the feedback parameter b from $I \Rightarrow E$. In the next two sections the reason for requiring *low frequency* input will be made clear.

2. Reduction to a first-order equation. In this section it will be shown that by placing a single restriction on the four feedback parameters a, b, c, and d, the study of the behavior of the forced Wilson–Cowan system can be reduced to the study of a single periodic differential equation of first order. Some known properties of periodic first-order differential equations are stated in the appendix; and these are referred to, when needed, in this section and in section 3.

We define a new parameter $\lambda \equiv c/a$; that is, λ is the ratio of the $E \Rightarrow I$ feedback parameter c to the recurrent excitatory $E \Rightarrow E$ feedback parameter a. We then make the assumption that the inhibitory feedback parameters are similarly related; that is, we assume that $d/b = \lambda$. This choice of parameter values was motivated by the experimental data used in the examples in [12]. By proving the structural stability of the system, the results obtained under this restriction will subsequently be shown to hold for all d/b sufficiently close to λ. In the numerical simulations in section 5 the meaning of "sufficiently close" will be quantified for a set of representative systems.

Assume $\tau_x = \tau_y = 1$ and $d = b\lambda$. Then the system (1.1) becomes

$$\begin{aligned} x' &= -x + \mathbf{S}(ax - by - \theta_x), \\ y' &= -y + \mathbf{S}(\lambda ax - \lambda by - \theta_y + i(t)). \end{aligned} \tag{2.1}$$

We now make a linear change of variables, where (t, x, y) maps to (t, u, x), with $u = ax - by$. This linear map induces a topological conjugacy between system (2.1) and the system

$$\begin{aligned} u' &= ax' - by' \\ &= f(t, u) = -u + a\mathbf{S}(u - \theta_x) - b\mathbf{S}(\lambda u - \theta_y + i(t)), \\ x' &= k(t, u, x) = -x + \mathbf{S}(u - \theta_x). \end{aligned} \tag{2.2}$$

Let $P = 2\pi/\omega$ be the period of $i(t)$. Since the linear conjugacy of (2.1) with (2.2) respects the identification of (t, x, y) with $(t + P, x, y)$ and (t, u, x) with $(t + P, u, x)$, it follows that it passes to a linear conjugacy of the induced flows on the quotient spaces $S^1 \times \mathbf{R}^2$. Therefore there is a one-to-one correspondence between periodic orbits of (2.1) and periodic orbits of (2.2). Moreover, this correspondence respects the type (attractor, repeller, saddle) of the periodic orbit.

Consider now system (2.2) on $S^1 \times \mathbf{R}^2$, and the single differential equation

$$u' = f(t, u) \tag{2.3}$$

on $S^1 \times \mathbf{R}$. There is a one-to-one correspondence of periodic solutions of (2.2) and (2.3) which takes attractors of (2.2) to attractors of (2.3) and saddles of (2.2) to repellers of (2.3). Periodic orbits of (2.2) induce periodic orbits of (2.3) simply by projecting on the u variable. Going the other way, assume that $\hat{u}(t)$ is a periodic solution of (2.3). Since the response function $\mathbf{S}$ is bounded, it follows that there exist constants $r < 0 < s$ such that the slope function k is positive on the line $x = r$ and negative on the line $x = s$. Furthermore, the derivative k_x is -1. The region $\{(t, x) | r \leq x \leq s\}$ is a forward invariant contracting region for the equation $x' = k(t, \hat{u}, x)$ and therefore contains a unique periodic attractor for the equation (see Lemma A.4 in the appendix). It is clear that $x' = k(t, \hat{u}, x)$ has no periodic solution outside this invariant region. So we have shown how a periodic solution of (2.3) corresponds to a periodic solution of system (2.2).

To simplify the analysis of (2.3) we define a new time variable $\tau = \omega t/(2\pi)$. In terms of τ, (2.3) becomes

$$\begin{aligned} \frac{du}{d\tau} &= \frac{2\pi}{\omega} f(\tau, u) \\ &= \frac{2\pi}{\omega} [-u + a\mathbf{S}(u - \theta_x) - b\mathbf{S}(\lambda u - \theta_y + \alpha \sin(2\pi\tau))]. \end{aligned} \tag{2.4}$$

Equation (2.4) is periodic of period 1 in τ, and the period in t, $P = 2\pi/\omega$, acts as a multiplier of the slopes. This is the reason why low frequency input has a large effect on the system behavior, whereas high frequency input does not. Using the fact that $\mathbf{S}$ takes values only between 0 and 1, it can be seen that $f(\tau, u) < 0$ whenever $u \geq a$, and $f(\tau, u) > 0$ whenever $u \leq -b$. This implies that, as $\tau \to \infty$, all solutions of (2.4) ultimately end up inside the region $-b \leq u \leq a$. In the next section we will examine the slope field for (2.4), restricted to the invariant cylinder $\mathcal{H} = \{0 \leq \tau < 1 \pmod 1, -b \leq u \leq a\} \subset S^1 \times \mathbf{R}$, and determine the behavior of solutions under various conditions on the parameters.

3. Periodic solutions of the equation in u. In this section it will be shown that there are three distinct regions of parameter space in which periodic forcing to the inhibitory cells can significantly alter the behavior of solutions of (2.4), and consequently the behavior of solutions of the Wilson–Cowan system (2.1).

In order to determine the number and position of periodic solutions of (2.4), we need to examine the nullclines $f(\tau, u) = 0$, and also the curves $f_u(\tau, u) = 0$. The former divide the region $\mathcal{H}$ into subregions of positive or negative slopes, and the latter determine where trajectories are getting closer together or further apart.

The function $f_u(\tau, u) = -1 + a\mathbf{S}'(u - \theta_x) - b\lambda\mathbf{S}'(\lambda u - \theta_y + \alpha \sin(2\pi\tau))$. Since $\mathbf{S}'(z)$ is everywhere positive and assumes its maximum value at $z = 0$, if $a < 1/\mathbf{S}'(0)$, $f_u(\tau, u)$ is negative for all τ and u, and the cylinder $\mathcal{H}$ is an invariant contracting region. In this case, (2.4) has a single periodic attractor (see Lemma A.4 in the appendix). The more interesting behavior occurs when $a > 1/\mathbf{S}'(0)$, and it will be assumed that this is the case in what follows.

It should be noted that all of the graphs in this section were produced by MAPLE. The $\mathbf{S}$-function used in drawing these graphs is $\mathbf{S}(z) = 1/(1 + e^{-z})$; however, the conclusions reached require only that $\mathbf{S}$ satisfy the three properties for a generic $\mathbf{S}$-function, which were given in section 1.

Consider the function f as the difference between two functions $g(u) = -u + a\mathbf{S}(u - \theta_x)$ and the sigmoidally shaped time-dependent function $h(\tau, u) = b\mathbf{S}(\lambda u -$

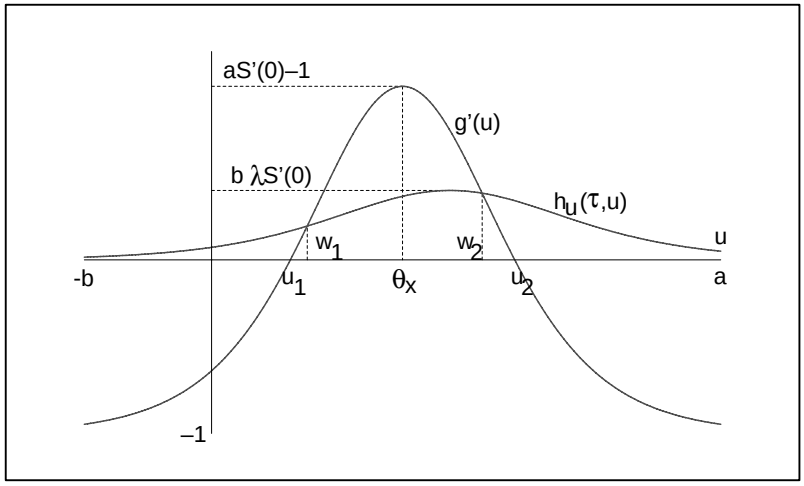


FIG. 3.1. *Zeros of* $f_u(\tau,u) = g'(u) - h_u(\tau,u)$.

$\theta_y + \alpha\sin(2\pi\tau)$). The function $f(\tau,u)$ is zero wherever $g(u)$ and $h(\tau,u)$ are equal; similarly, the function $f_u(\tau,u)$ is zero where $g'(u)$ and $h_u(\tau,u)$ are equal.

We look first for the zeros of $f_u(\tau,u)$. Again, using the fact that $\mathbf{S}'(z)$ has its maximum at $z=0$ and is strictly monotonic on $(-\infty,0)$ and $(0,\infty)$, it can be seen in Figure 3.1 that the function $g'(u) = -1 + a\mathbf{S}'(u-\theta_x)$ always has exactly two zeros when $a > 1/\mathbf{S}'(0)$. We define u_1 and u_2 to be the values of u at which $g'(u) = 0$; that is,

$$u_1 < u_2 \text{ are the two solutions of } \mathbf{S}'(u-\theta_x) = \frac{1}{a}. \tag{3.1}$$

As τ varies between 0 and 1, the graph of g' remains fixed, while the graph of $h_u(\tau,u) = b\lambda\mathbf{S}'(\lambda u - \theta_y + \alpha\sin(2\pi\tau))$ moves left and right (see Figure 3.1); and if $b\lambda\mathbf{S}'(0) < a\mathbf{S}'(0) - 1$, the graph of $h_u(\tau,u)$ must intersect the graph of $g'(u)$ at least twice, but always between u_1 and u_2.

Assume that $\lambda < \frac{a\mathbf{S}'(0)-1}{b\mathbf{S}'(0)} \equiv \lambda_b$. Notice that for large a the condition $\lambda < \lambda_b$ is essentially the condition $c/a < a/b$, which requires the product of the two external feedback parameters to be small relative to the recurrent excitation. With this restriction on λ, there will be *exactly* two intersections if we can show that the slope of $g'(u)$ is always steeper than the slope of $h_u(\tau,u)$ at a point of intersection, that is, if $|g''(u)| > |h_{uu}|$. Consider the closed u-intervals $I_1 = \{u | 0 \le g'(u) \le b\lambda\mathbf{S}'(0), u < \theta_x\}$ and $I_2 = \{u | 0 \le g'(u) \le b\lambda\mathbf{S}'(0), u > \theta_x\}$. The continuous function $\mathbf{S}''(z)$ is zero only at $z=0$, and therefore $|\mathbf{S}''(u-\theta_x)|$ has strictly positive minimum values m_1 on I_1, and m_2 on I_2. Thus $g'(u)$ has slope $a\mathbf{S}''(u-\theta_x) \ge am_1$ in I_1 and $\le -am_2$ in I_2. For any fixed value of τ, the slope of $h_u(\tau,u)$ is $h_{uu}(\tau,u) = b\lambda^2\mathbf{S}''(\lambda u - \theta_y + \alpha\sin(2\pi\tau))$; therefore, for any u in the closed interval $[-b,a]$, $|h_{uu}(\tau,u)|$ is less than or equal to $b\lambda^2 M$, where M is the maximum of $|\mathbf{S}''(z)|$ on the interval $-b\lambda - \theta_y - \alpha \le z \le a\lambda - \theta_y + \alpha$. Therefore, if $b\lambda^2 M < am$, where $m = \min(m_1, m_2)$, at any point of intersection of the curves g' and h_u we have $|g''(u)| > |h_{uu}(u)|$.

Now consider the region bounded by the curve $g'(u)$ and the interval $u_1 \le u \le u_2$ on the u-axis. This region is divided by the vertical line through θ_x into a left subregion L and a right subregion R. Orienting h_u in the direction of increasing u, the result

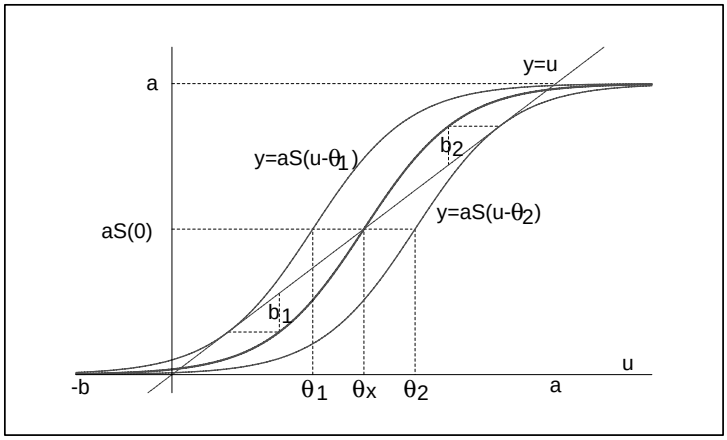


FIG. 3.2. *Values of θ_x for which $a\mathbf{S}(u-\theta_x)$ is tangent to $y=u$.*

$|g''(u)| > |h_{uu}(u)|$ implies that any intersection of h_u with the left boundary of L is an entrance to L. There is a unique exit from L at the point where h_u crosses the vertical line through θ_x. Since the number of entrances to L must equal the number of exits from L, it follows that there is only one intersection of h_u with the left boundary of L. Similarly there is only one intersection of h_u with the right boundary of R. These two intersections are the unique intersections of g' with h_u.

We have therefore shown that if $\lambda < \min(\lambda_b, \sqrt{\frac{am}{bM}}) \equiv \Lambda$, the function $f_u(\tau, u)$ will have exactly two zeros $w_1(\tau)$ and $w_2(\tau)$, lying between u_1 and u_2 for each value of τ. Note that this also implies that the function $f(\tau, u)$ has at most three zeros. Furthermore, since w_1 and w_2 are solutions of $f_u(\tau, u) = 0$, by implicit differentiation they are seen to be differentiable functions of τ, with $dw_i/d\tau = -f_{u\tau}/f_{uu}$. The second partial derivative $f_{uu}(\tau, u)$ is never zero at the points $w_1(\tau)$ or $w_2(\tau)$, since it is equal to the *difference* of the slopes of $g'(u)$ and $h_u(\tau, u)$ at those points.

The condition $\lambda < \sqrt{\frac{am}{bM}}$ appears to be more restrictive than the condition $\lambda < \lambda_b$. The ratio m/M is less than 1 and approaches 1 as the derivative $\mathbf{S}'$ approaches a piecewise linear tent function with its peak at 0. However, our proof of the structural stability of the Wilson–Cowan system in section 4 shows that "sufficiently small" perturbations of the **S**-function can be made without altering the general behavior of the system. Therefore, in biological terms, the condition $\lambda < \Lambda$ simply requires that the external feedback between the two populations be small enough relative to the recurrent excitation in E.

Next consider the curves where $f(\tau, u) = 0$. With $a > 1/\mathbf{S}'(0)$, the graph of $y = a\mathbf{S}(u-\theta_x)$ has its maximum slope, greater than 1, at $u = \theta_x$ and intersects the line $y = u$ in $1, 2$, or 3 points depending on the value of θ_x. Define θ_1 and θ_2 to be the lower and upper values of θ_x for which the curve $y = a\mathbf{S}(u-\theta_x)$ is tangent to the line $y = u$ (see Figure 3.2); that is, $\theta_1 < \theta_2$ are the solutions θ of the simultaneous equations

$$\begin{cases} a\mathbf{S}(u-\theta) = u, \\ a\mathbf{S}'(u-\theta) = 1. \end{cases} \tag{3.2}$$

Then if $\theta_1 < \theta_x < \theta_2$, $g(u) = -u + a\mathbf{S}(u-\theta_x)$ has exactly three zeros, and if $\theta_x < \theta_1$ or $\theta_x > \theta_2$, $g(u)$ has only one zero.

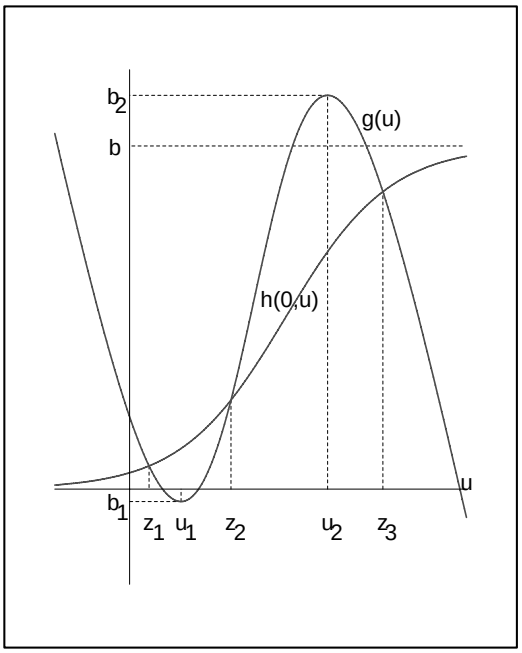


FIG. 3.3. *Zeros of the function* $f(0,u) = g(u) - h(0,u)$, $\theta_1 < \theta_x < \theta_2$, $b < b_2$.

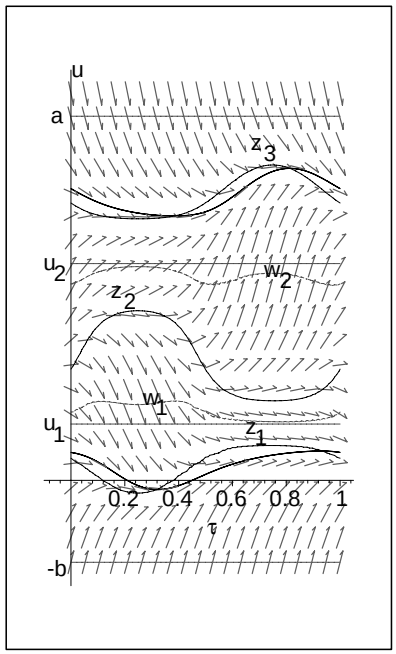


FIG. 3.4. *Slope field in* $\mathcal{H}$, *with* $\theta_1 < \theta_x < \theta_2$, $b < b_2$.

We consider first the case of $\theta_1 < \theta_x < \theta_2$. In Figure 3.3, graphs of $g(u)$ and $h(0,u)$ are drawn. From our discussion of the function $g'(u)$, we know that a local minimum and a local maximum of g occur at u_1 and u_2, respectively. Define $b_2 = g(u_2)$ to be the local maximum of g, and $b_1 = g(u_1)$ to be the local minimum. Notice that $b_2 = 0$ when $\theta_x = \theta_2$. As θ_x moves to the left or right of θ_2 on the u-axis, the position of the maximum of $g(u)$, and its magnitude, both change by the same amount $\theta_2 - \theta_x$ (see Figure 3.2); therefore, $b_2 = \theta_2 - \theta_x$. Similarly, it can be seen that the minimum $b_1 = g(u_1) = \theta_1 - \theta_x$.

If $b < b_2$, as the curve $h(\tau,u)$ moves left and right with τ, for any τ between 0 and 1, it always intersects the curve $g(u)$ in exactly three points $z_i(\tau)$. It can be seen from Figure 3.3 that for all τ these intersection points satisfy $z_1(\tau) < u_1 < z_2(\tau) < u_2 < z_3(\tau)$. From our previous analysis of the function f_u, we know that the curve $z_2(\tau)$ lies in the region between the curves $w_1(\tau)$ and $w_2(\tau)$.

For a representative set of parameter values $a > 1/\mathbf{S}'(0), \theta_1 < \theta_x < \theta_2$, $b < b_2$, and $\lambda < \Lambda$, the position of these curves in $\mathcal{H}$ is shown in Figure 3.4. Since f_u is negative for $u < u_1$ and $u > u_2$, and $f(\tau,u_1) < 0, f(\tau,u_2) > 0$ for any τ, the subregions $\mathcal{H}_1 = \{0 \le \tau < 1 \pmod 1, -b \le u \le u_1\}$ and $\mathcal{H}_2 = \{0 \le \tau < 1 \pmod 1, u_2 \le u \le a\}$ are invariant *contracting* regions for (2.4), so that Lemma A.4 (appendix) shows that there always exists a unique attracting period-1 solution $u^-(\tau)$ in $\mathcal{H}_1$ and $u^+(\tau)$ in $\mathcal{H}_2$. These are the two darker curves seen in Figure 3.4.

By making the input frequency ω small enough, the absolute value of the slope function $(2\pi/\omega)f(\tau,u)$ can be made larger than the absolute value of the derivative of both w_1 and w_2 at every value of τ. Under these conditions, the region $w_1(\tau) < u < w_2(\tau)$ is a repelling region for $u(\tau)$; that is, it is a contracting forward invariant

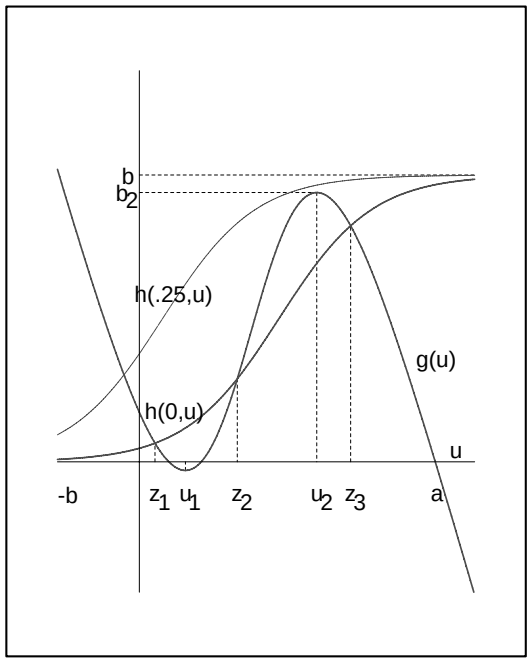


FIG. 3.5. *Zeros of the function* $f(\tau,u) = g(u) - h(\tau,u)$, $\theta_1 < \theta_x < \theta_2$, $b > b_2$.

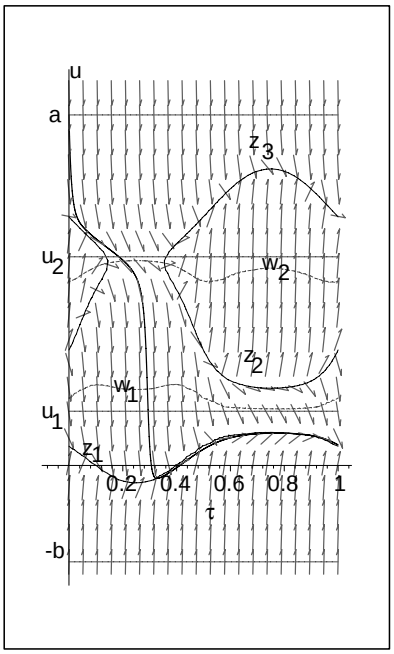


FIG. 3.6. *Slope field in* $\mathcal{H}$, *with* $\theta_1 < \theta_x < \theta_2$, $b > b_2$.

region when τ is replaced by $-\tau$. In this case there is a unique period-1 repeller in the interior of this subregion of $\mathcal{H}$ (see the appendix). All of the other solutions are attracted to either u^- or u^+. The above analysis allows us to state the following theorem.

THEOREM 3.1. *If* $a > 1/\mathbf{S}'(0)$, $\theta_1 < \theta_x < \theta_2$, $b < b_2 = \theta_2 - \theta_x$, *and* $\lambda < \Lambda$, *then, for all* ω *small enough, for any value of the amplitude* α *the equation* $u' = (2\pi/\omega)f(\tau,u)$ *has exactly three period-*1 *solutions; attractors* $u^-(\tau)$ *and* $u^+(\tau)$ *and repeller* $u^0(\tau)$ *satisfying* $u^-(\tau) < u_1 < u^0(\tau) < u_2 < u^+(\tau)$ *for all* τ. *The corresponding Wilson–Cowan system* (1.1), *with* $d = \lambda b$, $\tau_x = \tau_y = 1$, *has a low activity attracting state* $(x^-(t), y^-(t))$ *corresponding to* $u^-(\tau)$ *and a high activity attracting state* $(x^+(t), y^+(t))$ *corresponding to* $u^+(\tau)$. *Every solution with initial values* $(x(0), y(0))$ *satisfying* $ax(0) - by(0) > u^0(0)$ *is attracted to* (x^+, y^+), *and solutions with* $(x(0), y(0))$ *satisfying* $ax(0) - by(0) < u^0(0)$ *are attracted to* (x^-, y^-). *The period of oscillation, in* t, *of each of these attracting solutions is* $2\pi/\omega$.

With the parameter values a, b, θ_x, and λ as hypothesized in Theorem 3.1, low frequency external input $\alpha \sin(\omega t)$ to the inhibitory cells cannot significantly change the behavior of the system. In order for the input to cause one of the two attractors to lose stability, and hence *switch* the system from one attracting state to the other, it is necessary that either (1) the feedback parameter b from $I \Rightarrow E$ be increased to a value greater than the bifurcation value b_2, or (2) the excitatory threshold θ_x be decreased to a value less than θ_1. If $\theta_x > \theta_2$, the local maximum of the function $f(\tau,u)$ is negative for all values of τ, so that only the lower attractor $u^-(\tau)$ exists for any input.

Consider now the case in which $\theta_1 < \theta_x < \theta_2$ and $b > b_2$. In this case it can be seen in Figure 3.5 that there exists an α^* such that for all $\alpha > \alpha^*$ the curve $h(\tau,u)$

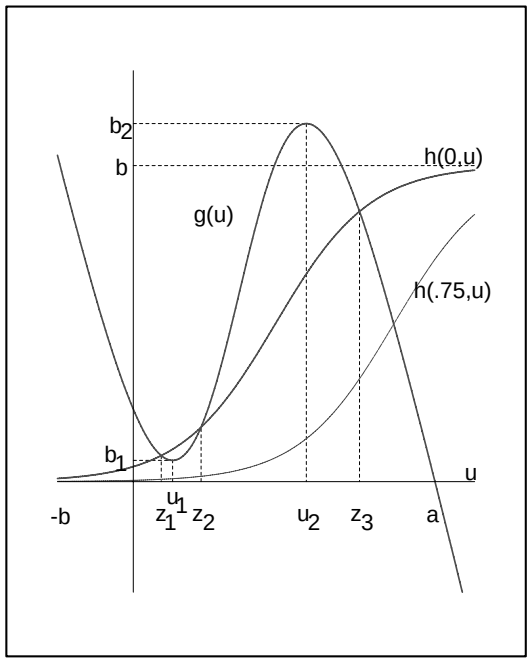


FIG. 3.7. *Zeros of the function* $f(\tau, u) = g(u) - h(\tau, u)$, $\theta_x < \theta_1$, $b_1 < b < b_2$.

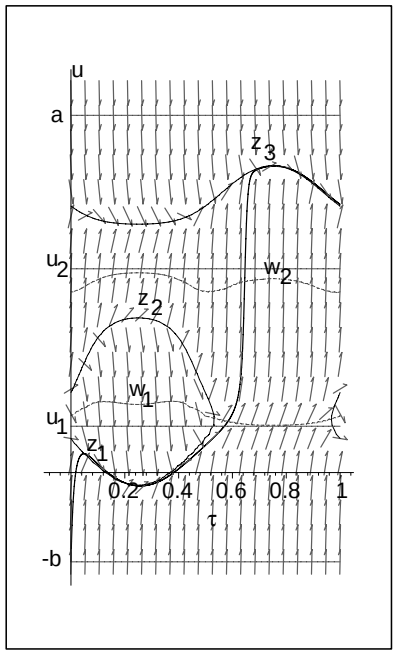


FIG. 3.8. *Slope field in* $\mathcal{H}$, *with* $\theta_x < \theta_1$, $b_1 < b < b_2$.

moves far enough to the left so that there will be an interval of τ (around $\tau = 0.25$) in which the intersections $z_2(\tau)$ and $z_3(\tau)$ no longer exist. Figure 3.6 shows the corresponding arrangement of the curves $z_i(\tau)$ in $\mathcal{H}$. The lower invariant contracting region $\mathcal{H}_1$ still exists, but by making the input frequency ω small enough, the solution $u(\tau)$ of (2.4) with $u(0) = a$ can be made to decrease below the minimum u-value on the curve $z_2(\tau)$, and $u(\tau)$ will therefore be attracted to the lower period-1 solution $u^-(\tau)$. Using the uniqueness of the solutions, it can be seen that all solutions of (2.4) are then attracted to $u^-(\tau)$ as $\tau \to \infty$. The solution starting at $u(0) = a$ is seen as the darker curve in Figure 3.6.

THEOREM 3.2. *If* $a > 1/\mathbf{S}'(0), \theta_1 < \theta_x < \theta_2, b > b_2$, *and* $\lambda < \Lambda$, *there exist positive constants* α^*, *and* ω^* *depending on* α, *such that for all* $\alpha > \alpha^*$ *and* $\omega < \omega^*(\alpha)$ *all solutions of* (2.4) *tend to the period*-1 *solution* $u^-(\tau)$, *with* $-b < u^-(\tau) < u_1$. *For these values of the parameters, external input* $i(t) = \alpha \sin(\omega t)$ *to the inhibitory cells forces all solutions of the Wilson–Cowan system* (2.1) *to the low activity state* $(x^-(t), y^-(t))$, *which is periodic with period* $2\pi/\omega$.

If $\theta_x < \theta_1$, the external input can produce two different types of behavior. If b is less than the value of the local minimum $b_1 = g(u_1) = \theta_1 - \theta_x$, which is now positive, then only the upper period-1 solution $u^+(\tau)$ exists for any input $\alpha \sin(\omega t)$; but if $b_1 < b < b_2$, then for all α large enough there will be an interval of τ around $\tau = 0.75$ for which $z_3(\tau)$ is the only solution of $f(\tau, u) = 0$ (see Figure 3.7). In this case, only the upper invariant contracting region $\mathcal{H}_2$ exists, and the input $\alpha \sin(\omega t)$ can be used to drive all solutions to the upper attractor $u^+(\tau)$ if $\alpha > \alpha^*$ and $\omega < \omega^*(\alpha)$. Such a solution can be seen in the slope field in Figure 3.8, where we have used the initial value $u(0) = -b$.

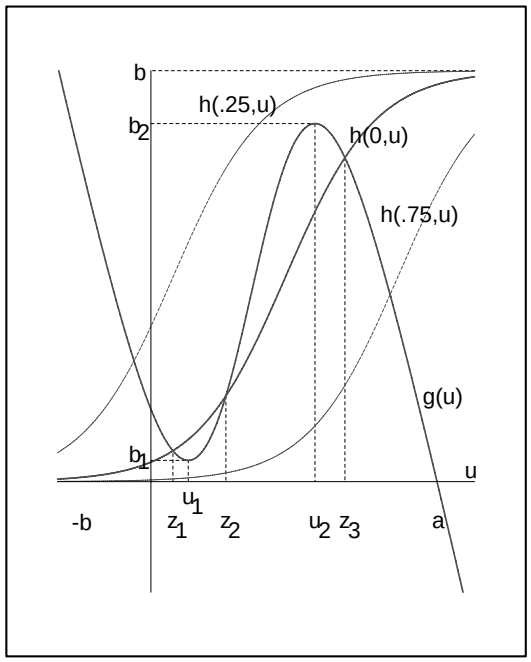


FIG. 3.9. *Zeros of the function* $f(\tau, u) = g(u) - h(\tau, u)$, $\theta_x < \theta_1$, $b > b_2$.

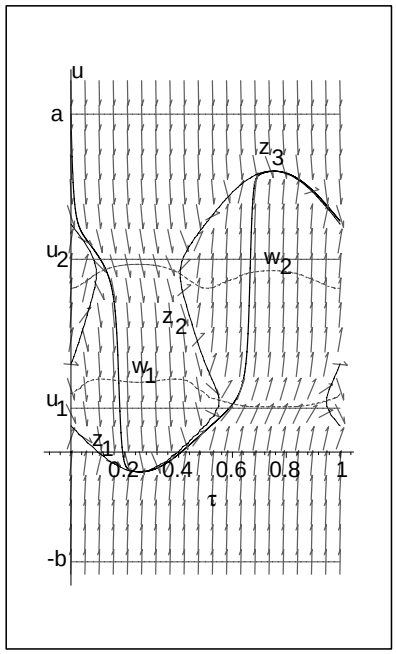


FIG. 3.10. *Slope field in* $\mathcal{H}$, *with* $\theta_x < \theta_1$, $b > b_2$.

THEOREM 3.3. *If* $a > 1/\mathbf{S}'(0), \theta_x < \theta_1, b_1 < b < b_2$, *and* $\lambda < \Lambda$, *there exist positive constants* α^*, *and* ω^* *depending on* α, *such that for all* $\alpha > \alpha^*$ *and* $\omega < \omega^*(\alpha)$ *all solutions of* (2.4) *tend to the period-*1 *solution* $u^+(\tau)$, *with* $u_2 < u^+(\tau) < a$. *For these values of the parameters, external input* $i(t) = \alpha \sin(\omega t)$ *to the inhibitory cells drives all solutions of the Wilson–Cowan system* (2.1) *to the high activity state* $(x^+(t), y^+(t))$, *which is periodic with period* $2\pi/\omega$ *in* t.

It can be seen in Figure 3.9, however, that if $\theta_x < \theta_1$ and $b > b_2$, then for large enough α there can be intervals of τ where z_3 is the only solution of $f(\tau, u) = 0$, as well as intervals of τ where z_1 is the only solution. In this case, with large amplitude, low frequency input $\alpha \sin(\omega t)$, both the upper and lower attractors can lose stability, and the solutions of (2.4) can be made to converge to an attracting periodic solution which oscillates alternately between the high and low activity states. Such a solution is shown in the slope field in Figure 3.10. With b large enough, our results in this region of parameter space resemble those in [12]. As hypothesized in [12], $x(t)$ and $y(t)$ appear to oscillate *in phase* in this region. This can be seen in the graph of the numerical solution pictured in Figure 5.4 in section 5.

The above theorems describe the behavior of the system in three distinct regions of parameter space. In each of these regions, the system has only hyperbolic periodic solutions, and in section 4 it will be shown to be a structurally stable system. This in turn implies that the behavior of the system in each of these regions remains similar for small perturbations of the parameters d, τ_x, and τ_y and also for small perturbations of the **S**-function and the input $i(t)$. In all of the systems we have simulated numerically, the value of d can be varied by a large percentage before any system bifurcation occurs. Note that the parameter d determines the magnitude of

the feedback among the cells in the inhibitory subpopulation, and the systems seem not to be very sensitive to this particular parameter in the regions of parameter space under consideration.

4. Structural stability of the periodically forced Wilson–Cowan system. In this section we show that in each of the three regions of parameter space described in Theorems 3.1, 3.2, and 3.3 the periodic Wilson–Cowan system (1.1) is structurally stable. This allows us to conclude that the results for the constrained system, with $d/b = c/a$ and $\tau_x = \tau_y$, can be extended to systems where the values of these parameters lie in sufficiently small neighborhoods of the constrained values, but in which the constraints are relaxed. In section 5, numerical simulations will be used to show what this means quantitatively.

We first give the definition of a structurally stable system and state the classical theorem which applies to three-dimensional autonomous systems on compact manifolds [8], [9], [11]. We then prove in Theorem 4.2 that, under certain conditions, the Wilson–Cowan system (1.1) satisfies the Morse–Smale conditions for structural stability. The corollary of Theorem 4.2 below shows that in each of the regions described in Theorems 3.1, 3.2, and 3.3, system (1.1) satisfies all of the hypotheses of Theorem 4.2.

DEFINITION 4.1. *A system of differential equations $x' = F(x)$ on a smooth manifold M (with or without boundary) is C^r* structurally stable *if, for all sufficiently small perturbations of the vector field F and its partial derivatives of order $\leq r$, the unperturbed and perturbed systems are topologically equivalent; that is, there exists a homeomorphism $h : M \to M$ which takes trajectories of the unperturbed system to trajectories of the perturbed system, preserving the time orientation but not necessarily the parameterization.*

Starting with the periodically forced Wilson–Cowan system (1.1) on $\mathbf{R}^2$, we add a new variable s and a new equation to form a related autonomous system on $\mathbf{R} \times \mathbf{R}^2$. We then identify the point (s, x, y) with the point $(s+P, x, y)$ to induce on the manifold $S^1 \times \mathbf{R}^2$ the system

$$\left\{ \begin{array}{rl} s'(t) & = 1, \\ \tau_x x'(t) & = -x(t) + \mathbf{S}(ax(t) - by(t) - \theta_x), \\ \tau_y y'(t) & = -y(t) + \mathbf{S}(cx(t) - dy(t) - \theta_y + i(s)). \end{array} \right. \tag{4.1}$$

THEOREM 4.2. *Suppose that system* (4.1) *has finitely many periodic orbits on the manifold $S^1 \times \mathbf{R}^2$, all of which are hyperbolic, and that every point approaches a periodic orbit in forward time, while in backward time every point approaches a periodic orbit or escapes to infinity, but no nonperiodic point approaches a saddle orbit in forward time and a saddle orbit in backward time. Further, suppose that from the Wilson–Cowan system a sufficiently small perturbation is made of the positive constants $\tau_x, \tau_y, a, b, c, d, \theta_x$, and θ_y; a C^1 sufficiently small perturbation is made of $i(t)$, preserving the period P; and a C^1 sufficiently small perturbation is made of the function $\mathbf{S}$. Then there exists a closed disk $D \subseteq \mathbf{R}^2$ such that any trajectory of the original or perturbed system either lies entirely in the interior of the manifold with boundary $M = S^1 \times D$ or else enters M and never leaves. In particular, all periodic orbits of the perturbed and unperturbed system are contained in the interior of M. The perturbed system is topologically equivalent on M to the unperturbed system, and the topological equivalence $h : M \to M$ approaches the identity as the perturbation and its first-order partial derivatives approach the unperturbed system and its corresponding partial derivatives. In particular, the perturbed system has no equilibria, all the*

periodic orbits of the perturbed system are hyperbolic, the number of periodic orbits in the unperturbed and perturbed systems is the same, and corresponding periodic orbits in the two systems have the same period and type (i.e., attractor, repeller, saddle).

This theorem will be proved by invoking a theorem of Palis, which requires the following definition.

DEFINITION 4.3 (see [10, pp. 319, 322]). *On a compact manifold M (with or without boundary), we call a differential equation* Morse–Smale *if*

(a) *the vector field of the differential equation is nowhere zero on the boundary and nowhere tangent to the boundary;*

(b) *there are finitely many equilibria and periodic orbits;*

(c) *all equilibria and periodic orbits are hyperbolic;*

(d) *all stable and unstable manifolds intersect transversally;*

(e) *the nonwandering set coincides with the set of equilibria and periodic orbits.*

The following theorem is usually stated for a compact manifold without boundary, but the proof carries over for a compact manifold with boundary, provided that one stipulates the tangency condition (a) given in the above definition.

THEOREM 4.4 (see [7]). *On a compact manifold M (with or without boundary), a Morse–Smale differential equation is C^1 structurally stable, and the equivalence $h : M \to M$ approaches the identity as the perturbation and its first partial derivatives approach the unperturbed system and its first partial derivatives.*

Proof of Theorem 4.2. Let us first rewrite system (1.1) in the form

$$x' = F_1(t,x,y) = -mx + m\mathbf{S}(f_1(t,x,y)),$$
$$y' = F_2(t,x,y) = -ny + n\mathbf{S}(f_2(t,x,y)),$$

where $m = 1/\tau_x, n = 1/\tau_y, f_1(t,x,y) = ax - by - \theta_x$, and $f_2(t,x,y) = cx - dy - \theta_y + i(t)$. The same differential equation can also be written in vector form as $(x,y)' = F(t,x,y)$, where the nonautonomous vector field $F(t,x,y) = (F_1(t,x,y), F_2(t,x,y))$.

As was shown previously, the nonautonomous system (1.1) on $\mathbf{R} \times \mathbf{R}^2$ can be rewritten as an autonomous system (4.1) on $S^1 \times \mathbf{R}^2$, which in vector form can be denoted by $(s,x,y)' = F^{\#}(s,x,y)$.

Choose $r > (m+n)/\min(m,n)$, and let D be a disk in the (x,y) plane of radius r, centered at the origin. The vector $\nu = (x,y)$ is an outward-pointing normal vector to the boundary circle of D. For any fixed t, the dot product $F \bullet \nu = -mx^2 - ny^2 + mx\mathbf{S}(f_1(t,x,y)) + ny\mathbf{S}(f_2(t,x,y)) \le -\min(m,n)r^2 + (m+n)r$. Our choice of r implies that $F \bullet \nu < 0$, which shows that for any fixed t, the vector field F points into the disk D. As a consequence, the vector field $F^{\#}$ for the system (4.1) on the manifold $S^1 \times \mathbf{R}^2$ points into the manifold with boundary $M = S^1 \times D$. The same is true for any vector field $G^{\#}$ on M that is a sufficiently small C^0 perturbation of $F^{\#}$.

Since the first equation in system (4.1) is $s'(t) = 1$, it follows that the system has no equilibria. By hypothesis, it has only finitely many periodic orbits, all of which are hyperbolic.

The condition of Theorem 4.2 that no nonperiodic point is both backward and forward asymptotic to a saddle implies that the only possible intersection of stable and unstable manifolds occurs on a periodic orbit. But by the hyperbolicity of periodic orbits, an intersection of stable and unstable manifolds along a periodic orbit is transverse. Therefore all intersections of stable and unstable manifolds are transverse.

If a point does not escape to infinity in backward time, it approaches a repeller in backward time or an attractor in forward time. Therefore no nonperiodic point

can be nonwandering. It follows that the nonwandering set coincides with the set of periodic orbits.

We have shown above that the conditions of Theorem 4.4 are satisfied. It follows that system (4.1) on M is C^1 structurally stable. Let $G^{\#}$ be a sufficiently small C^1 perturbation of the unperturbed vector field $F^{\#}$. Then the systems determined by $F^{\#}$ and $G^{\#}$ are topologically equivalent on M; that is, both systems have the same number of periodic orbits, and corresponding orbits have the same type. □

COROLLARY OF THEOREM 4.2. Consider any set of parameters $\tau_x = \tau_y = 1$, $a, b, c, \lambda, d = \lambda b, \theta_x, \theta_y, \alpha$, and ω, with $i(t) = \alpha \sin(\omega t)$ such that the resulting Wilson–Cowan system satisfies the conditions for Theorem 3.1, Theorem 3.2, or Theorem 3.3. Suppose a sufficiently small perturbation is made of the parameters $\tau_x, \tau_y, a, b, c, d, \theta_x$, and θ_y, together with a C^1 sufficiently small perturbation of $i(t)$ preserving the period $P = 2\pi/\omega$, and a C^1 sufficiently small perturbation of the function $\mathbf{S}$. It follows that the perturbed system has the same properties as the properties demonstrated for the unperturbed system in Theorems 3.1, 3.2, and 3.3.

Proof of the corollary. Given a case of the Wilson–Cowan system whose properties are demonstrated in Theorem 3.1, 3.2, or 3.3, we need to show that this unperturbed system satisfies the conditions needed to apply Theorem 4.2.

In the ensuing argument we will need to refer to system (4.1), to the linearly equivalent system

$$s' = 1, \quad u' = f(s,u), \quad x' = k(s,u,x), \tag{4.2}$$

and to system

$$s' = 1, \quad u' = f(s,u). \tag{4.3}$$

The latter are autonomous systems corresponding to systems (2.2) and (2.3) in section 2, with $u = ax - by$.

The system of Theorem 3.1 has three periodic orbits, and the systems of Theorem 3.2 and Theorem 3.3 have one periodic orbit. Therefore each one has at most finitely many periodic orbits. We first show that each of these periodic orbits is hyperbolic. In system (4.2), because u' depends only on s and u, it follows that a Poincaré map $H : \mathbf{R}^2 \to \mathbf{R}^2$ has the form $H(u_0, x_0) = (H_1(u_0), H_2(u_0, x_0))$. The Jacobian matrix of this map has the form

$$\begin{pmatrix} dH_1/du_0 & 0 \\ \partial H_2/\partial u_0 & \partial H_2/\partial x_0 \end{pmatrix}.$$

Therefore the eigenvalues of the Jacobian map are $\lambda_1 = dH_1/du_0$ and $\lambda_2 = \partial H_2/\partial x_0$. We now compute those eigenvalues.

A Poincaré map H of a periodic differential equation maps the initial value of each trajectory into its value one period later. We will use the Poincaré map with initial value at $t = 0$. It is known [3, pp. 129–130] that the derivative $H_1'(u_0) = \exp[\int_0^1 f_u(t, u(t))dt]$, where $u(t)$ is the solution of the differential equation $u' = f(t,u)$, with $u(0) = u_0$. Since the periodic orbits determined by Theorems 3.1, 3.2, and 3.3 lie entirely in invariant regions where f_u has constant sign, it follows that $0 < \lambda_1 = H_1'(u_0) \neq 1$ for any u_0 which is the initial point of a periodic orbit. In particular, for an attracting periodic orbit of the equation $u' = f(t,u)$ we have $0 < \lambda_1 < 1$, and for a repelling periodic orbit we have $\lambda_1 > 1$. For a fixed u_0, the equation $x' = -x + S(u -$

θ_x) is a first-order linear equation with solution $x(t) = e^{-t}\int_0^t e^s S(u(s)-\theta_x)ds + e^{-t}x_0$, and therefore, $\lambda_2 = \partial H_2/\partial x_0 = 1/e$.

It follows that corresponding to an attracting periodic orbit for system (4.3) we have lying above it an attractor in system (4.2), and this attractor has eigenvalues $0 < \lambda_1 < 1$ and $\lambda_2 = 1/e$. Corresponding to a repelling periodic orbit for system (4.3), we have lying above it a saddle in system (4.2), and this saddle has eigenvalues $\lambda_1 > 1$ and $\lambda_2 = 1/e$. Since the eigenvalues are preserved under the linear change of variables going from system (4.1) to system (4.2), it follows that the periodic orbits shown to exist in Theorems 3.1, 3.2, and 3.3 are in fact hyperbolic.

Every initial point (s_0, u_0, x_0) in system (4.2) approaches a periodic orbit in forward time. In backward time it either approaches a periodic orbit or escapes to infinity.

In the cases of Theorems 3.2 and 3.3, there is no saddle orbit. In the case of Theorem 3.1, there is a unique saddle orbit α for (4.2) which projects to the repeller α^* of (4.3). The points attracted to α in forward time are precisely the points which project to α^*. However, for such a point which is not on α, in backwards time, the x value of the trajectory goes to $\pm\infty$. Therefore, there is no nonperiodic point of (4.2) which in both backwards and forwards time is attracted to a saddle orbit.

We have thus shown that any system satisfying the conditions for Theorem 3.1, 3.2, or 3.3 also satisfies the conditions for Theorem 4.2 on structural stability. Therefore these systems are structurally stable in the sense of Theorem 4.2. In particular, all the properties demonstrated for the systems in Theorems 3.1, 3.2, or 3.3 also hold under a sufficiently small C^1 perturbation of the type referred to in the corollary. □

5. Examples. This section contains numerical solutions of the Wilson–Cowan system (2.1) for some suitably chosen parameter values, to demonstrate the types of behavior described in section 3. The examples are specifically chosen to illustrate the behavior in four different regions of the bifurcation diagram pictured in section 6, and are labelled according to the region in which they lie. In each of these examples we are assuming that $d/b = c/a = \lambda$. For simplicity we use the sigmoidal response function $\mathbf{S}(z) = \frac{1}{(1+e^{-z})}$. For this particular $\mathbf{S}$-function, the derivative $\mathbf{S}'(z) = \mathbf{S}(z)(1-\mathbf{S}(z))$, and $\mathbf{S}'(0) = 1/4$. The excitatory feedback parameter a needs only to satisfy $a > 1/\mathbf{S}'(0) = 4$, and we will arbitrarily choose $a = 8$. For the chosen $\mathbf{S}$-function the values of θ_1 and θ_2, which depend only on $\mathbf{S}$ and a, can be found from (3.2) by using the inverse function $\mathbf{S}^{-1}(z) = \ln(z/(1-z))$:

$$\theta_1 = \frac{a}{2} - \left[\left(\frac{a}{2}\right)\sqrt{1-\frac{4}{a}} + \ln\left(\frac{1-\sqrt{1-4/a}}{1+\sqrt{1-4/a}}\right)\right] \approx 2.9343, \tag{5.1}$$

$$\theta_2 = \frac{a}{2} + \left[\left(\frac{a}{2}\right)\sqrt{1-\frac{4}{a}} + \ln\left(\frac{1-\sqrt{1-4/a}}{1+\sqrt{1-4/a}}\right)\right] \approx 5.0657.$$

We first consider a system with $\theta_1 < \theta_x < \theta_2$, by letting $\theta_x = 3$. The bifurcation value $b_2 = \theta_2 - \theta_x$ is approximately equal to 2.0657. The parameter λ is required to be less than $(a\mathbf{S}'(0)-1)/(b\mathbf{S}'(0)) = 4/b$, and we will arbitrarily choose $\lambda = 0.8$, which is well within the required limit when b is close to b_2. This value of λ can also be shown to satisfy the condition $\lambda < \sqrt{\frac{am}{bM}}$ for this particular $\mathbf{S}$-function and the values of b being considered. The inhibitory threshold θ_y is arbitrary, but both θ_y and λ will affect the bifurcation value α^*, since they each affect where the function $h(\tau, u)$ has

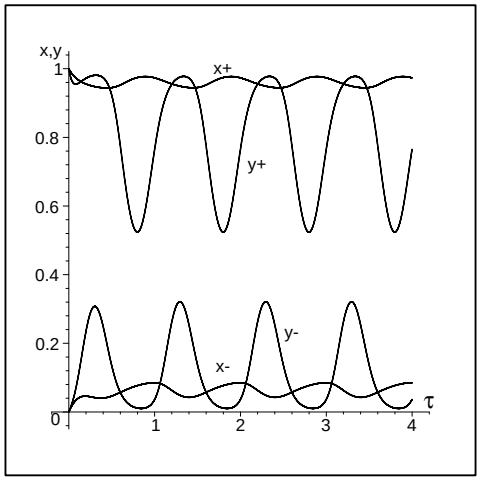


FIG. 5.1. *Solutions* (x, y) *in Region* 1*:* $\theta_x = 3$*,* $b = 1.8$.

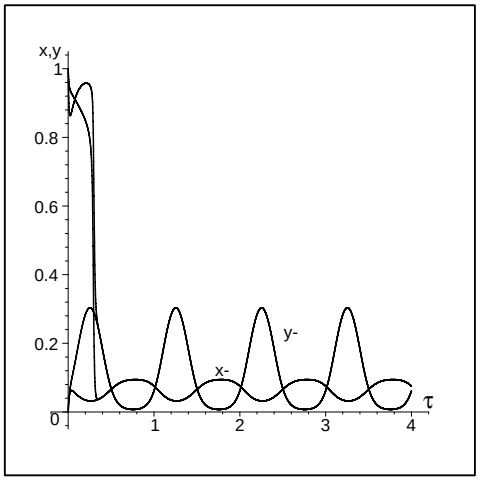


FIG. 5.2. *Solutions* (x, y) *in Region* 2*:* $\theta_x = 3$*,* $b = 2.2$.

its point of inflection. We arbitrarily set $\theta_y = 3$. The resulting Wilson–Cowan system is

$$\begin{cases} x' &= -x + \mathbf{S}(8x - by - \theta_x), \\ y' &= -y + \mathbf{S}(0.8(8x - by) - 3 + \alpha \sin(\omega t)). \end{cases} \tag{5.2}$$

Region 1. With $\theta_x = 3$ and $b = 1.8 < b_2$, we know from Theorem 3.1 that for *any* input $\alpha \sin(\omega t)$, with small enough frequency ω, almost every solution $(x(t), y(t))$ will approach one of the two attracting states $(x^-, y^-), (x^+, y^+)$, depending only on the values of $x(0)$ and $y(0)$. Figure 5.1 shows solutions of (5.2) with $b = 1.8$ and $i(t) = 2.5 \sin(0.5t)$. Two initial conditions, $(x(0), y(0)) = (0, 0)$ and $(x(0), y(0)) = (1, 1)$, are used, and the corresponding solutions are attracted, respectively, to the lower and upper periodic solutions. In this case, simulation shows that, except for the change in period, the periodic solutions are very similar for *all* values of ω.

Region 2. For the same threshold value $\theta_x = 3$, with $b = 2.2 > b_2$ and $i(t) = 2.5 \sin(0.05t)$, the same two sets of initial conditions were used. It can be seen in Figure 5.2 that, with $\alpha = 2.5$, the frequency $\omega = 0.05$ is small enough to drive *all* solutions to the lower attractor (x^-, y^-); that is, $2.5 > \alpha^*$ and $0.05 < \omega^*(2.5)$.

The parameter values used in the examples described here are of the same order of magnitude as those used by Tsodyks et al. [12]; therefore, assuming that the unit of time is milliseconds, the frequency $\omega = 0.05$ corresponds to a periodic input of approximately 8 Hertz, whereas $\omega = 0.5$ corresponds to an input of 80 Hertz. In the case in which $\theta_x = 3$ and $b = 2.2$, further simulations were done to find the bifurcation value of ω when α was equal to 2.5. The upper periodic solution lost stability at $\omega \approx 0.085$. This means that the periodic input must cycle at approximately 13 Hertz or less to cause all of the solutions to tend to the lower attractor (x^-, y^-), that is, to turn off the excitation.

Region 3. When $\theta_x = 2.8 < \theta_1$, the corresponding value of $b_2 = \theta_2 - \theta_x \approx 2.2657$. Now the value $b = 2.1$ is less than b_2 and greater than $b_1 = \theta_1 - \theta_x \approx 0.0134$. With initial conditions as before, and $i(t)$ again taken to be equal to $2.5 \sin(0.05t)$, Figure 5.3 shows both solutions being driven to the upper attractor (x^+, y^+). By simulation,

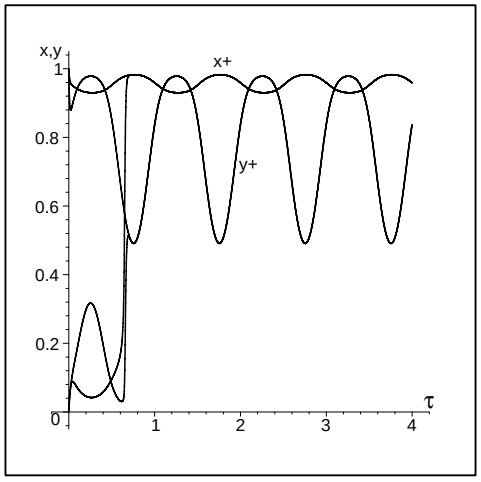


FIG. 5.3. *Solutions* (x, y) *in Region* 3*:* $\theta_x = 2.8$*,* $b = 2.1$.

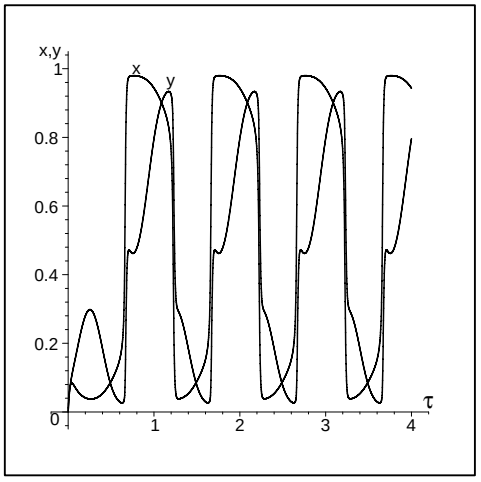


FIG. 5.4. *Wide oscillation in Region* 4*:* $\theta_x = 2.8$*,* $b = 2.5$.

the bifurcation value in this case was found to be $\omega^*(2.5) \approx 0.247$. This corresponds to an input of about 25 Hertz.

Region 4. In Figure 5.4, $\theta_x = 2.8$ and $b = 2.5 > b_2$, and with the same input $i(t)$ one sees large oscillations; that is, the solutions oscillate alternately between the high and low activity states. It can be seen in Figure 5.4 that $x(t)$ and $y(t)$ are oscillating *in phase* in this region, whereas in the other three regions they can be seen to oscillate nearly 180 degrees out of phase. This tends to reinforce the paradoxical behavior observed by Tsodyks et al. [12]. For this set of parameter values, additional simulations show that the lower solution loses stability at $\omega \approx 0.223$, and then the upper solution loses stability at $\omega \approx 0.156$. For any $\omega < 0.156$, the system has the oscillatory behavior shown in Figure 5.4.

A sensitivity analysis was done in each of the four regions by varying d above and below the value λb. Decreasing d all the way to zero has no significant effect on the behavior of the system in any of the regions. However, when d is increased, there is a reduction in the general level of activity in I. This causes the upper solution to regain stability in Region 2 after a 20 percent increase in d, and in Region 4 after an increase of 35 percent in d. An increase in d produces no significant change in Regions 1 and 3.

It is interesting to note from the above results that if b is approximately equal to $\theta_2 - \theta_1$, periodic forcing of low frequency to the inhibitory cells can be used to drive the system to *either* its high or low activity state, depending only on whether the excitatory threshold θ_x is above or below θ_1. If one assumes that, after a period of high excitation, θ_x increases (for example, if we had included a refractory period for the excitatory cells, it might have had this effect), then it can be shown that, in the range of parameter space where $b \approx \theta_2 - \theta_1$, the system can be forced to produce "bursting" activity. To demonstrate this by simulation, the system (5.2) with $b = 2.1$, $i(t) = 2.5\sin(0.05t)$, and a third differential equation for θ_x of the form

$$\frac{d\theta_x}{dt} = K_1(2.7 - \theta_x(t)) + \delta(x(t) - 0.5)[K_2(3.2 - \theta_x(t)) - K_1(2.7 - \theta_x(t))]$$

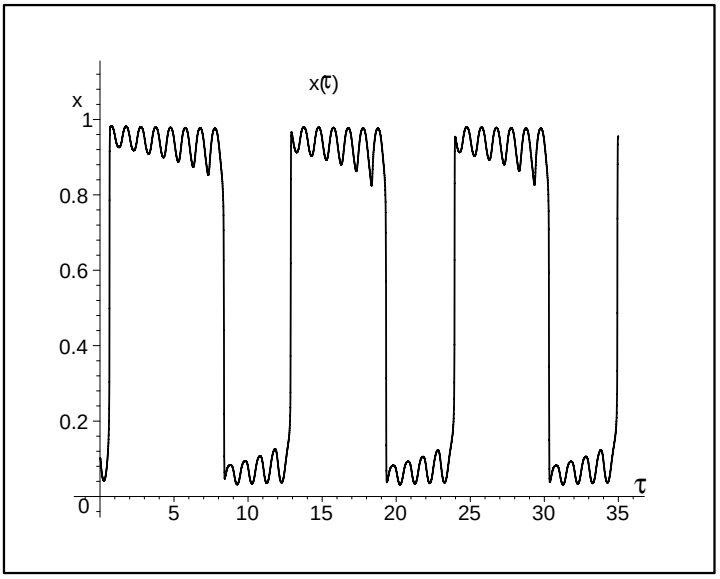


FIG. 5.5. *Bursting response to input, with variable* θ_x.

was solved numerically using MAPLE. The delta function used in the calculation is $\delta(z) = 1/(1 + \exp(-50z))$, and K_1 and K_2 were arbitrarily chosen to be 0.15. The value of θ_x was initialized to 2.8. While the excitatory activity x is greater than 50 percent, θ_x increases toward 3.2, but it then decreases again when the activity becomes less than 50 percent. A solution of this system is shown in Figure 5.5. Further simulations also showed that if d is varied about the value $\lambda b = 1.68$, the bursting activity persists, but the length of the bursts increases as d increases. In contrast, increasing the amplitude α of the forcing function shortened the length of the bursts.

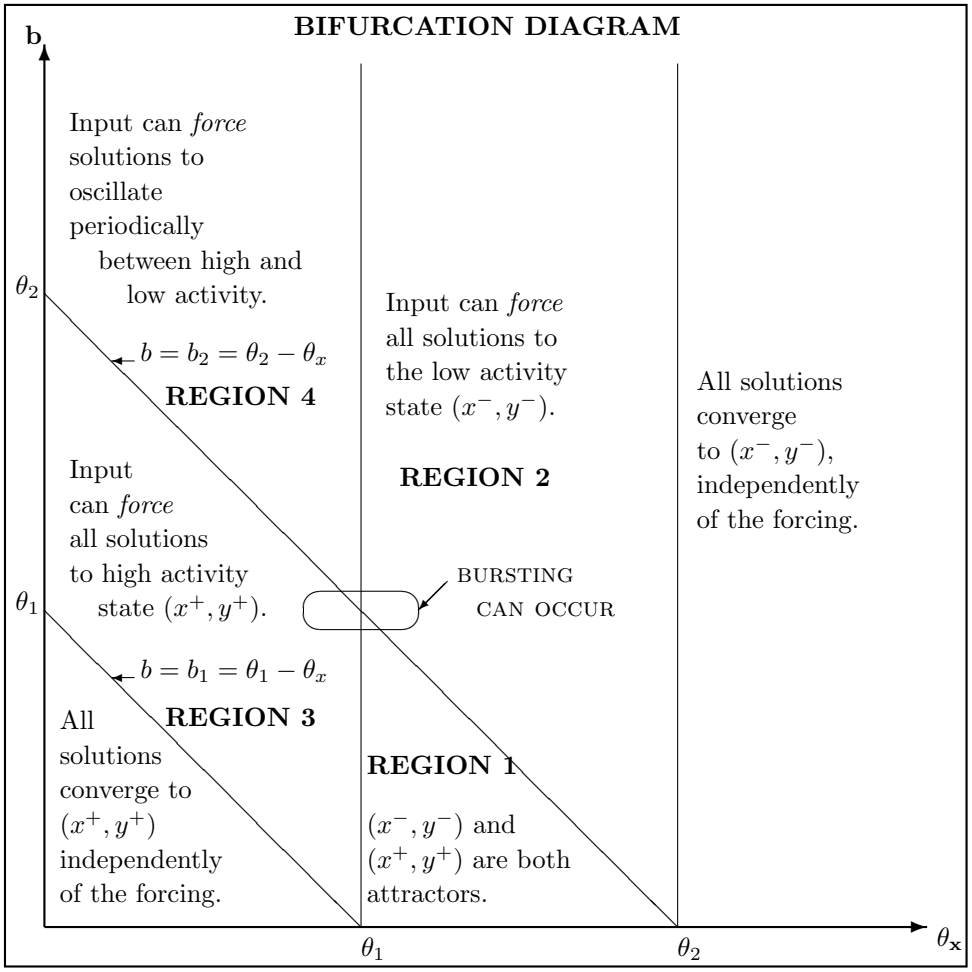
6. Summary. By reducing the Wilson–Cowan system to a single first-order differential equation, one is able to bring several known results for periodic first-order differential equations to bear on the problem of the behavior of the forced two-dimensional system.

Our results can be summarized in the bifurcation diagram below, using b and θ_x as the two bifurcation parameters. In all cases it is assumed that $a > 1/\mathbf{S}'(0)$, $\lambda \equiv c/a < \Lambda$, $d/b \approx c/a$, and the system is being forced with *low* frequency periodic input.

Note that in only three of the six regions of parameter space is it possible to use periodic input to the inhibitory cells to force the system to alter its behavior significantly, and then only if the frequency of the input is low and its amplitude is large enough. The bursting behavior, noted in the bifurcation diagram, appears to be a new result for a system of this type.

It seems clear that our results hold not only for the input function $i(t) = \alpha \sin(\omega t)$ but also for any P-periodic function which has a single maximum and single minimum between 0 and P. This will be an area of further study, to determine just exactly what type of periodic input functions produce the effects noted in this paper.

Another problem for further study involves the question of whether or not the first-order periodic equation (2.4), with parameters in our parameter range, can *ever*



have more than three periodic solutions for *any* values of ω. In the slope field region between the two invariant subregions $\mathcal{H}_1$ and $\mathcal{H}_2$ it seems to be the case, from simulations, that only the single periodic repeller exists. This is definitely true if λ is small enough, and it can be proved to be true in general for small ω and also for large ω, but in a middle region both proofs break down. For a slope function $f(\tau, u)$ which is a cubic in u, with coefficient of u^3 negative for all τ, there is a theorem (see [2], [6]) which shows that the differential equation can have *at most* three periodic solutions. Our slope function is similar to a cubic, for $-b < u < a$, in the sense that it has two turning points, a local minimum and then a local maximum. There may be a way to extend the theorem to such functions, but this has so far proved difficult to do.

Appendix. Properties of solutions of $y' = f(t, y)$. This section contains statements of some well-known properties (see, for example, [3], [4]) of solutions of the first-order differential equation

$$\frac{dy}{dt} = f(t, y). \tag{A.1}$$

A solution of an initial-value problem for (A.1) will be denoted by $\phi \equiv \phi(t, t_0, y_0)$, where $\phi(t_0, t_0, y_0) = y_0$ and $d\phi/dt \equiv f(t, \phi(t, t_0, y_0))$ for $t \geq t_0$. We will assume that

the slope function $f(t, y)$ is continuous and continuously differentiable in y; therefore, for any initial value the solution of (A.1) is unique. Furthermore, if $y_2 > y_1$, then the uniqueness and existence theorem for first-order differential equations implies that $\phi(t, t_0, y_2) > \phi(t, t_0, y_1)$ for all $t > t_0$.

DEFINITION A.1. *A* forward invariant region *for* (A.1), *in the* (t, y)*-plane, is a region* $V = \{t_0 \le t < \infty, v_1(t) < y < v_2(t)\}$, *where* v_1 *and* v_2 *are continuously differentiable curves such that* $f(t, v_1(t)) > v_1'(t)$ *and* $f(t, v_2(t)) < v_2'(t)$ *for all* $t \ge t_0$. *This implies that the vector field* f *is entering the region* V *everywhere along the curves* v_1 *and* v_2.

Once a solution of (A.1) enters a forward invariant region V, it can never cross the curves v_1 or v_2, and therefore it must remain in V for all $t > t_0$.

DEFINITION A.2. *A forward invariant region* V *for* (A.1) *is called* contracting *if* $f_y(t, y) \le -\delta < 0$, *for some positive constant* δ, *for all* $(t, y) \in V$.

DEFINITION A.3. *The slope function* $f(t, y)$ *is called* periodic of period P *if* $f(t + P, y) = f(t, y)$ *for all* t.

LEMMA A.4. *If the slope function* $f(t, y)$ *in* (A.1) *is periodic of period* P *and* V *is a contracting forward invariant region for* (A.1), *then there exists a unique period-*P *attractor* $\hat{\phi}$ *in* V, *and all solutions of* (A.1) *that enter* V *are attracted to* $\hat{\phi}$ *as* $t \to \infty$.

For a nice proof of this lemma, see p. 115 in [3].

If W is a contracting forward invariant region for (A.1) when the independent variable t is replaced by $-t$, a similar proof can be used to show that there exists a unique *repelling* solution for (A.1) in W.

Acknowledgments. The authors would like to thank the following people for their helpful suggestions: Zbigniew Nitecki, Clark Robinson, Paul Bugl, and Mako Haruta.

REFERENCES

[1] J. EGGERT AND J. L. VAN HEMMEN, *Modeling neuronal assemblies: Theory and implementation*, Neural Comput., 13 (2001), pp. 1923–1974.

[2] A. GASULL AND J. LLIBRE, *Limit cycles for a class of Abel equations*, SIAM J. Math. Anal., 21 (1990), pp. 1235–1244.

[3] J. HALE AND H. KOÇAK, *Dynamics and Bifurcations*, Texts Appl. Math. 3, Springer-Verlag, New York, 1991.

[4] M. W. HIRSCH AND S. SMALE, *Differential Equations, Dynamical Systems, and Linear Algebra*, Academic Press, San Diego, CA, 1974.

[5] F. C. HOPPENSTEADT AND E. M. IZHIKEVICH, *Weakly Connected Neural Networks*, Appl. Math. Sci. 126, Springer-Verlag, New York, 1997.

[6] A. L. NETO, *On the number of solutions of the equation* $dx/dt = \sum_0^n a_j(t)x^j, 0 \le t \le 1$, *for which* $x(0) = x(1)$, Invent. Math., 59 (1980), pp. 67–76.

[7] J. PALIS, *On Morse–Smale dynamical systems*, Topology, 8 (1969), pp. 385–404.

[8] J. PALIS AND S.SMALE, *Structural stability theorems*, in Global Analysis, Proceedings of the 14th Symposium in Pure Mathematics, AMS, Providence, RI, 1970, pp. 223–231.

[9] J. PALIS AND W. DE MELO, *Geometric Theory of Dynamical Systems: An Introduction*, Springer-Verlag, New York, 1982.

[10] C. ROBINSON, *Dynamical Systems: Stability, Symbolic Dynamics, and Chaos*, CRC Press, Boca Raton, FL, 1995.

[11] C. ROBINSON, *Structural stability on manifolds with boundary*, J. Differential Equations, 37 (1980), pp. 1–11.

[12] M. V. TSODYKS, W. E. SKAGGS, T. J. SEJNOWSKI, AND B. L. MCNAUGHTON, *Paradoxical effects of external modulation of inhibitory interneurons*, J. Neurosci., 17 (1997), pp. 4382–4388.

[13] P. WHITTLE, *Neural Nets and Chaotic Carriers,* John Wiley, London, 1998.

[14] H. R. WILSON AND J. D. COWAN, *Excitatory and inhibitory interactions in localized populations of model neurons*, Biophys. J., 12 (1972), pp. 1–24.

SIAM J. APPL. MATH.
Vol. 63, No. 5, pp. 1604–1614

OSCILLATORY FLOW NEAR A STAGNATION POINT*

M. G. BLYTH† AND P. HALL‡

Abstract. The classical Hiemenz solution describes incompressible two-dimensional stagnation point flow at a solid wall. We consider an unsteady version of this problem, examining particularly the response close to the wall when the solution at infinity is modulated in time by a periodic factor of specified amplitude and frequency. While this problem has already been tackled in the literature for general frequency in cases when the amplitude of the time-periodic factor is either large or small, we compute the flow for arbitrary values of both these parameters. For any given amplitude, we find that there exists a threshold frequency above which the flow is regular and periodic, with the same period as the modulation factor, and beneath which the solution terminates in a finite time singularity. The dividing line in parameter space between these two possibilities is identified and favorably compared with the predictions of asymptotic analyses in the small and large frequency limits.

Key words. stagnation point, boundary layer, steady streaming

AMS subject classification. 76D10

DOI. 10.1137/S0036139902408175

1. Introduction.

The classical Hiemenz solution describes the flow near a stagnation point on a plane wall and may be found in Batchelor [1]. When this flow is modulated in time by a periodic multiplicative factor on the streamfunction at infinity, the resulting solution can be used to describe the local dynamics around a stagnation point on an oscillating body. Equally, the body may be thought of as fixed, with the flow out in the far field varying periodically in time. The steady streaming motion established by the Reynolds stresses associated with the oscillatory motion (Stuart [2]) is an important feature of such time-periodic flows. This effect occurs, for example, around a transversely oscillating circular cylinder (Schlichting [3]). When the amplitude of the far field fluctuation is small, modified Hiemenz flow can be used to model the local effects of disturbances such as acoustic noise impinging on the boundary layer around a translating bluff body. More generally, such problems fall within the purview of receptivity theory (e.g., Erturk and Corke [4], Morkovin [5]). We are concerned with the behavior in the vicinity of the stagnation point, where the body surface may be considered to be locally flat. In this context we allow for fluctuations of arbitrary amplitude and frequency.

Other studies pertinent to this modified Hiemenz problem include those by Grosch and Salwen [6] and Merchant and Davis [7]; among earlier investigations, we mention those by Matunobu [8], Pedley [9], and Ishigaki [10]; see also Lighthill [11]. Grosch and Salwen, while confining their attention to small free stream fluctuations, examined the flow when the frequency of these disturbances is either small or large. Expanding in power series in the disturbance amplitude and Fourier series in time, they showed that to leading order the low frequency case is merely a quasi-steady version of the classical Hiemenz solution, while the high frequency case exhibits a double boundary layer structure similar to that first discussed by Riley [12] and Stuart [2] for oscillating

*Received by the editors May 20, 2002; accepted for publication (in revised form) December 3, 2002; published electronically June 12, 2003.
http://www.siam.org/journals/siap/63-5/40817.html
†School of Mathematics, University of East Anglia, Norwich, UK, NR4 7TJ (m.blyth@uea.ac.uk).
‡Mathematics Department, Imperial College, 180 Queen's Gate, London, UK, SW7 2BZ (philhall@ic.ac.uk).

flows. Merchant and Davis tackled the same unsteady Hiemenz problem, but also examined the flow structure when the mean component of the free stream is very much smaller than the oscillatory part. In this case a double boundary layer structure is once again revealed, although the authors showed that no solutions exist when the mean component of the free stream drops below a certain cut-off point. This is not to say that solutions with a different asymptotic form do not exist at smaller values of the free stream mean. We address this point in the current work.

Our interest in the problem was prompted by Hall and Papageorgiou's [13] recent study of unsteady incompressible flow induced in an infinite channel when the walls pulsate uniformly in space and periodically in time. Assuming a stagnation point structure for this flow, they demonstrated numerically the existence of purely periodic, quasi-periodic, and even chaotic flow solutions, depending on the frequency and amplitude of the wall motion. Our aim was to investigate the possibility of such varied dynamics for a periodically forced stagnation point flow in a semi-infinite domain. In fact, numerical solution of the unsteady Hiemenz problem shows that for many parameter values a singularity is encountered at a finite time. This eventuality is perhaps unsurprising in light of previous studies of colliding boundary layers or those in which the external flow reverses direction, where such singularities may also be found. One example is the flow over a rotating disk in a counter-rotating fluid, whose near-singularity structure was described by [14]. Another study of particular relevance to the current work is that by Riley and Vasantha [15], who considered the same problem as ours but with a purely oscillatory (zero mean) free stream. In this case the equations break down in finite time for any value of the forcing frequency. The breakdown was interpreted by Riley and Vasantha as the result of drifting fluid particles in the steady streaming layer accumulating at the stagnation point and ultimately causing an eruption of fluid from the boundary layer. The nature of the finite time singularity was found to have the same form as that occurring near the equator of an impulsively started sphere, as studied by Banks and Zaturska [16]. When the flow in the far field has nonzero mean, as is to be discussed here, the near-singularity structure is also described by Banks and Zaturska's analysis. The blow-up is not localized in space but occurs over the entire flow domain.

That the flow can break down when the mean component of the free stream is nonzero is not mentioned by either Grosch and Salwen or Merchant and Davis. However, the former authors demonstrate that the inclusion of a nonzero mean component in the free stream can allow the solution to be continued indefinitely without breakdown. We have found that this is true, provided that a condition between the fluctuation amplitude and frequency is not violated. The condition amounts to a threshold frequency, at any given amplitude, below which blow-up will occur but above which the solution remains regular. It is still conceivable that aperiodic or even chaotic solutions might exist in the large frequency limit, so long as this condition is satisfied. However, we have not been able to identify any such solutions despite extensive numerical searches. At all candidate parameter values tested, the solutions remain periodic with frequency equal to that of the free stream fluctuation. Nevertheless, the nature of the condition under which breakdown occurs is of interest, and in this sense our work constitutes a worthwhile extension of the previous studies.

We begin with a problem description, followed by a brief discussion of the numerical methods utilized to solve the governing equations. Results are then presented together with asymptotic analyses in the small and large frequency limits, and comparison is made between the two. We conclude with a short note on the axisymmetric version of the problem.

2. Problem statement and numerical method. The problem under consideration is that of two-dimensional unsteady Hiemenz flow approaching a flat plate. Referring to a set of Cartesian axes (x, y), the flat plate occupies $-\infty < x < \infty$, $y = 0$. We define velocity components $U_\infty u(x, y, t)$, $U_\infty v(x, y, t)$ in the x, y directions, respectively, where U_∞ is the typical flow speed in the far field. In this region the flow is potential, with $u = (x/l)U(t)$, $v = -(y/l)U(t)$, where

$$U(t) = 1 + \Delta \cos(\omega t) \tag{2.1}$$

for a chosen amplitude Δ and frequency ω. The remaining parameters are the kinematic viscosity ν and an arbitrarily chosen length scale l.

As in the steady case, it is reasonable to assume that the same simple velocity dependence on the x coordinate also applies in the viscous layer close to the plate. In this region we introduce the new coordinate $\eta = (y/l)R^{1/2}$ and set

$$u = \left(\frac{x}{l}\right) F_\eta(\eta, \tau), \qquad v = -R^{-1/2} F(\eta, \tau)$$

defining the Reynolds number to be $R = U_\infty l/\nu$ and introducing the new time variable $\tau = \omega t$. Note that we do not require R to be large. Finally, defining the Strouhal number

$$\sigma = \frac{\omega l}{U_\infty}$$

and setting $a(\tau) = 1 + \Delta \cos \tau$, we may express the wall layer system as

$$\sigma F_{\eta\tau} + F_\eta^2 - F F_{\eta\eta} = \sigma a_\tau + a^2 + F_{\eta\eta\eta}, \tag{2.2}$$

with

$$F(0, \tau) = F_\eta(0, \tau) = 0, \qquad F_\eta \to a(\tau) \quad \text{as} \quad \eta \to \infty, \tag{2.3}$$

to satisfy the solid wall boundary conditions and to match to the outer potential solution. When $\Delta = 0$, the problem reduces to that of classical steady Hiemenz flow. The temporally periodic part of (2.1) represents a superimposed disturbance on the steady far field solution. The response to this disturbance close to the plate is quantified by solving (2.2) and (2.3) for different values of the fluctuation amplitude Δ and the Strouhal number σ. We emphasize that Δ is not restricted to being small.

At this stage we note that functions satisfying (2.2) and (2.3) represent exact solutions of the Navier–Stokes equations since no approximations have been made. In addition, while the Reynolds number is not required to be large in this analysis, for ease of description we shall refer to the main flow governed by (2.2) as the Hiemenz boundary layer, even though no conventional boundary layer approximation is necessary.

Asymptotic solutions are possible in the limits of small and large frequency, and these will be discussed in a later section. For general values of the parameters (Δ, σ) numerical methods must be used to solve the wall layer equations. To expedite the numerical solution, we introduce the function $G = F_\eta$ and write the equations in the form

$$\sigma G_\tau + G^2 - F G_\eta = \sigma a_\tau + a^2 + G_{\eta\eta}, \tag{2.4a}$$

$$G = F_\eta, \tag{2.4b}$$

with

$$F(0,\tau) = G(0,\tau) = 0, \qquad G \to a(\tau) \quad \text{as} \quad \eta \to \infty. \tag{2.5}$$

We found that calculations should be started at $\tau = \pi/2$ rather than $\tau = 0$; otherwise, when $\Delta \gg 1$, the forcing is too large when the integration is initiated and the numerical solution blows up instantaneously. In practice, the initial profile makes negligible difference to the computation of the singular time, and calculations were always begun with $F = G = 0$. To march forwards in time we used either the second order accurate Crank–Nicholson method or the fourth order accurate Runge–Kutta integration, both with a second order accurate spatial discretization. Thus some independent numerical check on our results was available. Most of the results presented in this paper were computed using the latter of the two schemes. However, at various stages the computations were repeated using the Crank–Nicholson method, and these always provided good agreement. For large σ, we found it useful to introduce a stretched grid in order to insert many more points in the Stokes shear-wave layer next to the wall, where the most significant changes in the flow are concentrated. We present some of our results in the next section.

3. Analytical discussion and results. Grosch and Salwen [6] show that, for sufficiently small amplitude and in the limit of vanishing Strouhal number, the quasi-steady Hiemenz solution describes the flow to leading order. Specifically, to $O(\sigma)$, the solution may be written as $F = a(\tau)^{1/2} f(a^{1/2}\eta)$, where f satisfies the usual steady Hiemenz equation and boundary conditions. When $\sigma \to \infty$, again for small enough Δ, the solution adopts a double boundary layer structure, similar to that analyzed by Riley [12] and Stuart [2]. In this case the boundary layer splits into two regions. In the lower region, the Stokes layer, the solution is periodic to leading order. The nonlinear terms in the equations generate a small steady component, which persists to the upper reaches of the Stokes layer and acts to drive a steady streaming flow in the outer part of the Hiemenz boundary layer. For both small and large σ, Grosch and Salwen expanded in Taylor series in the amplitude Δ and Fourier series in time. While they were not able to determine the radius of convergence of their series exactly, they estimated that for large σ the series should converge when $\Delta < \sigma$. Later Merchant and Davis [7] showed that if both the amplitude and the Strouhal number are large, and if the thickness of the main boundary layer and the induced steady-streaming layer are chosen to coincide, then no solutions exist in our notation when $\Delta > 1.289\,\sigma^{1/2}$. This derives from the fact that the leading order steady-streaming equation has no solution when the amplitude exceeds this bound. However, this does not deny the existence of other solutions with a different boundary layer structure in these limits. In the Merchant and Davis flow structure the Stokes layer is linear to leading order. If instead we hypothesize that solutions exist wherein the nonlinear terms are of the same order of magnitude as the unsteady terms, corresponding to the scaling $\Delta \sim \sigma$ as $\sigma \to \infty$, we find that the equations for the leading order Stokes layer problem, with the appropriate matching condition at infinity, are the same as those studied by Riley and Vasantha [15]. These correspond to (2.2) and (2.3) with $a(\tau) = \cos\tau$. Riley and Vasantha's results show that these slightly reduced equations terminate in a singularity at a finite time for all values of σ. In due course we shall present numerical evidence that no regular solutions exist above the limit laid down by Merchant and Davis.

As a preliminary test of our codes, we computed small amplitude solutions and obtained results in excellent agreement with those of Grosch and Salwen. For large σ,

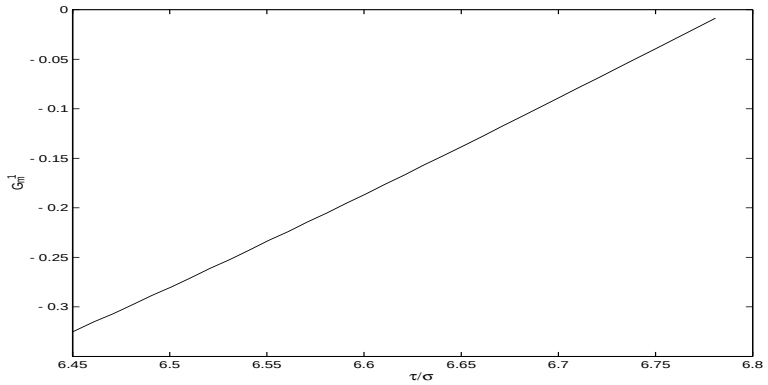


FIG. 3.1. $\Delta = 2.0$, $\sigma = 0.5$: *The locus of* $G_m^{-1} = 1/\min_\eta\{G\}$ *in time,* τ*, close to the singularity at* $\tau_s = 3.39$.

and for Δ of typical size $O(\sigma^{1/2})$, we compared our results with the first two terms in the asymptotic expansion of Merchant and Davis and again found that our numerics were in excellent agreement with the theory. For a fixed σ, however, we discovered that when the value of Δ exceeds a certain limit, the integration blows up at a finite time singularity. We shall henceforth label the singular time τ_s. The flow structure in the vicinity of τ_s is the same as that arising near the equator of an impulsively started sphere, a problem analyzed by Banks and Zaturska [16]. As the singularity is approached in that situation, the most important terms in the near-equator equations over the main part of the flow combine to mimic those of our own equation. Close to the singular time the flow acquires a three-tiered structure, with viscous zones at the wall and infinity sandwiching an inviscid core region. In this sense the flow is similar to that attained just prior to the breakdown which occurs a short time after the direction of a rotating disk is suddenly reversed. The near-singularity structure for this flow has been examined by Stewartson and Bodonyi [17] and corrected by Stewartson, Simpson, and Bodonyi [14]. (See also Ockendon [18] for a discussion of a steady rotating disk flow with a similar three-zone structure.) While Banks and Zaturska do not give the details of the flow in the upper and lower viscous regions, we have confirmed that the arguments of Stewartson, Simpson, and Bodonyi may be adapted accordingly. Further discussion of the viscous zones is suppressed. Instead we demonstrate that the flow behavior in the middle region is consistent with that of Banks and Zaturska. By comparison with their theory, in the main part of the flow we expect η to scale like $(\tau_s - \tau)^{-1/2}$, and thus we write $\widehat{\eta} = \eta(\tau_s - \tau)^{1/2}$. From their predictions we anticipate that $F \propto (\tau_s - \tau)^{-3/2}\phi(\widehat{\eta})$, and thus $G \propto (\tau_s - \tau)^{-1}$ as $\tau \to \tau_s^-$. As time progresses we observe that, at a given τ, the $G(\eta, \tau)$ profile has at most one local minimum G_m at $\eta = \eta_m$. Tracking the inverse of this minimum value up to the singular time for the case $\Delta = 2.0$, $\sigma = 0.5$, we plot the graph shown in Figure 3.1. The relationship between G_m^{-1} and τ appears convincingly linear. Assuming this to be the case, a more accurate value for τ_s may be predicted by means of linear interpolation. Similarly, plotting η_m^{-2} against τ close to the singular time reveals an apparent linear dependence which is equally compelling. The critical point dividing singular and periodic solutions at this frequency is $\Delta \approx 1.537$. As the amplitude approaches this value from above, we find that the singular time τ_s increases without bound.

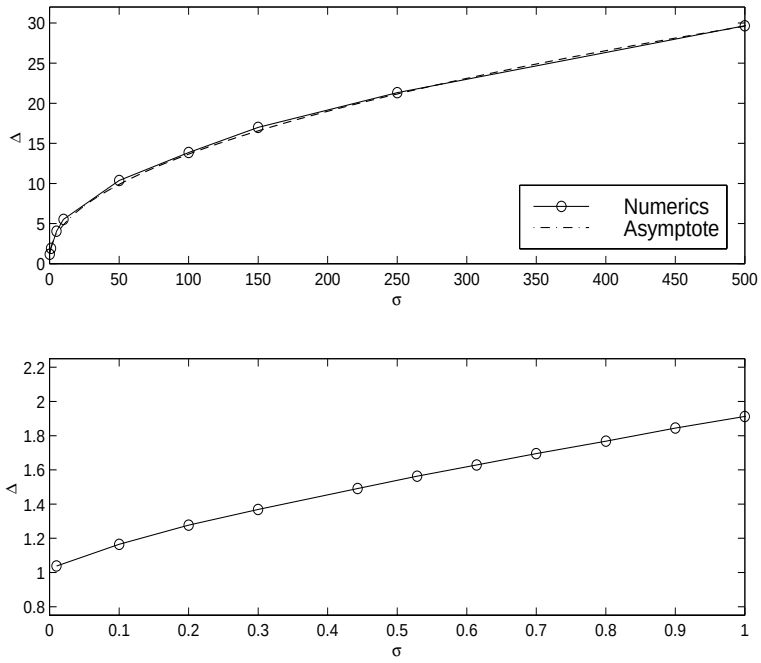


FIG. 3.2. *The barrier in* (Δ, σ) *parameter space between regular periodic solutions (below the line) and those which blow up in finite time (above the line). The lower graph shows the barrier close to* $\sigma = 0$. *The upper graph includes the asymptotic approximation* $\Delta \sim 1.289\sigma^{1/2} + 0.76$ *for large* σ *(see section* 3.1*).*

We now turn to the general picture in (Δ, σ) parameter space. Using the numerical procedure described above, we have identified the parametric regions in which the solutions either remain regular and periodic or else become singular in finite time. The parameter space is divided neatly into two regions by the curve depicted in Figure 3.2. We believe that points on this curve have been calculated accurately to within 0.1%. Flows corresponding to points (Δ, σ) lying below the curve are periodic, while those lying above eventually reach a singularity. We have calculated solutions up to and including $\sigma = 1000$ and, for the regular solutions, have encountered no bifurcations leading to aperiodic flows. Rather, the solutions remain periodic with the same period as that of the far field disturbance.

Also shown in Figure 3.2 is the asymptotic approximation valid for large values of σ, to be discussed shortly. Magnifying the curve in the region of zero frequency shows that it approaches unity as $\sigma \to 0$, suggesting that, for temporally slowly varying flows of this kind, only a small amount of flow reversal can be tolerated if the solution is to remain regular. This limit is discussed in due course.

When the disturbance frequency is large, the boundary layer supports a small steady flow component, driven by a residual slip velocity from a Stokes layer beneath. The analysis of Merchant and Davis suggests that no regular solutions exist when the amplitude is also large, specifically when $\Delta > 1.289\sigma^{1/2}$, since then the system governing the steady streaming component has no solution. In order to provide the best agreement between our numerical calculations and the large amplitude theory, we have found it worthwhile to compute the next term in this asymptotic expansion.

In what follows, we therefore supply a brief description of the asymptotic flow at large frequencies.

3.1. Large amplitude and frequency. It is convenient at this stage to rescale (2.2), (2.3) by writing $\eta = \Delta^{-1/2}\widehat{\eta}$, $F = \Delta^{-1/2}\widehat{F}$, $\sigma = \Delta\Omega$ and introducing the small parameter $\varepsilon = 1/\Delta$. This leads to the equivalent system:

$$\Omega\widehat{F}_{\widehat{\eta}\tau} + \widehat{F}_{\widehat{\eta}}^2 - \widehat{F}\widehat{F}_{\widehat{\eta}\widehat{\eta}} = -\Omega\sin\tau + (\varepsilon + \cos\tau)^2 + \widehat{F}_{\widehat{\eta}\widehat{\eta}\widehat{\eta}}, \tag{3.1}$$

with

$$\widehat{F}(0,\tau) = \widehat{F}_{\widehat{\eta}}(0,\tau) = 0, \quad \widehat{F}_{\widehat{\eta}} \to \varepsilon + \cos\tau \quad \text{as} \quad \widehat{\eta} \to \infty. \tag{3.2}$$

Now, when ε is small and Ω is large, we look for solutions with $\varepsilon = a_0\Omega^{-1} + a_1\Omega^{-2} + \cdots$. In the Stokes layer, of thickness $O(\Omega^{-1/2})$, the expansion proceeds as

$$\widehat{F} = \Omega^{-1/2}\phi_0(\xi,\tau) + \Omega^{-3/2}\phi_1(\xi,\tau) + \cdots, \tag{3.3}$$

where $\xi = \Omega^{1/2}\widehat{\eta}$ is a scaled coordinate normal to the wall. The first order solution is given by

$$\phi_0(\xi,\tau) = \xi\cos\tau - \cos\left(\tau - \frac{\pi}{4}\right) + \mathrm{e}^{-\xi/\sqrt{2}}\cos\left(\tau - \frac{\xi}{\sqrt{2}} - \frac{\pi}{4}\right). \tag{3.4}$$

At second order, the solution may be written as $\phi_1(\xi,\tau) = \frac{1}{2}\phi_M(\xi,\tau)$, where ϕ_M is a somewhat lengthy expression appearing as formula (3.20c) of Merchant and Davis's paper. As pointed out by Stuart [2], it is not possible to satisfy the infinity condition at this order; rather a steady slip velocity persists at the top of the Stokes layer, driving a steady streaming motion above. Therefore we simply note at this stage that $\phi_{1\xi}(\infty,\tau) = -3/4$.

With the current choice of scaling, the streaming layer has the same thickness as the Hiemenz boundary layer (of order $\Omega^{1/2}$ in this notation). Introducing the new coordinate $\zeta = \Omega^{-1/2}\widehat{\eta}$, the relevant expansion is

$$\widehat{G} = \Omega^{-1/2}\{\psi_0(\zeta,\tau) + f_0(\zeta)\} + \Omega^{-3/2}\{\psi_1(\zeta,\tau) + f_1(\zeta)\} + \cdots, \tag{3.5}$$

where $\widehat{G} = \widehat{F} - \Omega^{1/2}\zeta\cos(\tau) - \Omega^{-1/2}\cos(\tau - \pi/4)$. The functions ψ_i equal zero when averaged over a single time period. In fact $\psi_0 \equiv 0$, and the first order streaming problem is given by

$$f_0''' + f_0 f_0'' - f_0'^2 + a_0^2 = 0, \tag{3.6}$$

with

$$f_0(0) = 0, \quad f_0'(0) = \frac{-3}{4}, \quad f_0'(\infty) = a_0. \tag{3.7}$$

A numerical treatment of this problem by both Merchant and Davis and also by Riley and Weidman [19] indicates that a unique solution exists when $0 < a_0 < 3/4$, two solutions exist when $3/4 < a_0 < a_{0c}$, and no solutions exist when $a_0 > a_{0c}$, where $a_{0c} \approx 0.602$. Continuing, we derive the second order streaming problem:

$$f_1''' + f_0 f_1'' - 2f_0' f_1' + f_0'' f_1 = \frac{1}{2\sqrt{2}}f_0'' - 2a_0 a_1, \tag{3.8}$$

with

$$f_1(0) = \frac{13}{4\sqrt{2}}, \quad f_1'(0) = 0, \quad f_1'(\infty) = a_1. \tag{3.9}$$

Taking $a_1 = 0$ should reduce the problem to the corresponding equation of Merchant and Davis. However, we remark that this leaves a right-hand side proportional to f_0'', which is absent in that paper and which we believe should be included.

Our concern now is to calculate a_1 when $a_0 = a_{0c}$. Numerical trials suggest that a solution exists for the value $a_1 = -0.55$ and is unique. Reverting to our original notation, the asymptotic approximation for the critical amplitude proceeds as

$$\Delta = 1.29\sigma^{1/2} + 0.76 + O(\sigma^{-1/2}) \qquad \text{as} \quad \sigma \to \infty.$$

This is plotted in Figure 3.2 along with our full numerical results. It provides strong evidence that it is this asymptote which defines the barrier between large frequency solutions remaining regular and periodic and those encountering a finite time singularity.

3.2. Small frequency. We now focus our attention on small frequency solutions. The numerical calculations suggest that the barrier between regular and singular solutions approaches $\Delta = 1$ as $\sigma \to 0$ (see Figure 3.2). Naturally, when $\sigma = 0$ we obtain classical Hiemenz flow and therefore expect steady solutions at any amplitude. In this sense we expect $\sigma \to 0$ to be a singular limit. While considering a similar flow but with zero mean at infinity, Riley and Vasantha [15] showed that the singular time grows without bound as the fluctuation frequency tends to zero. Paralleling their analysis, we now examine the solution close to $\sigma = 0$.

When both the frequency and amplitude are small, the flow follows the quasi-steady solution given by Grosch and Salwen [6] and mentioned above in section 3. However, if the amplitude equals or exceeds unity, $a(\tau)$ has a zero at $\tau = \tau_0$, where $\tau_0 = \pi - \cos^{-1}(1/\Delta)$. In what follows, $\Delta - 1$ is assumed to be nonnegative (but not necessarily small) so that $a(\tau)$ has such a zero. The quasi-steady approximation will break down when τ approaches τ_0, as the unsteady terms, which were hitherto small, grow to become comparable in size with the others. A consideration of the relative magnitudes of terms in (2.2) suggests that the quasi-steady approximation will become invalid when $a(\tau) = O(\sigma^{1/2})$.

In the vicinity of $\tau = \tau_0$, a balance of the terms in (2.2) suggests the scalings

$$(\tau_0 - \tau) = \sigma^{1/2}T, \quad \eta = \sigma^{-1/4}Y, \quad F = \sigma^{1/4}\widetilde{F}(Y,T)$$

for new order one variables T, Y, $\widetilde{F}$. Neglecting terms of order $o(\sigma^{4/3})$, the governing system reduces to

$$-\widetilde{F}_{YT} + \widetilde{F}_Y^2 - \widetilde{F}\widetilde{F}_{YY} = -\mu + \mu^2 T^2 + \widetilde{F}_{YYY}, \tag{3.10a}$$

$$\widetilde{F}(0,T) = \widetilde{F}_Y(0,T) = 0, \quad \widetilde{F}_Y \to \mu T \quad \text{as} \quad Y \to \infty, \tag{3.10b}$$

where $\mu = (\Delta - 1)^{1/2}$. As $T \to \infty$, $\widetilde{F}$ must match to the quasi-steady Hiemenz solution. We initiate the calculation at $T = T_\infty$, where T_∞ is sufficiently large, with the profile

$$\widetilde{F} = \left(\frac{\mu T}{T_\infty}\right)^{1/2} f(x), \quad \text{with} \quad x = \left(\frac{\mu T}{T_\infty}\right)^{1/2} Y,$$

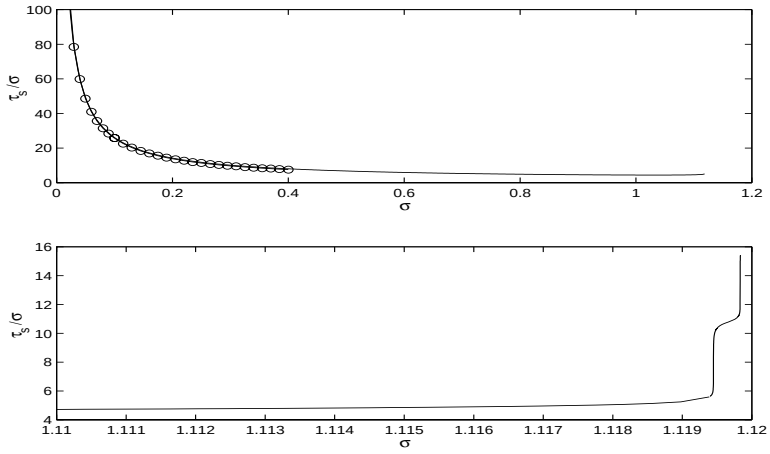


FIG. 3.3. *Variation of τ_s with σ when $\Delta = 2.0$. Top: τ_s in the singular window $\sigma \in (0, 1.12)$, plotted with the approximation $\tau_s/\sigma = 2\pi/3\sigma + 1.51/\sigma^{1/2}$ valid as $\sigma \to 0$ (shown as circles). Bottom: Magnification of the curve close to $\sigma = 1.12$.*

where $f(x)$ satisfies a scaled version of the steady Hiemenz equations. The system (3.10) is then integrated backwards in time T, using a Crank–Nicholson method. For all cases considered, the integration terminates at a singularity after a finite time. The singular time T_s is computed using an interpolation procedure similar to that mentioned above for the full numerics.

For the particular case $\Delta = 2.0$, we have computed $T_s = -1.51$. Thus we predict that

$$\frac{\tau_s}{\sigma} = \frac{2\pi}{3\sigma} + \frac{1.51}{\sigma^{1/2}} + \cdots \quad \text{as} \quad \sigma \to 0. \tag{3.11}$$

In Figure 3.3 we show how this estimate compares with the calculated values of τ_s at small frequency obtained by solving the full system (2.2), (2.3). The agreement is very satisfactory. The flow becomes singular for all frequencies in the window $\sigma \in (0, 1.12)$, beyond which we cross the barrier in Figure 3.2 and the solutions become periodic. It is interesting to note that as $\sigma \to 1.12^-$, the curve begins to wiggle, a behavior reminiscent of that seen in Riley and Vasantha's problem.

In summary of this short section, we remark that low frequency solutions can become singular as long as $\Delta \geq 1$. When $\Delta < 1$, this is not possible, and the flow is quasi-steady and periodic. These conclusions are in agreement with the picture presented in Figure 3.2, where the barrier between regular and irregular solutions approaches unit fluctuation amplitude as the frequency tends to zero.

4. Axisymmetric stagnation point. As a final note, we remark that behavior similar to that described in the previous sections is encountered at an axisymmetric stagnation point. In this case we envisage flow hitting a flat surface and spreading out radially from the stagnation point in the middle. An exact similarity solution with linear dependence in the radial coordinate may then be sought. If we confine our attention to the case in which there is no azimuthal variation, we find that the governing equation and boundary conditions for this flow are almost exactly the same as those for the two-dimensional case, the only difference arising in the nonlinear term,

where $-2FF_{\eta\eta}$ appears instead of $-FF_{\eta\eta}$. By analogy with our preceding work, an amplitude Δ and Strouhal number σ may be defined. Riley [20] has studied this flow in detail when the infinity condition has zero mean, and his results suggest that, in common with the two-dimensional stagnation point looked at by Riley and Vasantha, finite time breakdown occurs for all values of the frequency parameter σ. The flow with nonzero mean at infinity may also be studied and, as in the two-dimensional case, we find that the flow breaks down at a fixed σ as soon as a critical value of Δ is exceeded. A steady-streaming analysis analogous to that performed above again agrees well with the numerically computed results. Unfortunately we once more find no evidence of aperiodic solutions as σ is increased (with Δ remaining below the critical curve for regular solutions).

5. Concluding remarks. We have investigated unsteady stagnation point flow of a modified Hiemenz type. Previously Grosch and Salwen [6] investigated this problem for small fluctuation amplitudes in the low and high frequency limits. Merchant and Davis [7] also considered the large amplitude, high frequency limits, establishing an asymptotic structure in which the streaming region above the Stokes layer is the same thickness as the Hiemenz boundary layer. We have studied the same flow and, for general parameter values, provided numerical evidence that for all frequencies there exists a threshold value of the amplitude beyond which the flow will break down in finite time. The flow structure in the vicinity of the singularity is the same as that arising near the equator of an impulsively started sphere, reported by Banks and Zaturska [16]. Below the threshold value, the solutions are regular and periodic, with period equal to that of the free stream disturbance; in the limit of small frequency, they correspond to the solutions presented by Grosch and Salwen. We have also conducted asymptotic analyses at small and large frequency to predict the dividing line between singular and periodic solutions in these limits. Both have been successfully compared with the results of numerical simulations.

In an earlier study, Riley and Vasantha [15] showed that the same problem with zero mean flow in the free stream breaks down for all possible frequencies. However, when there exists a small mean flow in the free stream, corresponding to the limit of large disturbance amplitude in our problem, our work shows that the solution becomes singular in finite time only when the frequency is smaller than a given value proportional to the square of the amplitude. All frequencies exceeding this value lead to regular, periodic solutions.

REFERENCES

[1] G. K. Batchelor, *An Introduction to Fluid Dynamics*, Cambridge University Press, Cambridge, UK, 1967.

[2] J. T. Stuart, *Double boundary layers in oscillatory viscous flow*, J. Fluid Mech., 24 (1966), pp. 673–687.

[3] H. Schlichting, *Berechnung ebener periodischer Grenzschichtströmungen*, Phys. Z., 33 (1932), pp. 327–335.

[4] E. Erturk and T. C. Corke, *Boundary layer leading-edge receptivity to sound at incidence angles*, J. Fluid Mech., 444 (2001), pp. 383–407.

[5] M. V. Morkovin, *Instability, transition to turbulence and predictability*, AGARDograph, 236, 1978.

[6] C. E. Grosch and H. Salwen, *Oscillating stagnation point flow*, Proc. Roy. Soc. London A, 384 (1982), pp. 175–190.

[7] G. J. Merchant and S. H. Davis, *Modulated stagnation point flow and steady streaming*, J. Fluid Mech., 198 (1989), pp. 543–555.

[8] Y. Matunobu, *Structure of pulsatile Hiemenz flow and temporal variations of wall shear stress near the stagnation point* II, J. Phys. Soc. Japan, 43 (1977), pp. 326–329.

[9] T. J. Pedley, *Two-dimensional boundary layers in a free stream which oscillates without reversing*, J. Fluid Mech., 55 (1972), pp. 359–383.
[10] M. Ishigaki, *Periodic boundary layer near a two-dimensional stagnation point*, J. Fluid Mech., 43 (1970), pp. 477–486.
[11] M. J. Lighthill, *The response of laminar skin friction and heat transfer to fluctuations in stream velocity*, Proc. Roy. Soc. London A, 224 (1954), pp. 1–23.
[12] N. Riley, *Oscillating viscous flows*, Mathematika, 12 (1965), pp. 161–175.
[13] P. Hall and D. T. Papageorgiou, *The onset of chaos in a class of Navier–Stokes solutions*, J. Fluid Mech., 393 (1999), pp. 59–87.
[14] K. Stewartson, C. J. Simpson, and R. J. Bodonyi, *The unsteady boundary layer on a rotating disk in a counter-rotating fluid. Part* 2, J. Fluid Mech., 121 (1982), pp. 507–515.
[15] N. Riley and R. Vasantha, *An unsteady stagnation point flow*, Quart. J. Mech. Appl. Math., 42 (1988), pp. 511–521.
[16] W. H. H. Banks and M. B. Zaturska, *The collision of unsteady laminar boundary layers*, J. Engrg. Math., 13 (1979), pp. 193–212.
[17] K. Stewartson and R. J. Bodonyi, *The unsteady laminar boundary layer on a rotating disk in a counter-rotating fluid.*, J. Fluid Mech., 79 (1977), pp. 669–688.
[18] H. Ockendon, *An asymptotic solution for steady flow above an infinite rotating disc with suction*, Quart. J. Mech. Appl. Math., 25 (1972), pp. 291–301.
[19] N. Riley and P. D. Weidman, *Multiple solutions of the Falkner–Skan equation for flow past a stretching boundary*, SIAM J. Appl. Math., 49 (1989), pp. 1350–1358.
[20] N. Riley, *Unsteady flow at a stagnation point*, J. Fluid Mech., 256 (1993), pp. 487–498.

SIAM J. APPL. MATH.
Vol. 63, No. 5, pp. 1615–1626

SYSTEM OF PHASE OSCILLATORS WITH DIAGONALIZABLE INTERACTION*

TAKASHI NISHIKAWA† AND FRANK C. HOPPENSTEADT†

Abstract. We consider a system of N phase oscillators having randomly distributed natural frequencies and diagonalizable interactions among the oscillators. We show that, in the limit of $N \to \infty$, all solutions of such a system are incoherent with probability one for any strength of coupling, which implies that there is no sharp transition from incoherence to coherence as the coupling strength is increased, in striking contrast to Kuramoto's (special) oscillator system.

Key words. network of phase oscillators, Kuramoto model

AMS subject classifications. 34C15, 37N25, 37N20

DOI. 10.1137/S0036139902411132

1. Introduction.

Synchronization of coupled oscillators is a ubiquitous phenomenon in natural and artificial systems. Examples include synchronization of pacemaker cells of the heart [11, 12], rhythmic activities in the brain [4, 13], synchronous flashing of fireflies [1, 2], arrays of lasers [8, 9], and superconducting Josephson junctions [18, 19]. Characterization of the phenomenon using mathematical models has been a topic of great interest for researchers in various scientific and engineering disciplines.

Wiener [16, 17], who recognized the ubiquity of synchronization phenomena in the real world, made a first attempt at characterization using the Fourier integrals. A more successful approach was taken by Winfree [20], who used a population of interacting limit-cycle oscillators to describe synchronization properties. He realized that if the interactions among the oscillators are weak and the oscillators are nearly identical, the separation of fast and slow timescales leads to a reduced model that can be expressed in terms solely of the phase of each oscillator. Kuramoto [10] put this idea on a firmer foundation by employing a perturbation method to show that the reduced equation has a universal form. His analysis of this model in the case of mean-field coupling kicked off an avalanche of theoretical investigations of his model and its generalizations.

More generally and rigorously, if each oscillator has an exponentially stable limit-cycle and interactions among them are weak, the reduced phase equation can be shown (see [6, Theorem 9.1, p. 253]) to have the form

$$\dot{\theta} = \omega + \varepsilon f(\theta), \qquad \theta \in \mathbb{T}^N, \tag{1.1}$$

where $\omega \in \mathbb{R}^N$ is the vector of natural frequencies of the oscillators that are coupled to one another through the interaction function $f : \mathbb{T}^N \to \mathbb{R}^N$, and $\varepsilon > 0$ represents the overall strength of the coupling. The universal form of the interaction function, derived by Kuramoto [10] under the additional assumption that the oscillators are

*Received by the editors July 10, 2002; accepted for publication (in revised form) January 8, 2003; published electronically June 12, 2003.
http://www.siam.org/journals/siap/63-5/41113.html

†Department of Mathematics, Arizona State University, Tempe, AZ 85287-1804 (tnishi@chaos6.la.asu.edu, fchoppen@asu.edu). The research of the first author was supported by DARPA/ONR grant N00014-01-1-0943. The research of the second author was partially supported by NSF grant DMS-0109001.

almost identical, corresponds to the choice

$$f_i(\theta) = \sum_{j=1}^{N} h_{ij}(\theta_j - \theta_i), \qquad i = 1, \dots, N, \tag{1.2}$$

where $f(\theta) = (f_1(\theta), \dots, f_N(\theta))^T$. The mean-field model that he studied results when $h_{ij}(x) = \sin(x)/N$ for all i, j.

Let $\theta(t)$ be a solution of (1.1). The oscillators i and j are said to be *locked* if $\lim_{t\to\infty} \theta_i(t)/\theta_j(t) = 1$. The solution is said to be *coherent* if all pairs of oscillators are locked. If none of the oscillator pairs are locked, the solution is *incoherent*. A solution that is neither coherent nor incoherent is called *partially coherent*. The main conclusion of Kuramoto's work [10] on his mean-field model is that in the limit of $N \to \infty$ there exists a critical coupling strength ε_c such that for $\varepsilon < \varepsilon_c$ the solution is incoherent, but for $\varepsilon > \varepsilon_c$ partially coherent solutions appear, for which the fraction of locked oscillator pairs is nonzero. Although his result was important, since this behavior closely resembles the phase transition phenomena widely observed in statistical physics, his analysis is heuristic and makes assumptions about the symmetry of the distribution of natural frequencies, which might not be necessary for the results [14].

In this paper, we consider a class of *diagonalizable* interaction functions, in which separation of variables is possible after an appropriate coordinate transformation. This allows us to prove rigorously that for the system (1.1) with a generic diagonalizable interaction function, if the solution is partially coherent, then it is almost surely coherent. This, together with the fact that the probability of having a coherent solution goes to zero in the limit of $N \to \infty$, leads to our main conclusion. Namely, for any $\varepsilon > 0$, the solution is almost surely incoherent in the limit of $N \to \infty$. Our result shows that a diagonalizable system of phase oscillators cannot exhibit a sudden transition from incoherence to coherence, in sharp contrast to the mean-field model of Kuramoto. This implies that for the system (1.1) to exhibit a phase transition, the interaction function f cannot be diagonalized.

There is an alternative rigorous approach to Kuramoto's mean-field model, in which the partial differential equation for the density of oscillators with certain frequency, which is obtained by taking the continuum limit $N \to \infty$, is studied to analyze the stability of the solutions. See [15] for an excellent review in this direction.

The approach taken here is similar to that in [5, p. 80]. However, some conclusions made there might be misleading or lack detailed analysis. This paper is intended to correct and clarify those points.

The rest of the paper is organized as follows. In section 2, we introduce an appropriate change of variables to separate a time-like variable from the rest of the system. In section 3, we define diagonalizable interaction and show how complete separation of variables can be achieved. We also establish some properties of diagonalizable systems. Then, in section 4, we introduce randomness of the natural frequencies of the oscillators and state our main results. Finally, we discuss some approximate behavior of the system for large N in section 6, and section 7 is reserved for concluding remarks.

2. Separation of the time-like variable. In this and the following sections, we consider the system (1.1) of N phase oscillators, where the natural frequency vector ω and the coupling strength ε are fixed (nonrandom) constants. We will consider ω to be a random vector in section 4 in order to make probabilistic statements about the system.

Let us suppose that the interaction function f satisfies two conditions,

(C1) $\mathbf{1}^T f(\theta) = \mathbf{0}$ for all $\theta \in \mathbb{T}^N$ and
(C2) $f(v\mathbf{1} + \theta) = f(\theta)$ for all $\theta \in \mathbb{T}^N$,
where $\mathbf{1} = (1/\sqrt{N}, \dots, 1/\sqrt{N})^T$. The condition (C1) says that the interaction function is orthogonal to the vector $\mathbf{1}$. The second condition (C2) expresses the translation invariance of f along the direction of $\mathbf{1}$. If, for example, the interaction function has the form (1.2), these conditions are satisfied if the functions h_{ij} are odd. In particular, the mean-field model of Kuramoto does satisfy these conditions.

Under conditions (C1) and (C2), the system (1.1) can be separated into two independent systems—one for the time-like variable and the other for the phase deviations.

Let W be an $N \times (N-1)$ matrix whose columns, denoted by W_j, $j = 1, \dots, N-1$, form an orthonormal basis of the subspace $\mathbf{1}^\perp \equiv \{x \in \mathbb{R}^N : \mathbf{1}^T x = 0\}$. In other words, W is an $N \times (N-1)$ matrix that satisfies $\mathbf{1}^T W = \mathbf{0}$ and $W^T W = I_{N-1}$, where I_{N-1} is the $(N-1) \times (N-1)$ identity matrix. Then, the change of variable

$$\theta = v\mathbf{1} + Wu \tag{2.1}$$

converts the system (1.1) into two systems,

$$\dot{v} = \mathbf{1}^T \omega, \tag{2.2}$$

$$\dot{u} = W^T \omega + \varepsilon W^T f(Wu), \tag{2.3}$$

which can be solved separately.

Systems satisfying the conditions (C1) and (C2) arise in mathematical neuroscience [5, 6], in which θ often takes the form $\omega t + \phi$ in the limit $t \to \infty$, where ω is the vector of carrier frequencies and ϕ is the vector of phase deviations. The equation (2.3), in some sense, governs the behavior of the phase deviations.

The solution to (2.2) is $v(t) = v(0) + (\mathbf{1}^T \omega)t$, and hence the variable v is time-like if $\mathbf{1}^T \omega \neq 0$ or, equivalently, if the average natural frequency $\sum_i \omega_i / N$ is nonzero. Thus, the behavior of the solution of (1.1) is essentially determined by (2.3).

Recall that the solution is called coherent if $\lim_{t\to\infty} \theta_i(t)/\theta_j(t) = 1$. This can be rephrased in terms of the vector $\mu \equiv \lim_{t\to\infty} u(t)/t$ of output frequencies of the u-equation (2.3), if it exists.

LEMMA 2.1. *Let $u(t)$ be a solution of (2.3), and suppose that $\mu \equiv \lim_{t\to\infty} u(t)/t$ exists. Then, the solution of (1.1) is coherent if and only if $\mu = 0$.*

Proof. Let $\Omega = \lim_{t\to\infty} \theta(t)/t = (\mathbf{1}^T \omega)\mathbf{1} + W\mu$. Then $\mu = 0$ implies that $\Omega = (\mathbf{1}^T \omega)\mathbf{1}$, which in turn implies that $\lim_{t\to\infty} \theta_i(t)/\theta_j(t) = 1$.

Conversely, if $\lim_{t\to\infty} \theta_i(t)/\theta_j(t) = 1$, Ω must be a multiple of $\mathbf{1}$. Since W is orthogonal to $\mathbf{1}$, this implies that $W\mu = 0$. Since W is invertible, it follows that $\mu = 0$. ∎

As an immediate consequence of Lemma 2.1, if the solution of (2.3) tends to an equilibrium, then the corresponding solution of (1.1) is coherent. For example, if the interaction function f that satisfies (C1) and (C2) is a gradient vector field, i.e., $f(\theta) = -\nabla V_0(\theta)$ for some potential function $V_0 : \mathbb{T}^N \to \mathbb{R}$, then the u-equation (2.3) is also a gradient system:

$$\dot{u} = -\nabla \left[-\omega^T W u + \varepsilon V(u) \right], \tag{2.4}$$

where $V(u) = V_0(Wu)$. A minimum u^* of the potential function $-\omega^T W u + \varepsilon V(u)$ then corresponds to the vector of phase deviations for a coherent solution of the original oscillator system (1.1). As in the proof of Lemma 2.1, we have $\Omega = (\mathbf{1}^T \omega)\mathbf{1}$ in this case, meaning that the output frequency of every oscillator tends to the mean natural

frequency $\bar{\omega} \equiv \sum_i \omega_i/N$ of the oscillators. For the interaction of the form (1.2), the potential function takes the form

$$V(u) = -\frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{N} H_{ij}(\theta_i - \theta_j) = -\frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{N} H_{ij}\left(\sum_{k=1}^{N-1}[W_{ik} - W_{jk}]u_k\right),$$

where $H_{ij}(x) \equiv \int_0^x h_{ij}(y)dy$.

3. Diagonalizable interaction. Assuming that the interactions among the oscillators are diagonalizable enables us to carry out a rigorous analysis of the system.

DEFINITION 3.1. *We say that the system* (1.1) *(or the interaction function f) is* diagonalizable *if there exist an $N \times (N-1)$ matrix W and real, continuous, periodic functions p_j such that*

(i) $\mathbf{1}^T W = \mathbf{0}$,
(ii) $W^T W = I_{N-1}$, *and*
(iii) $f(Wu) = Wp(u)$ *with* $p(u) = (p_1(u_1), \dots, p_{N-1}(u_{N-1}))^T$.

For example, $W = W^{(N)}$ defined by

$$\begin{aligned} W_{jk}^{(N)} &= \frac{1}{\sqrt{N}}\left(\sin\frac{2\pi jk}{N} + \cos\frac{2\pi jk}{N}\right) \\ &= \frac{2}{\sqrt{N}}\sin\left(\frac{2\pi jk}{N} + \frac{\pi}{4}\right) \end{aligned} \tag{3.1}$$

satisfies these conditions.

When the system (1.1) is diagonalizable, the equations for the components of u become independent of other components:

$$\dot{u_j} = a_j + \varepsilon p_j(u_j), \qquad j = 1, \dots, N-1, \tag{3.2}$$

where we set $a_j = W_j^T \omega$. Thus, the problem is reduced to solving a scalar differential equation for each j. The following lemma applies to each equation in (3.2).

LEMMA 3.2. *Let $\varepsilon > 0$, and let a be a real number. Let $p(u)$ be a real, continuous, periodic function with period $L > 0$. Define $m = \min_{0\le u<L} p(u)$, $M = \max_{0\le u<L} p(u)$. For any solution $u(t)$ of $\dot{u} = a + \varepsilon p(u)$, the limit $\mu_p(a,\varepsilon) \equiv \lim_{t\to\infty} u(t)/t$ exists and*

$$\mu_p(a,\varepsilon) = \begin{cases} L/T(a,\varepsilon), & a < -\varepsilon M,\ a > -\varepsilon m, \\ 0, & -\varepsilon M \le a \le -\varepsilon m, \end{cases} \tag{3.3}$$

where

$$T(a,\varepsilon) \equiv \int_0^L \frac{du}{a + \varepsilon p(u)}$$

is the "period" of the solution in the case of $a < -\varepsilon M$ or $a > -\varepsilon m$, in the sense that $u(t + T(a,\varepsilon)) = u(t) + L$.

Proof. If $-\varepsilon M \le a \le -\varepsilon m$, then any solution $u(t)$ tends to a zero of the function $a + \varepsilon p(u)$. Hence, $\mu_p(a,\varepsilon) = 0$.

For notational simplicity, let us drop the dependence of $T(a,\varepsilon)$ on a and ε below. Suppose $a > -\varepsilon m$, so that $a + \varepsilon p(u) > 0$ for all u. It is straightforward to show that the function $u(t)$ defined implicitly by the formula

$$\int_{u_0}^{u(t)} \frac{du}{a + \varepsilon p(u)} = t$$

is the unique solution of $\dot{u} = a + \varepsilon p(u)$ with the initial condition $u(0) = u_0$, and that it satisfies $u(t+T) = u(t) + L$. We have

$$\begin{aligned}\mu_p(a,\varepsilon) &= \lim_{t\to\infty} \frac{u(t)}{t} \\ &= \lim_{t\to\infty} \frac{1}{t}\left(u_0 + \int_0^t [a + \varepsilon p(u(s))]ds\right) \\ &= a + \varepsilon \lim_{t\to\infty} \frac{1}{t}\int_0^t p(u(s))ds.\end{aligned}$$

Let n be the largest integer for which $nT \le t$. Then, by changing the variables in each integral using the translation by multiples of T, we see that

$$\begin{aligned}\int_0^t p(u(s))ds &= \sum_{k=1}^{n} \int_{(k-1)T}^{kT} p(u(s))ds + \int_{nT}^{t} p(u(s))ds \\ &= n\int_0^T p(u(s))ds + \int_0^{t-nT} p(u(s))ds.\end{aligned}$$

Consequently,

$$\begin{aligned}\left|\frac{1}{t}\int_0^t p(u(s))ds - \frac{1}{T}\int_0^T p(u(s))ds\right| &= \left|\left(\frac{n}{t} - \frac{1}{T}\right)\int_0^T p(u(s))ds + \frac{1}{t}\int_0^{t-nT} p(u(s))ds\right| \\ &\le \frac{|nT-t|}{tT}\int_0^T |p(u(s))|ds + \frac{1}{t}\int_0^{t-nT} |p(u(s))|ds \\ &\le \frac{T}{tT}\, T\max\{|m|,|M|\} + \frac{1}{t}\, T\max\{|m|,|M|\} \\ &\to 0\end{aligned}$$

as $t \to \infty$, showing that the limit $\lim_{t\to\infty} \frac{1}{t}\int_0^t p(u(s))ds$ exists and is equal to $\frac{1}{T}\int_0^T p(u(s))ds$. Thus, by changing variables from s to u and translating by u_0, we see that $\mu_p(a,\varepsilon)$ exists and

$$\begin{aligned}\mu_p(a,\varepsilon) &= a + \frac{\varepsilon}{T}\int_0^T p(u(s))ds \\ &= a + \frac{\varepsilon}{T}\int_0^L \frac{p(u)du}{a + \varepsilon p(u)} \\ &= \frac{L}{T}.\end{aligned}$$

If $a < -\varepsilon M$, then, by replacing u with $-u$, a with $-a$, ε with $-\varepsilon$, and m with M, the problem reduces to the previous case. The lemma is proved. □

The function $\mu_p(\cdot,\varepsilon)$ in Lemma 3.2, which can easily be shown to be differentiable with positive derivative outside the interval $[-\varepsilon M, \varepsilon M]$, determines the relationship between the input frequency a and the output frequency $\mu_p(a,\varepsilon)$. If we take $p(u) = \sin(u)$, for example, the integration in the expression of μ_p can be carried out, and we get

$$\mu_p(a,\varepsilon) = \begin{cases} -\sqrt{a^2-\varepsilon^2}, & a < -\varepsilon, \\ 0, & -\varepsilon \le a < \varepsilon, \\ \sqrt{a^2-\varepsilon^2}, & a \ge \varepsilon, \end{cases}$$

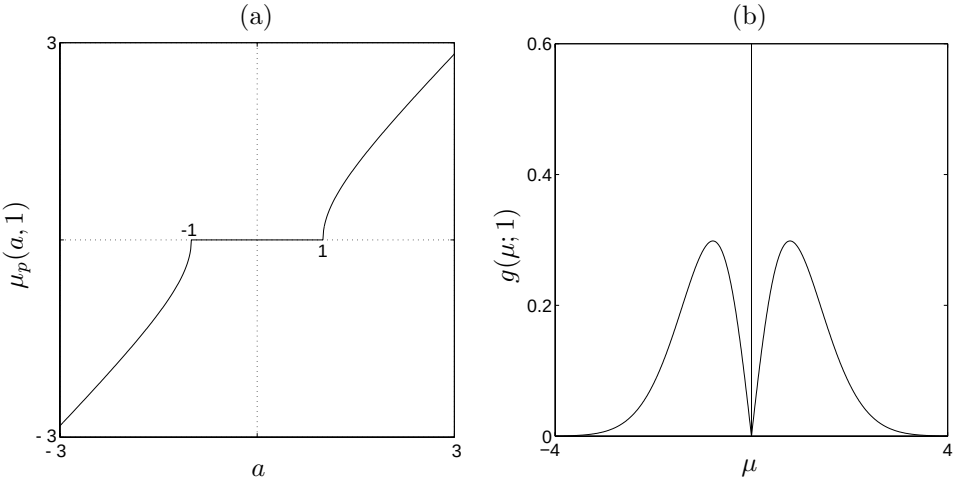


FIG. 3.1. (a) *The graph of the input-output frequency function* $\mu_p(a,\varepsilon)$ *versus* a *for* $\varepsilon = 1$ *and* $p(u) = \sin(u)$. (b) *The corresponding density* $g(\mu;1)$ *when* a *is the standard Gaussian random variable.*

the graph of which is given in Figure 3.1(a) for $\varepsilon = 1$. With this function, the output frequency vector μ of the u-equation (2.3) can be written as

$$\mu = \mu(a) = (\mu_{p_1}(a_1,\varepsilon), \dots, \mu_{p_{N-1}}(a_{N-1},\varepsilon))^T.$$

It is important to note here that μ does *not* depend on the initial condition $u(0) = u_0$, which implies that it is also independent of the initial condition for θ. In other words, the initial condition for the system (1.1) does not affect the behavior of its solution, as far as its coherence properties are concerned. Therefore, in this sense, coherence, partial coherence, and incoherence are properties of the system rather than of individual solutions for a diagonalizable system.

4. Randomly distributed frequencies. In this section we consider ω to be a random vector in $\mathbb{R}^N$. We take the components $\omega_1, \dots, \omega_N$ of ω to be independent and identically distributed (i.i.d.) random variables with mean 0 and variance $\sigma^2 > 0$. In the general case of mean $\omega_0 \neq 0$, the problem can always be reduced to the zero-mean case by the translation of θ by $-\omega_0 t$.

Since ω is random, the vectors a and μ are also random vectors in $\mathbb{R}^{N-1}$. Lemma 3.2 along with the relation $a = W^T\omega$ can be used to determine the distribution of μ from the distribution of ω. For example, if each ω_j is standard Gaussian, then so is each a_j, in which case the density $g(\mu;\varepsilon)$ for the random variable $\mu_p(a_j,1)$ when $p(u) = \sin(u)$ can be computed. The result is

$$g(\mu;\varepsilon) = \frac{|\mu| e^{-(\mu^2+\varepsilon^2)/2}}{\sqrt{2\pi(\mu^2+\varepsilon^2)}} + \delta(\mu)\,\mathrm{erf}\left(\frac{\varepsilon}{\sqrt{2}}\right),$$

where $\delta(\mu)$ is Dirac's delta function. The graph of this density is shown in Figure 3.1 for $\varepsilon = 1$.

Our main goal is this section is to compute the probabilities that the system (1.1) is coherent, partially coherent, or incoherent. The following theorem reveals a curious property of a generic diagonalizable system of phase oscillators.

THEOREM 4.1. *Let the natural frequency vector ω be a random vector in $\mathbb{R}^N$, whose components are i.i.d. with a common continuous distribution. Suppose that* (1.1) *is a diagonalizable system of N phase oscillators such that W satisfies the condition that $W_{ki} \neq W_{kj}$ for all $k = 1, 2, \dots, N$ and for all $i, j = 1, 2, \dots, N-1$ such that $i \neq j$. Then, the partial coherence of the system almost surely implies coherence; i.e., given that the system is partially coherent, the probability that it is coherent is one.*

Proof. Once again, let $\Omega(\omega) = \lim_{t\to\infty} \theta(t)/t = (\mathbf{1}^T\omega)\mathbf{1} + W\mu(\omega)$. Let S_c be the set of ω in $\mathbb{R}^N$ that corresponds to coherent systems, i.e., $S_c = \{\omega \in R^N : \Omega(\omega) = \mathbf{0}\}$. By Lemma 2.1, we may also write $S_c = \{\omega \in R^N : \mu(\omega) = \mathbf{0}\}$. Let S_{pc} be the set corresponding to partially coherent systems, that is, $S_{pc} = \{\omega \in R^N : \Omega_i(\omega) = \Omega_j(\omega)$ for some $i \neq j\}$. It is easy to see that we can also rewrite this in terms of μ as

$$\begin{aligned} S_{pc} &= \left\{ \omega \in R^N : \text{ There are } i \neq j \text{ s.t. } \sum_{k=1}^{N-1} (W_{ki} - W_{kj})\mu_k(\omega) = 0 \right\} \\ &= \bigcup_{i\neq j} \left\{ \omega \in R^N : \sum_{k=1}^{N-1} (W_{ki} - W_{kj})\mu_k(\omega) = 0 \right\} \equiv \bigcup_{i\neq j} S_{pc}^{(i,j)}. \end{aligned}$$

The probability that the system is coherent, given that the system is partially coherent, is $P(S_c)/P(S_{pc})$ since $S_c \subset S_{pc}$. This probability is one if and only if $P(S_{pc} \setminus S_c) = 0$, which would be satisfied if $P(S_{pc}^{(i,j)} \setminus S_c) = 0$ for every pair $i \neq j$. We shall show this next.

Let us fix i and j. For any $A \subset \{1, 2, \dots, N-1\}$, denote by Z_k the subspace $\{\mu \in \mathbb{R}^{N-1} : \mu_k = 0\}$, and let $Z_A = \bigcup_{k\in A} Z_k$ and $Z'_A = \bigcap_{k\notin A} Z_k$. Let R_A denote the subspace $\{\mu \in \mathbb{R}^{N-1} : \sum_{k\in A}(W_{ki} - W_{kj})\mu_k = 0\}$. Define $Q_A = R_A \cap Z'_A \setminus \{\mathbf{0}\}$. We will show that $P(Q_A) = 0$ for any choice of A. $P(S_{pc}^{(i,j)} \setminus S_c) = 0$ follows from this by taking $A = \{1, 2, \dots, N-1\}$.

We shall prove $P(Q_A) = 0$ by induction on $n = |A|$, the cardinality of A. Suppose first that $n = 1$ and, say, $A = \{1\}$. Since $W_{1i} - W_{1j} \neq 0$, we have $R_A = \{\mu \in \mathbb{R}^{N-1} : \mu_1 = 0\} = Z_1$ and $Z'_A = \bigcap_{k=2}^{N-1} Z_k$. Thus, $Q_A = \bigcap_{k=1}^{N-1} Z_k \setminus \{\mathbf{0}\} = \varnothing$, which implies $P(Q_A) = 0$. The same holds for any other A with $|A| = 1$.

Suppose that $P(Q_A) = 0$ for any A with $|A| = n-1$, and consider the case $|A| = n$. We have

$$\begin{aligned} P(Q_A) &= P(Q_A \cap Z_A) + P(Q_A \setminus Z_A) \\ &= P\left(\bigcup_{k\in A} Q_A \cap Z_k \right) + P(Q_A \setminus Z_A) \\ &\leq \sum_{k\in A} P(Q_A \cap Z_k) + P(Q_A \setminus Z_A). \end{aligned}$$

We see that $P(Q_A \cap Z_k) = 0$ for each $k \in A$, by the induction hypothesis, since we can write $Q_A \cap Z_k = R_A \cap Z'_A \cap Z_k \setminus \{\mathbf{0}\} = R_{A_k} \cap Z'_{A_k} \setminus \{\mathbf{0}\} = Q_{A_k}$ with $A_k = A \setminus \{k\}$, for which we have $|A_k| = n - 1$. Thus, if we can show $P(Q_A \setminus Z_A) = 0$, then we are done.

We show $P(Q_A \setminus Z_A) = 0$ in three steps. First, since Z'_A is an n-dimensional subspace and $Q_A \subset R_A \cap Z'_A$ is an $(n-1)$-dimensional subspace, the n-dimensional Lebesgue measure of Q_A in Z'_A must be zero.

Next, note that the conditional probability distribution of μ, given $\mu \in Z'_A$, is continuous with respect to the Lebesgue measure outside the set Z_A. This can be seen by noting the following: (1) each component can be written as $\mu_j = \mu_{p_j}(a_j)$ by Lemma 3.2, (2) $\mu_{p_j}^{-1}$ exists and is differentiable except at the origin, again by Lemma 3.2, and (3) the conditional distribution of $a_j = \sum_k W_{kj}\omega_k$, given that $\mu(\omega) \in Z'_A$ (which is equivalent to $-\varepsilon M_k \le \sum_l W_{lk}\omega_l \le -\varepsilon m_k$ for all $k \notin A$), is continuous everywhere.

Finally, combining these two observations, we see that $P(Q_A \setminus Z_A \mid Z'_A) = 0$, which implies that $P(Q_A \setminus Z_A) = P(Z'_A)P(Q_A \setminus Z_A \mid Z'_A) = 0$. This completes the proof of the theorem. □

We next describe the behavior of a generic diagonalizable system in the limit of $N \to \infty$. In order to formalize the process of taking the limit, we need to choose a sequence of systems of the form (1.1). Such a sequence can be characterized by the following:

1. Consider a sequence $\{W^{(N)}\}_{N=1,2,\dots}$ of matrices with the following properties:
   (a) Each $W^{(N)}$ is an $N \times (N-1)$ matrix with orthonormal columns.
   (b) $\mathbf{1}_N^T W^{(N)} = \mathbf{0}$ for all N.
   (c) Each $W^{(N)}$ satisfies the condition for W in Theorem 4.1.
   (d) $||W^{(N)}||_\infty \to 0$ as $N \to \infty$. (Here $||\cdot||_\infty$ denotes the maximum matrix norm defined by $||A||_\infty = \max|A_{ij}|$, where the maximum is taken over all elements of A.) This is like a mixing condition that will be necessary later in order to apply Proposition 4.3.
2. Consider a sequence $\{p_j\}_{j=1,2,\dots}$ of real, continuous, periodic functions such that the corresponding sequence of norms $||p_j|| \equiv \max_j |p_j(u)|$ is bounded.
3. Consider a sequence $\{\omega_j\}_{j=1,2,\dots}$ of i.i.d. random variables with mean ω_0 and variance σ^2.

The sequence of matrices $W^{(N)}$ defined by (3.1) satisfies the conditions above. Given such sequences, for each $\varepsilon > 0$ and N, we define $\mathcal{S}_{N,\varepsilon}$ to be the diagonalizable system (1.1) of phase oscillators using the natural frequency vector $\omega = (\omega_1, \dots, \omega_N)^T$, the functions $\{p_1, \dots, p_{N-1}\}$, and the matrix $W^{(N)}$. We are now ready to state and prove our main theorem.

THEOREM 4.2. *Let $\mathcal{S}_{N,\varepsilon}$ be defined as above. Then, for any fixed $\varepsilon > 0$, $\mathcal{S}_{N,\varepsilon}$ is almost surely incoherent as $N \to \infty$; i.e., the probability that $\mathcal{S}_{N,\varepsilon}$ is incoherent tends to one in the limit of $N \to \infty$.*

Proof. As mentioned before, we may assume $\omega_0 = 0$ without loss of generality, since the $\omega_0 \neq 0$ case can always be reduced to the $\omega_0 = 0$ case.

From Theorem 4.1, we know that the probability that $\mathcal{S}_{N,\varepsilon}$ is *not* incoherent is equal to the probability q_c that it is coherent. We need to show that $q_c \to 0$ as $N \to \infty$.

Let $N_0 < N$ be fixed. From Lemma 2.1, it follows that $q_c = P(\mu = \mathbf{0})$. Since the sequence $\{||p_j||\}_{j=1,2,\dots}$ is bounded, we can define $M = \sup M_j$ and $m = \inf m_j$, where $m_j = \min_{0\le u<L} p_j(u)$, $M_j = \max_{0\le u<L} p_j(u)$ for each j. Then, Lemma 3.2 implies

$$\begin{aligned} q_c &= P(\mu = \mathbf{0}) \\ &= P(-\varepsilon M_j \le a_j^{(N)} \le -\varepsilon m_j,\ j = 1, \dots, N-1) \\ &\le P(-\varepsilon M \le a_j^{(N)} \le -\varepsilon m,\ j = 1, \dots, N_0), \end{aligned}$$

where $a_j^{(N)} = (W_j^{(N)})^T \omega$.

The following Proposition shows that for each j, $a_j^{(N)}$ converges to a Gaussian random variable with mean 0 and variance σ^2.

PROPOSITION 4.3. *Let $X_1, X_2, \dots$ be a sequence of i.i.d. random variables with $EX_j = 0$ and $Var(X_j) = E(X_j^2) = \sigma^2$. Suppose that, for each N, real numbers $b_{N,1}, \dots, b_{N,N}$ satisfy $\sum_{j=1}^N b_{N,j}^2 = 1$. Also, suppose that*

$$\lim_{N\to\infty} \max_{1\le j\le N} |b_{N,j}| = 0.$$

Then we have

$$S_N = \sum_{j=1}^N b_{N,j} X_j \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

as $N \to \infty$.

Proof. Let $Y_{N,j} = b_{N,j} X_j$. We will apply the Lindeberg–Feller central limit theorem (see, for example, [3, p. 98]) to $Y_{N,j}$. For this we need to check three conditions. The first is $EY_{N,j} = b_{N,j} EX_j = 0$. The second condition is satisfied because $\sum_{j=1}^N EY_{N,j}^2 = \sum_{j=1}^N b_{N,j}^2 EX_j^2 = \sigma^2 > 0$. To show that the third condition is satisfied, let $\varepsilon > 0$ be fixed. We have

$$\sum_{j=1}^N E\Big(|Y_{N,j}|^2 \,\Big|\, |Y_{N,j}| > \varepsilon\Big) = \sum_{j=1}^N b_{N,j}^2 E\left(|X_j|^2 \,\middle|\, |X_j| > \frac{\varepsilon}{|b_{N,j}|}\right),$$

where $E(X|A)$ denotes the conditional expectation of X, given A. Let j be fixed. For each N, set $Z_N = |X_j|^2$ if $|X_j| > \varepsilon/|b_{N,j}|$, and 0 otherwise. Since $|b_{N,j}| \to 0$, $Z_N \le |X_j|^2$ for each N, and $Z_N \to 0$ almost surely, we may use the dominated convergence theorem to show that for each $j = 1, 2, \dots$, $EZ_N = E\big(|X_j|^2 \mid |X_j| > \varepsilon/|b_{N,j}|\big) \to 0$ as $N \to \infty$. Thus, the third condition $\sum_{j=1}^N E\big(|Y_{N,j}|^2 \mid |Y_{N,j}| > \varepsilon\big) \to 0$ is satisfied. The conclusion now follows directly from application of the Lindeberg–Feller theorem. ∎

For each $j = 1, \dots, N-1$, we take $b_{N,i} = W_{ij}^{(N)}$ and $X_i = \omega_i$ in Proposition 4.3, and we see that $a_j^{(N)}$ converges in distribution to $a_j^{(\infty)}$ as $N \to \infty$, where $a_j^{(\infty)}$ is a Gaussian random variable. Moreover, due to the orthogonality of $W^{(N)}$, $a_1^{(N)}, \dots, a_{N-1}^{(N)}$, in some sense, become independent in the limit.

LEMMA 4.4. *The random variables $a_1^{(\infty)}, a_2^{(\infty)}, \dots$ are independent.*

Proof. We need to show that for any finite $A \subset \mathbb{N}$, the collection $\{a_k^{(\infty)}\}_{k\in A}$ is a set of independent random variables. For simplicity, we prove this only for $A = \{1, 2\}$, but a similar argument works for a general case.

Let t_1 and t_2 be given. Set

$$b_{N,j} = \frac{t_1 W_{1j}^{(N)} + t_2 W_{2j}^{(N)}}{\sqrt{t_1^2 + t_2^2}}.$$

Then, as $N \to \infty$, $\max_j |b_{N,j}|$ approaches zero because $\max_j |W_{1j}|$ and $\max_j |W_{2j}|$ go to zero. Also, it is easy to check that $\sum_{j=1}^N b_{N,j}^2 = 1$ for all N. Applying Proposition 4.3, we see that

$$\frac{t_1 a_1^{(N)} + t_2 a_2^{(N)}}{\sqrt{t_1^2 + t_2^2}} = \sum_{j=1}^N b_{N,j}\omega_j \xrightarrow{d} \mathcal{N}(0, \sigma^2),$$

which implies that $t_1 a_1^{(N)} + t_2 a_2^{(N)} \xrightarrow{d} \mathcal{N}(0, \sigma^2(t_1^2+t_2^2))$, which in turn implies the convergence of the joint characteristic function of $a_1^{(N)}$ and $a_2^{(N)}$ as $N \to \infty$. Specifically,

$$Ee^{it_1 a_1^{(\infty)} + it_2 a_2^{(\infty)}} = \lim_{N\to\infty} Ee^{i(t_1 a_1^{(N)} + t_2 a_2^{(N)})} = e^{-\sigma^2(t_1^2+t_2^2)/2} = e^{-\sigma^2 t_1^2/2}\, e^{-\sigma^2 t_2^2/2}.$$

Therefore, $a_1^{(\infty)}$ and $a_2^{(\infty)}$ are independent. □

Let us come back to the proof of Theorem 4.2. As a consequence of $a_1^{(\infty)}, a_2^{(\infty)}, \ldots$ being independent Gaussian random variables, we have

$$\begin{aligned} q_c &\le P(\{-\varepsilon M \le a_j^{(N)} \le -\varepsilon m,\ j = 1, \ldots, N_0\}) \\ &\to \left[\frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\varepsilon M}^{-\varepsilon m} e^{-x^2/2\sigma^2} dx\right]^{N_0} \end{aligned}$$

as $N \to \infty$. Since this holds for any fixed N_0, and since the right-hand side goes to zero as $N_0 \to \infty$, we conclude that $q_c \to 0$. This completes the proof of Theorem 4.2. □

5. Example. A network of voltage-controlled oscillator (VCO) devices can be built as an example of systems with diagonalizable interaction. The behavior of the jth VCO in the network is described by its phase variable θ_j, which satisfies [5, 7]:

$$\dot{\theta_j} = \omega_j + I_j(t),$$

where ω_j is the center frequency and $I_j(t)$ is the input signal from other VCOs. The system is diagonalizable if, for example, $I_j(t)$ has the form

$$I_j(t) = \varepsilon \sum_{k=1}^{N-1} W_{jk}^{(N)} \sin\left(\sum_{\ell=1}^{N} W_{\ell k}^{(N)} \theta_\ell\right),$$

with $W^{(N)}$ defined by (3.1).

This type of interaction can be implemented using commercially available circuit elements, as follows. The sine terms on the right-hand side can be constructed as sums and products of output voltages:

$$\sin\left(\sum_{\ell=1}^{N} W_{\ell k}^{(N)} \theta_\ell\right) = \sum_b \prod_{\ell=1}^{N} \cos\left(W_{\ell k}^{(N)} \theta_\ell - \frac{b_\ell \pi}{2}\right),$$

where the sum is taken over all (ordered) binary N-tuples $b = (b_1, \ldots, b_N)$, $b_\ell = 0, 1$, such that $\sum_\ell b_\ell$ is odd. This means that $I_j(t)$ is a sum of terms that are products of $\sin(W_{\ell k}^{(N)} \theta_\ell)$ and $\cos(W_{\ell k}^{(N)} \theta_\ell)$. Signals of the form $\sin(W_{\ell k}^{(N)} \theta_\ell)$ can be obtained by using the amplified version of the input $W_{\ell k}^{(N)} I_\ell(t)$ as the controlling voltage in a separate VCO with center frequency $W_{\ell k}^{(N)} \omega_\ell$. From these we can get $\cos(W_{\ell k}^{(N)} \theta_\ell)$ by the phase shift of $\pi/2$. Finally, $I_j(t)$ is obtained by putting these signals through multipliers and adding the outputs.

6. Discussion. In order to gain additional insights, let us consider the case when the functions $p_j = p$ have the range $[-1, 1]$ and do not depend on j. The arguments in the proof of Theorem 4.2 suggest that for a diagonalizable system (1.1) with large N, the probability q_c that it is coherent is approximately

$$q_c \approx \left[\frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\varepsilon}^{\varepsilon} e^{-x^2/2\sigma^2} dx\right]^{N-1} = \left[\operatorname{erf}\left(\frac{\varepsilon}{\sigma\sqrt{2}}\right)\right]^{N-1} \equiv \tilde{q}_c(\varepsilon; N),$$

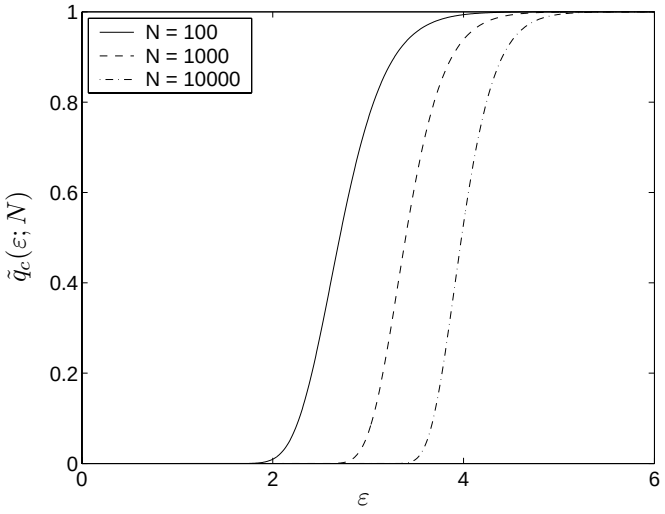


FIG. 6.1. *Probability of coherence* $\tilde{q}_c(\varepsilon; N)$ *as a function of* ε *for* $\sigma = 1$ *and* $N = 100, 1000, 10000$.

where $\mathrm{erf}(x)$ is the error function. Typical graphs of $\tilde{q}_c(\varepsilon; N)$ are plotted in Figure 6.1. One can see that for any finite value of N, there seems to be a sharp transition point through which $\tilde{q}_c(\varepsilon; N)$ changes from 0 to 1. However, unlike the mean-field model of Kuramoto, this point keeps shifting to the right as N increases, and tends to ∞ in the limit of $N \to \infty$, although it can be shown that this increase is at most $O(\sqrt{\ln N})$.

LEMMA 6.1. *Let* $\sigma > 0$ *and* $0 < q < 1$ *be fixed. Define* $\varepsilon_{q,\sigma}(N)$ *implicitly by* $q = \tilde{q}_c(\varepsilon_{q,\sigma}(N); N)$. *Then* $\varepsilon_{q,\sigma}(N) = O(\sqrt{\ln N})$ *as* $N \to \infty$*; i.e.,* $\varepsilon_{q,\sigma}(N)/\sqrt{\ln N}$ *is bounded as* $N \to \infty$.

Proof. Let $x = (\sigma\sqrt{2})^{-1}\varepsilon_{q,\sigma}(N)$. Then, using a known estimate for the error function, we have for $x \geq 1$

$$\mathrm{erf}(x) \geq 1 - \frac{2e^{-x^2}}{\sqrt{\pi}x + \sqrt{\pi x^2 + 4}} \geq 1 - \frac{2e^{-x^2}}{\sqrt{\pi} + \sqrt{\pi + 4}}.$$

Hence, we have the estimate

$$q \geq (1 - C_0 e^{-x^2})^{N-1} \geq 1 - (N-1)C_0 e^{-x^2},$$

where

$$C_0 = \frac{2}{\sqrt{\pi} + \sqrt{\pi + 4}}.$$

Here we used the relation $(1-x)^n \geq 1 - nx$, which is valid for $n \geq 0$ and $0 \leq x \leq 1$. The estimate for $\varepsilon_{q,\sigma}(N)$ can be obtained by rearranging:

$$\varepsilon_{q,\sigma}(N) \leq \sigma\sqrt{C_1 + 2\ln(N-1)},$$

where $C_1 = 2\ln C_0 - \ln(1-q)$. This implies $\varepsilon_{q,\sigma}(N) = O(\sqrt{\ln N})$. ∎

7. Conclusions. In this paper, we have defined a class of systems of phase oscillators characterized by having diagonalizable interactions. For a system in this class, complete separation of variables through appropriate changes of variable is possible, which enables us to draw rigorous conclusions about the probabilistic properties of the system. In particular, we have shown that partial coherence of the system almost surely implies coherence and, in the limit of large system size, the system is almost surely incoherent. A major implication of our result is that, unlike the mean-field model of Kuramoto, diagonalizable systems cannot exhibit a sharp transition from incoherence to coherence. This provides some insight into what is necessary to see such a transition in a system of phase oscillators.

REFERENCES

[1] J. Buck, *Synchronous fireflies*, Sci. Amer., 234 (1976), pp. 74–85.
[2] J. Buck, *Synchronous rhythmic flashing of fireflies.* 2, Quart. Rev. Biol., 63 (1988), pp. 265–289.
[3] R. Durrett, *Probability: Theory and Examples*, Brooks/Cole, Pacific Grove, CA, 1991.
[4] C. M. Gray, *Synchronous oscillations in neuronal systems: Mechanism and functions*, J. Comput. Neurosci., 1 (1994), pp. 11–38.
[5] F. C. Hoppensteadt, *An Introduction to the Mathematics of Neurons*, Cambridge University Press, Cambridge, UK, 1997.
[6] F. C. Hoppensteadt and E. M. Izhikevich, *Weakly Connected Neural Networks*, Springer-Verlag, New York, 1997.
[7] P. Horowitz and W. Hill, *The Art of Electronics*, 2nd ed., Cambridge University Press, Cambridge, UK, 1989.
[8] Z. Jiang and M. McCall, *Numerical-simulation of a large number of coupled lasers*, J. Opt. Soc. Amer. B Opt. Phys., 10 (1993), pp. 155–163.
[9] S. Y. Kourtchatov, V. V. Likhanskii, A. P. Napartovich, F. T. Arecchi, and A. Lapucci, *Theory of phase locking of globally coupled laser arrays*, Phys. Rev. A, 52 (1995), pp. 4089–4094.
[10] Y. Kuramoto, *Chemical Oscillations, Waves and Turbulence*, Springer-Verlag, Berlin, 1984.
[11] D. C. Michaels, E. P. Matyas, and J. Jalife, *Mechanisms of sinoatrial pacemaker synchronization—A new hypothesis*, Circulation Res., 61 (1987), pp. 704–714.
[12] C. S. Peskin, *Mathematical Aspects of Heart Physiology*, Courant Institute of Mathematical Science Publication, New York, 1975.
[13] W. Singer and C. M. Gray, *Visual feature integration and the temporal correlation hypothesis*, Ann. Rev. Neurosci., 18 (1995), pp. 555–586.
[14] S. H. Strogatz, *Norbert Wiener's Brain Waves*, Lecture Notes in Biomath. 100, Springer-Verlag, Berlin, 1994.
[15] S. H. Strogatz, *From Kuramoto to Crawford: Exploring the onset of synchronization in populations of coupled oscillators*, Phys. D, 143 (2000), pp. 1–20.
[16] N. Wiener, *Nonlinear Problems in Random Theory*, MIT Press, Cambridge, MA, 1958.
[17] N. Wiener, *Cybernetics*, MIT Press, Cambridge, MA, 1961.
[18] K. Wiesenfeld, P. Colet, and S. H. Strogatz, *Synchronization transitions in a disordered Josephson series array*, Phys. Rev. Lett., 76 (1996), pp. 404–407.
[19] K. Wiesenfeld, P. Colet, and S. H. Strogatz, *Frequency locking in Josephson arrays: Connection with the Kuramoto model*, Phys. Rev. E, 57 (1998), pp. 1563–1569.
[20] A. T. Winfree, *Biological rhythms and behavior of populations of coupled oscillators*, J. Theoret. Biol., 16 (1967), pp. 15–42.

SIAM J. APPL. MATH.
Vol. 63, No. 5, pp. 1627–1647

NONLINEAR COUPLING NEAR A DEGENERATE HOPF (BAUTIN) BIFURCATION*

JONATHAN D. DROVER† AND BARD ERMENTROUT†

Abstract. A nonlinearly coupled system of bistable (fixed point and limit cycle) differential equations is analyzed. The nonlinear equations arise from the first several terms in the normal form expansion near a Bautin bifurcation. Existence and stability of in-phase and out-of-phase periodic solutions to a pair of identical systems are explored. Existence, uniqueness, and stability of traveling wave solutions from a stable rest state to a stable periodic solution are proved for the associated evolution/convolution equation. Numerical simulations suggest some interesting patterns in regimes where waves no longer exist. The results are shown to hold for a nonreduced conductance-based model.

Key words. Bautin bifurcation, subcritical Hopf bifurcation, bistability, traveling waves, localized pulses

AMS subject classifications. 37N25, 34C15, 45M99

DOI. 10.1137/S0036139902412617

1. Introduction.

The analysis of the behavior of coupled neuronal oscillators and "near" oscillators (i.e., excitable cells) has been the subject of many recent papers [2, 13, 15, 14]. Intrinsic neuronal oscillations arise primarily via two distinct mechanisms [21, 15]: (i) a saddle-node on a limit cycle bifurcation or (ii) a subcritical or supercritical Hopf bifurcation. In the saddle-node bifurcation, the oscillations that arise are large-amplitude, and so it is possible to study the effects of coupling between them by looking at certain normal forms that arise [14, 8]. Coupled systems near a supercritical Hopf bifurcation have been the subject of numerous studies (e.g., Aronson, Ermentrout, and Kopell [1]), most recently by Hoppensteadt and Izhikevich [13]. The problem with such analysis is that the coupling that has been analyzed is linear. This means that the coupling itself can determine whether or not the system is at rest or oscillates. Chemical synaptic coupling is inherently nonlinear, unlike coupling via diffusive-like interactions. This is because subthreshold oscillations and perturbations from rest are insufficient to excite the channels which release the chemical transmitters necessary for communication between neurons. Because synaptic coupling between neurons is nonlinear, the presence of coupling does not alter the stability of the resting state of such a neuron. This contrasts with diffusive or gap junctional coupling, which is linear and can therefore affect the stability of the resting state [7, 10, 19].

The oscillations emerging from a *subcritical* Hopf bifurcation generally "turn around" for neuronal models to become large-amplitude stable oscillations. Thus, like the saddle-node case, these large-amplitude oscillations are sufficient to excite chemical synapses. Furthermore, unlike a supercritical Hopf bifurcation, there is a range of parameters for which the system is intrinsically bistable: there is a stable equilibrium point and a stable oscillation. The goal of this paper is to study a reduced model (normal form) that nonlinearly couples systems with a subcritical Hopf bifurcation.

*Received by the editors August 2, 2002; accepted for publication (in revised form) February 3, 2003; published electronically July 26, 2003. Parts of this work comprised the Master's thesis for the first author. This work was supported in part by NSF grant DMS-9972913.
http://www.siam.org/journals/siap/63-5/41261.html
†Department of Mathematics, Thackeray 301, University of Pittsburgh, Pittsburgh, PA 15260 (jddst25@pitt.edu, bard@pitt.edu).

We model the bistable subcritical Hopf bifurcation by considering a degenerate form of the Hopf bifurcation, which arises at the transition between sub- and supercritical bifurcations. This is called the Bautin bifurcation, and the normal form for such a bifurcation is

$$z' = z(\lambda + b|z|^2 + f|z|^4) \equiv N(z), \tag{1.1}$$

where b, f are complex numbers and λ is the bifurcation parameter. The usual Hopf bifurcation does not involve the parameter f and is sub- or supercritical according to whether the real part of b is, respectively, positive or negative. In the Bautin bifurcation, the real part of the coefficient b vanishes. Thus, the degeneracy in the normal form arises from the nonlinear terms in the system. The advantage of the Bautin normal form is that it captures the bistability of the medium as well as the fact that the stable oscillation is bounded away from the rest state. Equation (1.1) can be derived from any nonlinear system near a degenerate Hopf bifurcation [17].

Consider, now, a pair of synaptically coupled neurons near this bifurcation. In the absence of coupling, they can be described in terms of the complex amplitudes z_1, z_2, where each z_j satisfies (1.1). Coupling alters the normal form by adding new terms, whose form can be deduced by using standard symmetry arguments [12] or by direct (albeit tedious) calculation. The coupled system has the form

$$\begin{aligned} z_1' = {} & N(z_1) + L_1 z_1 + L_2 z_2 + C_1 z_1|z_2|^2 + C_2 z_2|z_2|^2 \\ & + C_3 z_1^2 \bar{z}_2 + C_4 z_2^2 \bar{z}_1 + C_5 z_2|z_1|^2 + C_6 z_1|z_1|^2, \end{aligned} \tag{1.2}$$

with an analogous equation for z_2'. If the coupling between oscillators is through diffusion (e.g., gap junctions) then $L_1 = -L_2$. In [1, 13] only the linear coupling terms are kept. [16] studied (1.1) in the context of synchronization between bursters with linear coupling. We have also included nonlinear coupling terms up to order 3 in (1.2). The motivation for this is as follows. Consider the stability of the origin $z_j = 0$. The linearized equations have the form

$$\begin{aligned} z_1' &= (\lambda + L_1) z_1 + L_2 z_2, \\ z_2' &= (\lambda + L_1) z_2 + L_2 z_1, \end{aligned}$$

with eigenvalues $\lambda + L_1 \pm L_2$. Thus, linear coupling can alter the stability of the origin. This cannot happen in a system of synaptically coupled neurons except in very unusual circumstances. That is, the threshold for synaptic interactions would have to be nearly identical to the resting state of the neuron. For this reason, we will assume that the coupling between the two normal forms should not be linear. Thus, in this paper, we set $L_1 = L_2 = 0$.

We turn to the nonlinear coupling terms, of which there are six. Consider two neurons which are stably at rest ($z_j = 0$). Suppose that we excite one of them past threshold so that it begins to fire. Then if the neurons are coupled sufficiently strongly, this should induce the other neuron to begin firing. Of the six coupling terms in (1.2), only one term can cause this. Since $z_j = 0$ is the rest state, then any coupling term which includes z_j cannot contribute to pushing z_j away from rest. For example, if z_2 is oscillating, then the only term which can influence a *resting* z_1 is the second term:

$$z_2|z_2|^2.$$

While all terms are important once the neurons are both oscillating, the onset of oscillations is effected only through the second term. Thus, we will restrict the analysis

of coupled systems to nonlinear coupling with $C_n = 0$ except for C_2. If there is linear coupling and we are very close to the bifurcation, then the nonlinear coupling can be scaled out. Thus, one should regard our analysis, which includes nonlinear coupling terms, to hold for systems that are a small distance from the actual bifurcation. As the magnitude of the linear terms decreases (and for synaptic coupling, this is very small), the nonlinear terms have a stronger effect on the behavior, and thus in principle, we need not be very far from the critical bifurcation at all.

In the first part of the paper, we describe the bifurcations that occur in

$$\begin{aligned} z_1' &= z_1(\lambda + (iq)|z_1|^2 - |z_1|^4) + (c_1 + ic_2)z_2^2\bar{z}_2, \\ z_2' &= z_2(\lambda + (iq)|z_2|^2 - |z_2|^4) + (c_1 + ic_2)z_1^2\bar{z}_1, \end{aligned} \tag{1.3}$$

where $b = iq$ (at the Bautin bifurcation) and all parameters are real. We restrict our attention to the case in which the coefficient f in (1.1) is real and negative since we want the resulting large-amplitude oscillations to be stable. We set $f = -1$ with no loss in generality. We establish the existence and stability of in- and out-of-phase oscillations in preparation for section 3 in which we look at spatially distributed networks.

Spatially distributed neurons and waves have been the object of much recent analysis. Coupling between neurons is not restricted to nearest neighbors but instead can involve long spatial scales. Typically, when modeled as a continuum, interactions take the form of convolutions. The coupled pair of Bautin oscillators can be easily generalized to a continuous spatial model with convolution coupling:

$$z_t = z(\lambda + (iq)|z|^2 - |z|^4) + (c_1 + ic_2)\int_{-\infty}^{\infty} J(x-y)z^2(y)\bar{z}(y)dy. \tag{1.4}$$

In the second part of the paper, we will describe the existence of traveling wavefronts, which join the stable resting state to an oscillatory solution. The resulting waves are similar in structure to those found in [9] when coupling was diffusive. We analyze the existence and stability of plane waves for (1.4) as well. We use numerical simulations to find spatially localized patterns which may be analogous to patterns of activity used to model working memory. Finally, we close the paper with simulations of a simple biophysically based model in order to demonstrate that the behavior of the normal-form-based model holds in more "realistic" neural models.

2. Two coupled equations. In this section we determine under what conditions the symmetric and asymmetric solutions exist and are stable. For this analysis, it is easiest to put the equations into polar form. We let

$$z_j = r_j e^{i\theta_j}$$

and let $\phi = \theta_1 - \theta_2$. Then, substituting into (1.3), we obtain the equations

$$\begin{aligned} r_1' &= \lambda r_1 - r_1^5 + c_1 r_2^3\cos(\phi) + c_2 r_2^3\sin(\phi), \\ r_2' &= \lambda r_2 - r_2^5 + c_1 r_1^3\cos(\phi) - c_2 r_1^3\sin(\phi), \\ \phi' &= q(r_1^2 - r_2^2) - c_1\left(\frac{r_1^3}{r_2} + \frac{r_2^3}{r_1}\right)\sin(\phi) + c_2\left(\frac{r_2^3}{r_1} - \frac{r_1^3}{r_2}\right)\cos(\phi). \end{aligned} \tag{2.1}$$

Phase-locked solutions with constant r_1, r_2 to the coupled system of oscillators are fixed points of (2.1). The existence and stability of periodic solutions is readily obtained. We remark that $z_j = 0$ is an asymptotically stable solution to (1.3) if and only

if $\lambda < 0$. In this section, we will look for limit-cycle solutions to (1.3) with amplitudes r_j bounded away from zero so that the denominators in the equation for ϕ cause no problems.

2.1. Symmetric in-phase solutions. We look for solutions of the form $\rho = r_1 = r_2$ (symmetric) and $\phi = 0$. When $\phi = 0$, the sine terms in (2.1) vanish. Inserting the solution $\phi = 0$, (2.1) is reduced to the first order ODE

$$\rho' = \lambda\rho + c_1\rho^3 - \rho^5. \tag{2.2}$$

The positive roots are given by

$$\frac{1}{2}\sqrt{2c_1 \pm 2\sqrt{c_1^2 + 4\lambda}}. \tag{2.3}$$

First, we will establish when these solutions exist.

2.1.1. Existence. We want to determine under what conditions on the parameters the solutions (2.3) are real. Thus we want to know when the polynomial $f(x) = \lambda + c_1x^2 - x^4$ has two positive, real roots. For $\lambda > 0$ this polynomial will always have one positive root; thus the upper branch of solutions (2.3) will be real. Now, because this polynomial is symmetric about the y-axis, in order for two real, positive roots to exist it must be true that $f(0) < 0$, implying $\lambda < 0$. Also, there must be a relative maximum at a point $x = a$ so that $f(a) > 0$. This maximum occurs at $x = \sqrt{c_1/2}$. Thus, the polynomial $f(x)$ will have two positive, real roots if $c_1 > \sqrt{-2\lambda}$.

One of the goals of this paper is to prove the existence of traveling waves connecting a stable rest state to a stable plane wave. Thus, we must choose parameters so that equations lie in a bistable regime. For this reason, we will only consider parameter values that allow two stable states. Thus, we must have two nonzero solutions to separate stable regions. We will consider only $\lambda < 0$. The only requirement that follows is that $c_1 > \sqrt{-2\lambda}$ must hold.

2.1.2. Stability. The Jacobian for the polar coordinate system is

$$\begin{pmatrix} \alpha(r) & \beta(r) & \gamma(r) \\ \beta(r) & \alpha(r) & -\gamma(r) \\ a(r) & -a(r) & b(r) \end{pmatrix}, \tag{2.4}$$

where $\alpha(r) = \lambda - 5r^4$, $\beta(r) = 3c_1r^2$, $\gamma(r) = c_2r^2$, $a(r) = 2qr - 4c_2r$, and $b(r) = -2c_1r^2$.

The characteristic polynomial for (2.4) is

$$(x - \alpha - \beta)(x^2 + (\beta - b - \alpha) + \alpha b - \beta b - 2\gamma a).$$

Thus, a symmetric in-phase solution is stable if the following three conditions hold at the equilibrium:

$$\alpha + \beta < 0, \tag{2.5}$$

$$\beta - b - \alpha > 0, \tag{2.6}$$

$$\alpha b - \beta b - 2\gamma a > 0. \tag{2.7}$$

The lower branch of solutions is always unstable. Substituting the value of the lower branch solution into $\alpha + \beta$ gives

$$-4\lambda - c_1^2 + c_1\sqrt{c_1^2 + 4\lambda}.$$

By the requirement for existence, we have that $c_1 > \sqrt{-2\lambda}$, and so the above quantity is real. It is easily seen that the expression is the same as

$$\sqrt{c_1^2 + 4\lambda}\left[c_1 - \sqrt{c_1^2 + 4\lambda}\right] > 0,$$

and so the value (2.5) is positive. Thus this periodic solution to (2.2) is unstable and will act as the desired separatrix between the stable fixed point and the upper periodic branch. A similar proof is used to show that this eigenvalue is always negative for the upper branch of solutions; thus it will not prevent stability.

We turn now to the other two conditions. First, consider condition (2.6). Using the definitions of the various parameters, this condition becomes

$$5c_1^2 + 4\lambda + 5c_1 R > 0,$$

where $R = \sqrt{(c_1^2 + 4\lambda)} > 0$. Since $c_1^2 + 4\lambda > 0$, this expression is clearly positive, and thus the condition (2.6) holds on the upper branch. The last condition, (2.7), will not hold for all parameters, and in fact, as we will show, the left-hand side can become negative if the parameters c_2, q are sufficiently large. We can rewrite condition (2.7) as

$$F(c_2) \equiv Ac_2^2 + Bc_2 + C > 0,$$

where

$$\begin{aligned} A &= 4c_1^2 + 8\lambda + 4c_1 R, \\ B &= -q(c_1^2 + 4\lambda + c_1^2 + c_1 R) \equiv -qD, \\ C &= 4c_1 R(\lambda + 2c_1^2) + 4c_1^2(5\lambda + 2c_1^2). \end{aligned}$$

We note that since $c_1^2 + 4\lambda > 0$, $\lambda < 0$, and $c_1 > 0$, A, D, C are all positive. In particular, if $c_2 = 0$, then condition (2.7) holds and the upper branch is asymptotically stable. Suppose that c_2 is nonzero. The parameter q appears only in the coefficient B. Furthermore, $D > 0$ so that for fixed c_2 we can always find a sufficiently large value of $|q|$ so that $F(c_2) < 0$ and the upper branch is destabilized. We should choose c_2, q to have the same sign. The critical value of q leads to a zero eigenvalue for the stability matrix. Since the synchronous solution cannot disappear (its existence is independent of both c_2, q), the bifurcation occurring must be either a transcritical or a pitchfork. However, due to symmetry, the transcritical cannot occur, and the resulting bifurcation must be a pitchfork [23]. The critical value of q is readily found to be

$$q^* = \frac{Ac_2^2 + C}{Dc_2}.$$

Using AUTO [6], we have tracked the solution through this instability and verified that it is indeed a pitchfork bifurcation. For $\lambda = -1/5$, $c_1 = c_2 = 1$, we find that the synchronous state becomes unstable at $q^* = 5.4473$, in agreement with the above formula. Furthermore, the bifurcation is subcritical. Thus, once the critical value of q is exceeded, we find that the only stable state is the resting value, $r_1 = r_2 = 0$.

2.2. Symmetric out-of-phase solutions. In order for the out-of-phase solutions to exist we must have that $c_1 < 0$. In this paper we will assume that $c_1 > \sqrt{-2\lambda}$ and $\lambda < 0$, so these solutions are of little interest here.

2.3. Two-equation summary. For $\lambda < 0$ and $c_1 > \sqrt{-2\lambda}$ there are two in-phase symmetric periodic solutions. The top branch of solutions is stable as long as q is sufficiently small. The lower branch is unstable. These solutions, and the zero solution which is linearly stable, provide a bistable region. For the two-equation system this is of little consequence in the existence and stability of phase-locked periodic solutions; however, when modeled in continuum, the bistability leads to traveling waves and pulsing behaviors that connect the stable origin to the stable periodic solutions, as we shall see in the subsequent sections.

3. The convolution equation. As stated in the introduction, we will consider the equation

$$z_t = z[\lambda + (b_1 + iq)|z|^2 - |z|^4] + (c_1 + ic_2)\int_{-\infty}^{\infty} J(x-y)z^2(y)\overline{z}(y)dy,$$

where $\int_{-\infty}^{\infty} J(x)dx = 1$, $J(x) \geq 0$, and $||J||_\infty < \infty$. We remark that if $c_2 = 0$, then, no matter what the value of q, the synchronous periodic state $z(x,t) = \rho \exp(i\Omega t)$ is always stable when it exists. We will also assume that $b_1 \geq 0$ since we want the local Hopf bifurcation to be subcritical (and, in fact, throughout most of the paper, we assume $b_1 = 0$). We will first study the existence and stability of plane wave solutions to this system. We then turn to the existence of traveling fronts, which join the stable equilibrium point, $z = 0$, to a plane wave. Finally, we look at the behavior of these fronts as q increases. We present numerical evidence for the existence of localized regions of periodic activity surrounded by the near absence of activity.

3.1. Relationship of the normal form to a kinetic model. The relationship between the coupling in an actual kinetic model and the normal form given in (1.4) is derived here. Consider the following system consisting of two synaptically coupled neurons:

$$\begin{aligned}
\frac{dV_1}{dt} &= f(V_1) - gs_2(V_1 - V_{syn}),\\
\frac{ds_1}{dt} &= \alpha(V_1)(1-s_1) - \beta s_1,\\
\frac{dV_2}{dt} &= f(V_2) - gs_1(V_2 - V_{syn}),\\
\frac{ds_2}{dt} &= \alpha(V_2)(1-s_2) - \beta s_2.
\end{aligned}$$

Assume that $f(V_{rest}) = 0$ and that $\frac{\partial f}{\partial V}(V_{rest}) < 0$. Then the uncoupled system has a stable rest state at $V = V_{rest}$. Suppose that $\alpha(V) = \alpha'(V) = 0$ for V less than some critical value, which itself is greater than V_{rest}. Then V_{rest} will remain a stable equilibrium point for the coupled system. Hence, the coupling will not affect the existence or stability of the rest state. If we consider N neurons coupled similarly, we get a system

$$\begin{aligned}
\frac{dV_i}{dt} &= f(V_i) - g(V_i - V_{syn})\sum_j w_{i-j}s_j,\\
\frac{ds_i}{dt} &= \alpha(V_i)(1-s_i) - \beta s_i.
\end{aligned}$$

We may slave the coupling directly to the presynaptic potentials. Because we have that $\alpha(V_{rest}) = \alpha'(V_{rest}) = 0$, the expression must be at least second order, and we

get that

$$\frac{dV_i}{dt} = f(V_i) - g(V_i - V_{syn}) \sum_j w_{i-j} R(V_j - V_{rest}),$$

where R is a function with no linear component. Because we are allowing only terms that contribute to excitation away from rest, we may simplify even further:

$$\frac{dV_i}{dt} = f(V_i) - g(V_{rest} - V_{syn}) \sum_j w_{i-j} R(V_j - V_{rest}).$$

And finally, if modeled in continuum, we get that

$$V_t(x,t) = f(V(x,t)) - \hat{g} \int_{\infty}^{\infty} J(x-y) R(V(y,t) - V_{rest}) dy,$$

as desired. (Note that $\hat{g} = g(V_{rest} - V_{syn})$.)

3.2. Plane wave solutions. Consider solutions of (1.4) of the form $\rho e^{i(\Omega t - kx)}$. Substituting this in yields

$$\begin{aligned} i\rho\Omega e^{i(\Omega t-kx)} &= \rho e^{i(\Omega t-kx)}(\lambda + (b_1 + iq)\rho^2 - \rho^4) \\ &\quad + (c_1 + ic_2)\int_{-\infty}^{\infty} J(x-y)\rho^3 e^{i(\Omega t - ky)} dy, \end{aligned}$$

which implies

$$i\Omega = \lambda + (b_1 + iq)\rho^2 - \rho^4 + \rho^2(c_1 + ic_2)\int_{-\infty}^{\infty} J(x-y)e^{ik(x-y)} dy$$

and finally

$$i\Omega = \lambda + (b_1 + iq)\rho^2 - \rho^4 + \rho^2(c_1 + ic_1)\hat{J}(k), \tag{3.1}$$

where $\hat{J}(k) = \int_{-\infty}^{\infty} J(s)e^{iks} ds$. Taking the imaginary part of the right-hand side of (3.1), we find that

$$\Omega = q\rho^2 + c_1\rho^2 \mathrm{Im}(\hat{J}(k)) + c_2\rho^2 \mathrm{Re}(\hat{J}(k)), \tag{3.2}$$

with ρ determined so that

$$\lambda + b_1\rho^2 - \rho^4 + c_1\rho^2 \mathrm{Re}(\hat{J}(k)) - c_2\rho^2 \mathrm{Im}(\hat{J}(k)) = 0. \tag{3.3}$$

For the remainder of the paper, we will assume that $c_2 = 0$ and that $J(x)$ is a symmetric kernel; thus (3.2) and (3.3) become, respectively,

$$\Omega = q\rho^2 \tag{3.4}$$

and

$$\lambda - \rho^4 + c_1\rho^2 \hat{J}(k) = 0. \tag{3.5}$$

In the next subsection, we prove the asymptotic stability of these plane waves so that we can then show the existence of traveling wavefronts connecting the stable rest state to these oscillations. We remark that setting $c_2 = 0$ is a major simplification. It is well known that if $c_2 q$ is large enough, then the corresponding Ginzburg–Landau model has spatiotemporal chaos. (See, for example, [20], where spatiotemporal chaos is explored for $z_t = z(a + bz\bar{z}) + dz_{xx}$. See also section 3.4.3, where such a solution is exhibited for the present model.)

3.3. Plane wave stability. We claim that the plane wave solutions defined by (3.5), (3.4) are asymptotically stable for sufficiently small k. We will show this using a symmetric kernel $J(x)$ such that $\frac{dJ}{dx}(0) = 0$ and $\int_{-\infty}^{\infty} J(x)dx = 1$. The kernel we use is $(1/\sqrt{\pi})e^{-x^2}$, and the results can be generalized to any function satisfying the above requirements.

The plane wave has the form $z(x,t) = \rho e^{i(\Omega t - kx)}$, with ρ and Ω defined as in (3.5) and (3.4). We now examine the stability of the plane waves by linearizing about a solution. Let

$$z(x,t) = \rho e^{i(\Omega t - kx)} + w(x,t). \tag{3.6}$$

Substituting (3.6) into (1.4) and taking terms linear in w, we obtain

$$w_t = f(w) + c_1\rho^2 \int_{-\infty}^{\infty} J(x-y)\left[e^{2i(\Omega t - kx)}\bar{w} + 2w\right] dy, \tag{3.7}$$

where

$$f(w) = \lambda w + iq\rho^2\left[2w + e^{2i(\Omega t - kx)}\bar{w}\right] - \rho^4\left[3w + 2e^{2i(\Omega t - kx)}\bar{w}\right]. \tag{3.8}$$

Now, let $w = e^{i(\Omega t - kx)}v$ and $v = v_1 + iv_2$ and substitute into (3.7). This gives the equation

$$\begin{aligned} i\Omega(v_1 + iv_2) + (v_1 + iv_2)_t \\ = g(v_1, v_2) + c_1\rho^2 \int_{-\infty}^{\infty} J(s)e^{iks}\left[3v_1(x-s) + iv_2(x-s)\right] ds, \end{aligned} \tag{3.9}$$

where

$$g(v) = \lambda(v_1 + iv_2) + 2iq\rho^2(v_1 + iv_2) + iq\rho^2(v_1 - iv_2) - 3\rho^4(v_1 + iv_2) - 2\rho^4(v_1 - iv_2).$$

Expanding the integrand in (3.9) and separating the real and imaginary parts gives

$$c_1\rho^2 \int_{-\infty}^{\infty} J(s)\left[3\cos(ks)v_1(x-s) - \sin(ks)v_2(x-s)\right] ds \tag{3.10}$$

and

$$c_1\rho^2 \int_{-\infty}^{\infty} J(s)\left[\cos(ks)v_2(x-s) + 3\sin(ks)v_1(x-s)\right] ds. \tag{3.11}$$

Let $g = g_1 + ig_2$. Then these are

$$g_1(v) = \lambda v_1 - q\rho^2 v_2 - 5\rho^4 v_1 \tag{3.12}$$

and

$$g_2(v) = \lambda v_2 + 3q\rho^2 v_1 - \rho^4 v_2. \tag{3.13}$$

Because $\Omega = q\rho^2$, we can subtract this from both sides of (3.9). We define

$$h_1(v) = g_1(v) + q\rho^2 v_2 = \lambda v_1 - 5\rho^4 v_1, \tag{3.14}$$

$$h_2(v) = g_2(v) - q\rho^2 v_1 = \lambda v_2 + 2q\rho^2 v_1 - \rho^4 v_2. \tag{3.15}$$

Now, let $v_1 = e^{\alpha t}e^{ilx}u_1$ and $v_2 = e^{\alpha t}e^{ilx}u_2$. Then (3.10) becomes

$$c_1\rho^2 e^{\alpha t}e^{ilx}\int_{-\infty}^{\infty} J(s)\left[3\cos(ks)e^{-ils}u_1 - \sin(ks)e^{-ils}u_2\right]ds,$$

which, because J is assumed to be even, simplifies to

$$c_1\rho^2 e^{\alpha t}e^{ilx}\left[3\int_{-\infty}^{\infty} J(s)\cos(ks)\cos(ls)u_1 ds + \int_{-\infty}^{\infty} J(s)\sin(ks)\sin(ls)u_2 ds\right]. \tag{3.16}$$

Similarly, (3.11) becomes

$$c_1\rho^2 e^{\alpha t}e^{ilx}\left[\int_{-\infty}^{\infty} J(s)\cos(ks)\cos(ls)u_2 ds - 3\int_{-\infty}^{\infty} J(s)\sin(ks)\sin(ls)u_1 ds\right]. \tag{3.17}$$

Thus, (3.9) becomes

$$\alpha u_1 = \left[\lambda - 5\rho^4 + 3c_1\rho^2 J_{cc}\right]u_1 + iJ_{ss}u_2, \tag{3.18}$$

$$\alpha u_2 = \left[\lambda - \rho^4 + c_1\rho^2 J_{cc}\right]u_2 + \left[2q\rho^2 - iJ_{ss}\right]u_2, \tag{3.19}$$

where

$$J_{ss} = \int_{-\infty}^{\infty} J(s)\sin(ks)\sin(ls)ds$$

and

$$J_{cc} = \int_{-\infty}^{\infty} J(s)\cos(ks)\cos(ls)ds.$$

These can be expressed in terms of the Fourier transform $\hat{J}(k)$ of the kernel

$$J_{cc} = \frac{1}{2}\hat{J}(k+l) + \frac{1}{2}\hat{J}(k-l)$$

and

$$J_{ss} = -\frac{1}{2}\hat{J}(k+l) + \frac{1}{2}\hat{J}(k-l).$$

For the kernel J chosen, we have the transform $\hat{J}(k) = e^{-\frac{1}{4}k^2}$. The top panel of Figure 1 shows that, for $k = 0$, the real part of the greatest eigenvalue is negative for all $l \neq 0$. The middle panel shows that when $k = 1$, the plane wave solution is unstable, with the real part of the greatest eigenvalue positive. The bottom panel shows that there is a $k > 0$ such that the corresponding plane wave is stable. Thus there is an interval surrounding $k = 0$ such that the perturbation goes to zero, and hence the plane wave is stable.

This result is generalizable to any symmetric, nonnegative, integrable kernel qualitatively like a Gaussian. We need only that $\hat{J}$ decreases as $|k|$ gets farther from 0.

3.4. Traveling wavefronts. We will show that, under certain restrictions on the parameters, there exists a traveling wavefront that takes the system from the stable equilibrium point $z = 0$ to the plane wave solutions defined by (3.4), (3.5).

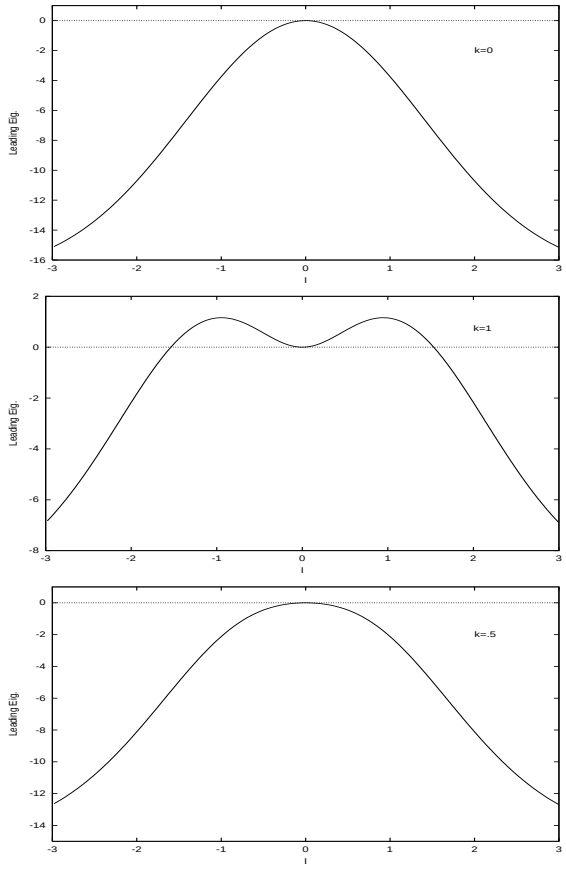


FIG. 1. *The real part of the leading eigenvalue for $k = 0$, $k = 1$, and $k = .5$. The horizontal axis is l, and the vertical axis is the value of the real part of the eigenvalue.*

Let $Z(k; x, t) = \rho(k) \exp(i[\Omega(k)t - kx])$ be a plane wave as constructed above. For $k = k^*(q)$ we seek traveling waves connecting $z(x, t) = 0$ to $Z(k^*; x, t)$. That is, there exists a real valued function $h(\xi)$ and a real c such that

$$z(x, t) = h(x - ct)Z(k^*; x, t),$$

where $h(-\infty) = 0$ and $h(\infty) = 1$. In this paper, we prove only the $q = 0$ case.

We put the system (1.4) into polar coordinates, making the substitution $z = re^{i\theta}$:

$$r_t = \lambda r + b_1 r - r^5 + c_1 G_1(x, t) + c_2 G_2(x, t), \tag{3.20}$$

$$r\theta_t = qr^3 + c_1 G_2(x, t) + c_2 G_1(x, t), \tag{3.21}$$

where

$$G_1(x, t) = \int_{-\infty}^{\infty} J(x - y)r^3(y)\cos(\theta(x) - \theta(y))dy, \tag{3.22}$$

$$G_2(x, t) = \int_{-\infty}^{\infty} J(x - y)r^3(y)\sin(\theta(x) - \theta(y))dy.$$

First, suppose that $b_1 = q = c_2 = 0$. Then $\theta = 0$ satisfies (3.21). This makes (3.20)

$$\rho' = \lambda\rho + b_1\rho^3 - \rho^5 + c_1 \int_{-\infty}^{\infty} J(x-y)\rho^3(y)dy. \tag{3.23}$$

3.4.1. Existence, uniqueness, asymptotic stability. Proving the existence, uniqueness, and asymptotic stability of a traveling wave connecting the stable equilibrium at $z = 0$ with a stable periodic orbit is done by appealing to the following theorem.

THEOREM 3.1 (see Chen [4]). *Consider the evolution equation*

$$u_t = Du_{xx} + G(u, J_1 * S_1(u), \ldots, J_n * S_n(u)), \tag{3.24}$$

where $J*S(u)$ stands for the convolution $\int_R J(x-y)S(u(y))dy$. Assume the following:

1. *For some $a \in (0,1)$, the function $f(u) = G(u, S_1(u), \ldots, S_n(u))$ satisfies $f > 0$ in $(-1,0) \cup (a,1)$, $f < 0$ in $(0,a) \cup (1,2)$, and $f'(0) < 0$, $f'(a) > 0$, and $f'(1) < 0$.*

2. *For each $i = 1, \ldots, n$, the kernel J_i is C^1 and satisfies $J_i(.) \geq 0$, $\int_R J_i(y)dy = 1$, and $\int_R |J_i(y)|dy < \infty$.*

3. *The functions $G(u,p)$ ($p = (p_1, \ldots, p_n)$) and $S_1(u), \ldots, S_n(u)$ are smooth functions satisfying for all $u \in [-1,2]$, $p \in [-1,2]^n$, $i = 1, \ldots, n$, $G_{p_i}(u,p) \geq 0$, $S_{u_i}(u) \geq 0$.*

4. *Either $D > 0$ or $G_u(u,p) < 0$ and $G_{p_1}(u,p)S_{u_1}(u) > 0$ on $[-1,2]^{n+1}$.*

If conditions 1–4 hold, there exists a unique (up to a translation) asymptotically stable, monotone traveling wave connecting 0 to 1.

Remark. In our system, the fixed point $\rho = 0$ corresponds to the zero fixed point of (1.4), and the fixed point $\rho > 0$ corresponds to the synchronous periodic solution to (1.4). Here $k^*(0) = 0$.

We now show that the following assumptions hold for (3.23).

1. The function $f(x) = \lambda x + b_1x^3 - x^5 + c_1x^3$ must satisfy a number of conditions regarding stability. This is to ensure that there is both a stable equilibrium and a stable periodic solution with an unstable periodic solution between the two acting as a separatrix. The first condition is $f(0) = 0$, $f(a) = 0$, and $f(b) = 0$, where b is the amplitude of the stable periodic orbit and a is the amplitude of the unstable separatrix. These values can be scaled so that $b = 1$; however, that is not necessary to satisfy the assumption. The requirements that $f'(0) < 0$ and $f'(b) < 0$ are simply to ensure that both the rest state and periodic solution are stable solutions. $f(x)$ does satisfy these conditions, as shown in section 1.2, so the first assumption in Chen's theorem is satisfied.

2. $\int_{-\infty}^{\infty} J(x-y)dy = 1$, $J(x) \geq 0$, and $J(x)$ is bounded for all x. We defined J in this way, so this assumption is satisfied.

3. $G_p(u,p) > 0$. For our purposes, this quantity is just c_1, so we will assume $c_1 > 0$. This is necessary for the existence of a nonzero root for f.

4. $G_u(u,p) < 0$ and $S_u(u) > 0$. The first of these is true for sufficiently small b_1. Because we have assumed that $b_1 = 0$, this condition is satisfied. The second quantity, $3c_1u^2$, is slightly problematic because it vanishes at $u = 0$; however, because it holds for all $u \neq 0$, the inequality that this condition is used to satisfy is still satisfied for our purposes (see the proof of Chen's theorem). Hence, this assumption is satisfied.

Because all four of the assumptions are valid for the scalar equation, there exists a unique and asymptotically stable traveling wave connecting the stable fixed point at $z = 0$ and the stable periodic solution of (3.23).

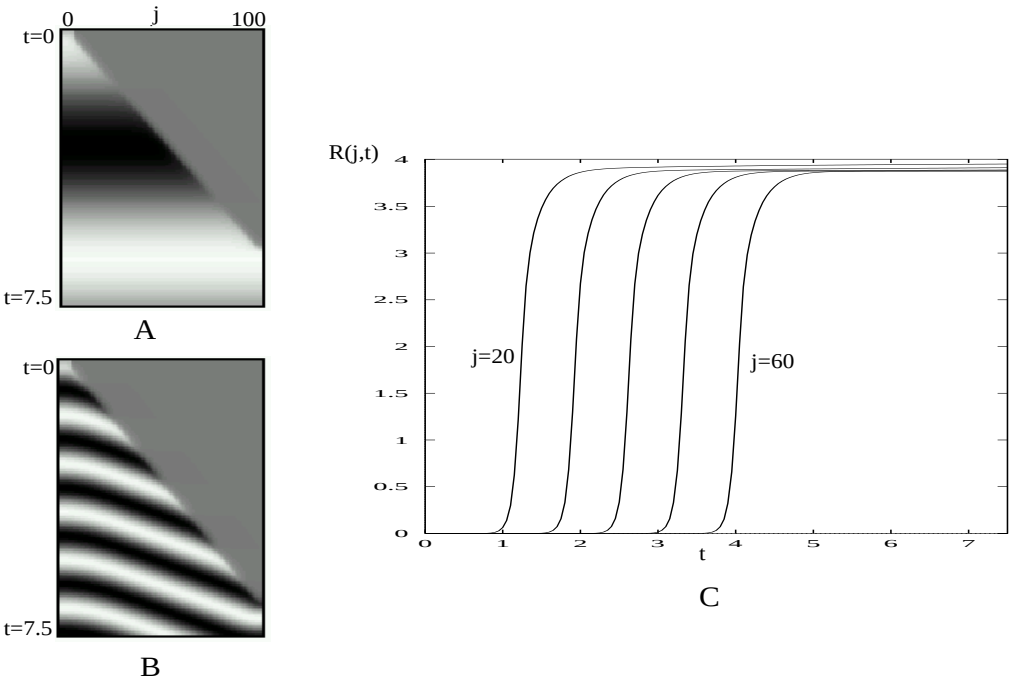


FIG. 2. *Traveling wave solution.* (A) $q = 0.0$, $\lambda = -0.2$, $c_1 = 1$. *This shows* $u(j,t)$ *with* j *horizontal and* t *vertical. Lightest color is* $u = 2$*, and darkest,* $u = -2$. (B) *Same as* (A) *but* $q = 1.0$. (C) *Amplitude* $R = u^2 + v^2$ *as a function of* t *for* $j = 20, 30, 40, 50, 60$*, for* $q = 1.0$.

Since $\theta = 0$, the imaginary part of z is never excited; hence the wave propagated will be from the stable solution $z(x) = 0$ for all $x > x_0$ (without loss of generality we may assume that the wave propagates from left to right) to the stable (parameter dependent) periodic $z(x) = \rho_0$ with zero imaginary part, as shown in Figure 2(A). There is no phase-gradient; all of the oscillators are perfectly synchronized after the wave passes. Because there is no imaginary component of the wave, it has wave number 0. Although we have no proof of existence of the wave for nonzero q, we expect that waves continue to exist at least for some finite range of q around 0. Figure 2(B,C) shows a simulation of the equations for $q = 1.0$. The magnitude, $R = u^2 + v^2$, travels as a front, but there is a clear phase-gradient in the wake of the wave. The frequency of the oscillations is also higher. More properties of the case $q \neq 0$ are considered in the next section.

3.5. The parameter q.

3.5.1. Small q. Assume that $b_1 = c_2 = 0$ and that q is sufficiently small. (Here, small depends on the parameters λ and c_1.) For nonzero q, we cannot assume that $\theta = 0$, so that a phase-gradient will appear (see Figure 2(B)). This, in turn, lowers the effective coupling strength between the oscillators since the coupling includes the term $\cos(\theta(y) - \theta(x))$, and for a large enough gradient, this can be quite small. Thus, one effect of increasing q is the appearance of a phase-gradient which, in turn, weakens the effective coupling strength and thus should slow the wave down.

Figure 3 shows some properties of the wavefront as a function of the magnitude q for $\lambda = -0.2, c_1 = 1$ fixed. As expected from the above discussion, the parameter q

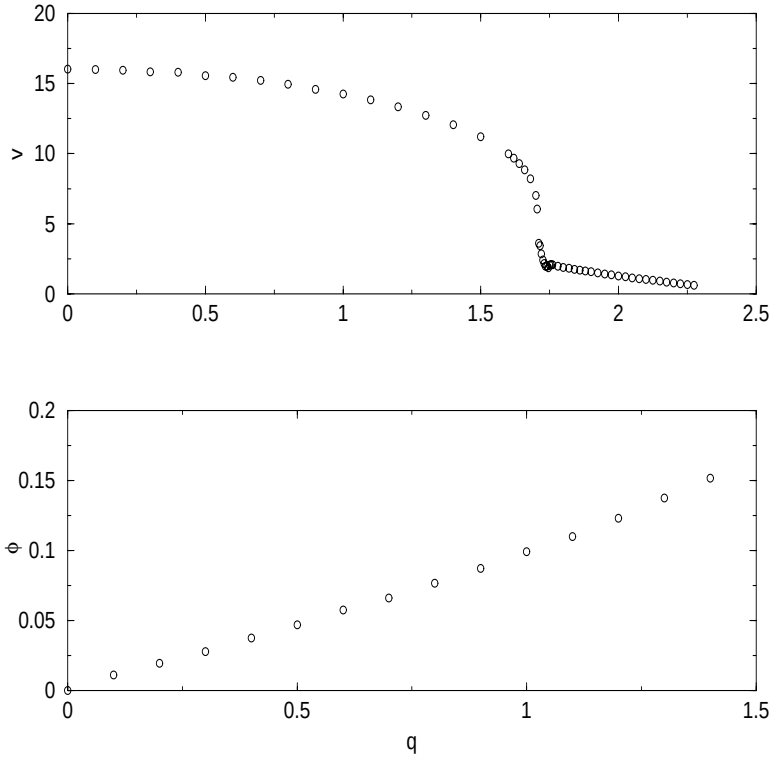


FIG. 3. *Dependence of the wave properties on the parameter q. Upper panel shows the velocity of the front for* $\lambda = -0.2$ *and* $c_1 = 1$. *Lower panel shows the phase-gradient. Note the different horizontal scales. The simulation consists of* 100 *cells coupled to* 20 *nearest neighbors with coupling strength* $\exp(-|j|/2)$. *Front velocity is measured as follows. Let* T_j *denote the time at which* $r_j = 1$. *Then the plotted velocity is just* $10/(T_{70} - T_{60})$, *that is,* $\Delta x/\Delta t$. *The phase-gradient is measured as follows. Let* $\theta_j = \arctan(v_j/u_j)$. *The quantity* $D(t) = (\theta_{60} - \theta_{40})/20 \approx \theta_x$ *is plotted, and the value after the second oscillation is taken to be an approximation of the phase gradient. In a finite domain,* $D(t)$ *always goes to zero since the network synchronizes.*

slows the wave down. For low values of q, this is a gradual decrease in speed. However, at $q \approx 1.75$ there is a precipitous drop in the velocity. For q larger than about 2.25, the wave ceases to exist and is replaced by a stable localized pulse (see below). The critical value of q depends on both the coupling strength c_1 and λ. The closer λ is to zero, the less excitation is required to cause propagation, due to the location of the unstable periodic orbit separating the rest state from the stable oscillation. Thus, for λ close to zero, wave propagation can occur for higher values of q. As λ gets larger in magnitude, the separatrix is farther from the stable rest state, and propagation of a wavefront requires more from the coupling, which large q prevents. This means that if q is fixed, we can achieve a similar slowing by altering the parameter λ. Indeed, we will exploit this in the next section, where we show similar phenomena for a conductance-based neural model. In addition to the drop in the velocity, the phase-gradient increases with q in an almost linear fashion. We compute the phase-gradient only up to about $q = 1.4$, as beyond that, the wave velocity becomes nearly zero, and whether or not there is a traveling front becomes ambiguous. The nearly linear dependence on q leads to a simplification for analyzing the effects of q on the wave velocity. Assume that the oscillations behind the front are quickly drawn to the plane

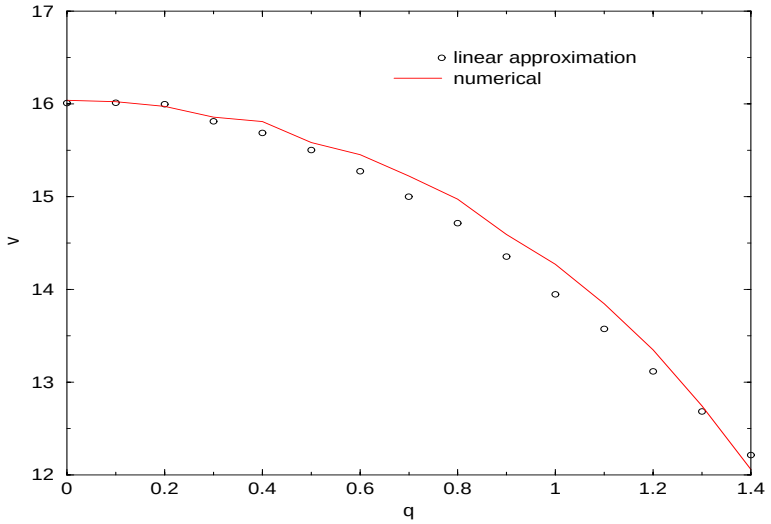


FIG. 4. *Velocity of the full equations (Figure* 3*, top) as* q *varies compared to the velocity computed from the approximate equation* (3.25) *using* $k(q) = q/10$*, which approximates the slope of the lower curve in Figure* 3.

wave and thus $\theta(x,t) = \Omega t + k(q)x$. Here $k(q) \approx mq$ is the asymptotic wave number, and m is the slope of the dependence. Then the amplitude evolves according to the equation

$$r_t = r(\lambda + b_1 r^2 - r^4) + c_1 \int_{-\infty}^{\infty} \tilde{J}(x-y) r^3(y)\, dy, \tag{3.25}$$

where $\tilde{J}(x) = J(x)\cos(k(q)x)$. This is identical (up to normalization) to the zero q model, for which we have proved the existence of a front. Thus, we expect that there will continue to be traveling fronts for q small enough. However, the new convolution kernel $\tilde{J}$ is narrower than the original, and thus we expect the velocity to decrease. Figure 4 shows a comparison of the front velocity for the approximation and the full equations. For values up to $q \approx 1.4$ the approximation is very good. (Recall that the computation of the phase became difficult beyond $q = 1.4$.) For larger values of q, the approximation was not very good.

3.5.2. Larger q and pulse formation. We saw above that the velocity of the front seems to go to zero as q approaches some fixed finite value. We can ask what happens for q beyond this point. One possibility is that the wave will not propagate at all and all initial data will return to the rest state, $z = 0$. However, numerical simulations indicate that rather than decay to rest, the medium remains excited locally and forms localized pulses. Because this excitation is necessary to get the θ variable excited, q large will prohibit excitation. The parameters are selected to lie in a bistable region, and thus the solution $z(t,x) = 0$ is stable. If there is not sufficient coupling, the wave will not propagate, and under sufficiently large initial conditions this can result in bumps or pulse solutions (see Figure 5). This figure shows the real part $u(x,t)$ in a simulation for $q = 3.25$, which is past the regime of existence of traveling fronts. The envelope $R = u^2 + v^2 = |z|^2$ of a pulse appears to be stationary (inset, Figure 5(B)), while the imaginary part appears to be periodic.

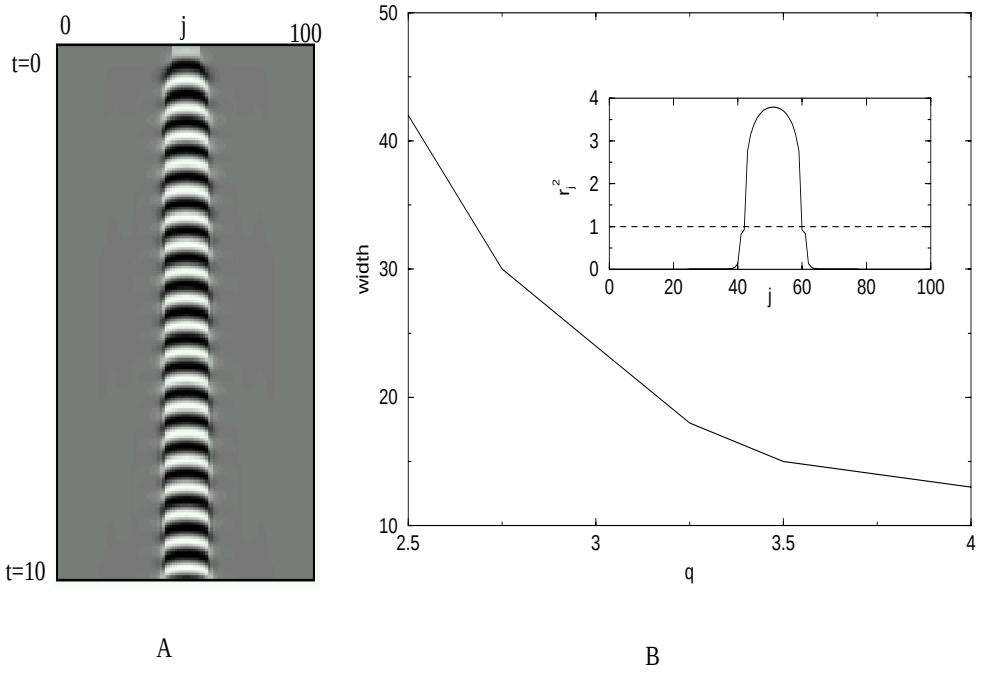


FIG. 5. (A) *Stable localized pulse with* $q = 3.25$, $\lambda = -.2$, $c_1 = 1$. *Grey scale indicates magnitude of* $u_j(t)$. *Index is horizontal, and time runs vertically from* 0 *to* 10. (B) *Pulse width versus* q. *Inset shows* r_j^2, *which appears to be stationary. The horizontal line at* 1 *shows the points at which the width is measured.*

This makes the pulses here quite different from those described in models for working memory (see the discussion and the next section), which have aperiodic behavior. As the parameter q decreases, the width of the pulses gets larger (Figure 5(B)). We conjecture that the width goes to infinity as $|q|$ decreases to q_∞. Figure 3 indicates that the velocity of the fronts goes to zero as q goes to a critical value, q_0. We conjecture that $q_\infty = q_0$, and for the present system that this critical value of q is around 2.5. The reason for this is as follows. Suppose that $q < q_\infty$. Then there are no finite-width pulses. That is, an initial stimulus in the middle of the medium will expand without bound. This is just a pair of wavefronts propagating outward. Similarly, if $q > q_0$, then there are no waves with a positive velocity; we expect that localized sufficiently large initial data will persist and not propagate.

3.5.3. Interactions of pulses. It is possible to initiate multiple pulses in the same medium; however, the behavior is quite dependent on the initial distance as well as the relative phases of the two initial conditions. For example, it is possible for two pulses to merge and form a single pulse, or they can merge and then expand to fill the medium. In the following simulations, the medium is started at rest except for two local regions in which u_j is either 1 or -1. When $u_j = 1$ for both regions, we call this an "in-phase" initial condition, and when $u_j = -1$ we call it "out-of-phase." Figure 6 shows some examples. In (A), a pair of in-phase initial data merge and then chaotically fill the medium. This suggests that the appearance of pulses and the steady-state behavior could depend on the amount of medium initially excited. Indeed, if we excite successively larger parts of the medium, there is a transition from

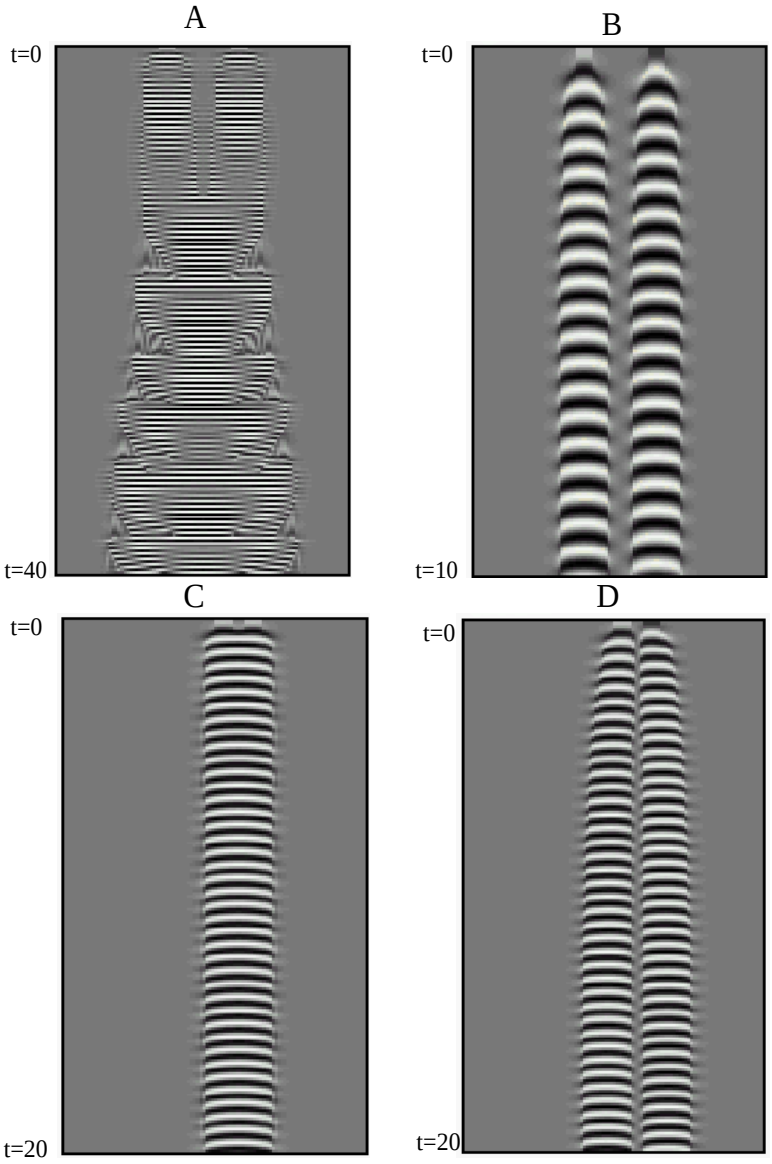


FIG. 6. *Interactions between pulses.* (A) *Two initiated in-phase lead to chaotic behavior;* (B) *same as* (A) *but initially out of phase;* (C) *same as* (A) *but started closer together, leading to merging;* (D) *two out-of-phase pulses split apart. All figures have* $q = 3.25, c_1 = 1, \lambda = -0.2$.

spatially localized behavior to chaotically expanding behavior. In Figure 6(B), the same initial data are given as in A, but the two are out of phase. This results in a pair of local pulses oscillating exactly a half-cycle out of phase. In Figure 6(C), an in-phase pair is started close to each other, leading to simple merging to a single pulse. Finally in (D), the same initial data as in (C) but in antiphase results in a splitting apart of the two pulses rather than a merging. There are many other aspects of interaction which remain to be explored. In particular, the transition between localized pulses and slow chaotic spreading is an intriguing problem.

4. Conductance-based models. The results described in the previous section apply to a very special system of equations, namely, the normal form for a Hopf bifurcation. Thus, a natural question to ask is whether a system away from the bifurcation can have the same behavior. In this section, we consider the behavior of a simple biophysical model with synaptic coupling. The Morris–Lecar (ML) model [21] is a membrane model with leak, calcium, and potassium currents. With the right choice of parameters, the isolated model undergoes a subcritical Hopf bifurcation and has a small regime of bistability between a resting state and a periodic solution. Thus, we will synaptically couple an array of ML neurons and, by altering the applied current, show that the network is able to produce both traveling fronts joining a fixed point to a periodic orbit and localized regions of activity. The equations for each cell are

$$\begin{aligned} C\frac{dV}{dt} &= I - g_l(V - E_l) - g_{Ca}m_\infty(V)(V - E_{Ca}) - g_K n(V - E_K) - I_{syn}, \\ \frac{dn}{dt} &= \frac{n_\infty(V) - n}{\tau_n(V)}, \\ \frac{ds}{dt} &= \alpha(V)(1 - s) - \frac{s}{\tau}, \end{aligned} \tag{4.1}$$

where I_{syn} is the total synaptic current applied to the ith neuron:

$$I_{syn,i} = \left(\sum_j W(i-j)s_j \right) (V_i - E_{syn}).$$

The functions and parameters used are in the appendix. Basically the g's are maximal conductances, the E's are reversal potentials, and $W(j)$ is the coupling strength between neurons and decays with distance. In the normal form, we are able to alter certain abstract parameters such as the imaginary part of the coupling and the nonlinear frequency parameter q. In the actual model, there is no direct analogue of these parameters. However, we can instead alter the applied current I to take the system into and out of the regime of bistability. In Figure 7, we choose I so that the network is near the onset of spontaneous periodicity but remains bistable. A shock at the left of the medium results in a propagating wave shown on the left. Decreasing the current (and making the network less excitable) results in a pulse.

Unlike in the simplified model, the pulse does not seem to be periodic. Rather, it is aperiodic, as is often the case for conductance-based models (see, e.g., [18]).

5. Discussion. We have used a simple canonical model of a bistable oscillatory system to study the propagation of periodic waves and the loss of these waves as either the threshold (λ) changes or the "twist"(q) varies. We have also studied the stability of plane waves in this system. In previous work, we have shown that diffusive coupling in a bistable (oscillatory and fixed point) system leads to a unique traveling front [9]. We also showed that as the threshold changed, the waves disappeared, leaving in their wake spatial patterns. Similar behavior is found in the present model. Unlike the results in [9], we have no closed form for the traveling waves. The existence of wavefronts for sufficiently small values of q remains an open problem; we have proved it only for $q = 0$.

The physiological motivation for this work comes from the behavior of disinhibited slices of cortical tissue [11, 5]. Shocking the tissue results in the propagation of a front

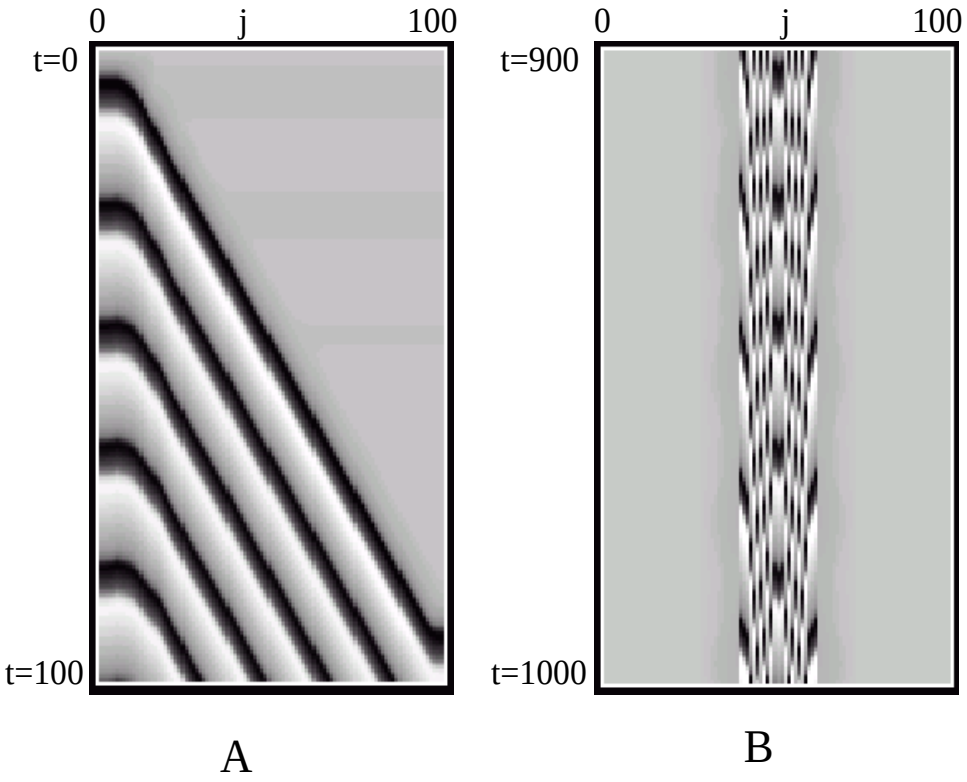


FIG. 7. *Evolution of the voltage for a network of* 100 *ML neurons. The horizontal axis is cell number, and the vertical axis is time in milliseconds. White is a potential of* -45 *mV, and black is a potential of* 20 *mV.* (A) *Traveling wave for* $I = 88$. (B) *Localized pulse for* $I = 82$.

of activity. In experiments the activity eventually terminates due to additional slow processes so that, rather than a front, one obtains a traveling pulse with a finite number of oscillations within the envelope. The present model can be augmented by the addition of a slow negative feedback term, which will terminate the activity, resulting in a spatially confined propagating pulse.

Stationary patterns of localized activity in networks of spiking models have been suggested as models for locally persistent neural activity known as working memory [18, 22]. In these models, the mechanism depends on recurrent excitation of the neurons coupled with *lateral inhibition*. In the context of these models, lateral inhibition means that the connection function $J(x)$ is positive for small $|x|$ and negative for $|x|$ sufficiently large. The present model provides another mechanism which does not depend on lateral inhibition. Rather it depends on bistability between an oscillatory and a rest state. [3] utilized bistability in a firing rate model to obtain localized structures, but, as in the above-mentioned models, they require lateral inhibition. The reason for this can be clarified by looking at our model without the oscillatory component:

$$r_t = r(\lambda + br^2 - r^4) + c\int_{-\infty}^{\infty} J(x-y)r^3(y,t)\ dy.$$

The results of [4], while not precluding localized structures, indicate that in the bistable case, the main type of behavior observed will be stable traveling fronts.

Thus, even though the system is bistable, localized pulses cannot be found with positive $J(x)$. However, the presence of oscillations enables local phase gradients to develop, which can prevent the expansion of the wavefronts beyond a certain range. The mechanism applies to whole classes of voltage-dependent models in which there is a subcritical Hopf bifurcation as current is applied to the system (so-called Type II excitability; see [21]).

It still remains to rigorously prove the existence of these stationary solutions. These satisfy

$$0 = R(\lambda + b_1 R^2 - R^4) + c_1 \int_{-\infty}^{\infty} J(x-y) R^3(y) \cos[\Theta(y) - \Theta(x)]\ dy,$$
$$R(x)\Omega = qR^3(x) + c_1 \int_{-\infty}^{\infty} J(x-y) R^3(y) \sin[\Theta(y) - \Theta(x)]\ dy,$$

where Ω is an unknown parameter, $R(\pm\infty) \to 0$, and $\Theta(0) = 0$. In the case in which $J(x) = \exp(-|x|)/2$, one can then convert the integral equations to a set of differential algebraic equations as follows. Let

$$C(x) = \int_{-\infty}^{\infty} J(x-y) R^3(y) \cos(\Theta(y))\ dy,$$
$$S(x) = \int_{-\infty}^{\infty} J(x-y) R^3(y) \sin(\Theta(y))\ dy,$$

so that we must solve

$$C - C_{xx} = R^3(x) \cos\Theta(x), \qquad S - S_{xx} = R^3(x) \sin\Theta(x),$$

with the constraints

$$0 = R(\lambda + b_1 R^2 - R^4) + c_1 \left(\cos\Theta(x) C(x) + \sin\Theta(x) S(x)\right),$$
$$R\Omega = qR^3 + c_1 \left(\cos\Theta(x) S(x) - \sin\Theta(x) C(x)\right).$$

We have made little progress on this open problem.

Appendix. In the ML equations (4.1), V denotes membrane potential, C is membrane capacitance, and m and h are gating variables. The g's and E_j's are the maximal conductances and reversal potentials, respectively, for calcium, potassium, and leak currents. The gating functions are as follows:

$$n_\infty(v) = .5\left(1 + \tanh\left(\frac{v - v_3}{v_4}\right)\right),$$

$$m_\infty(v) = .5\left(1 + \tanh\left(\frac{v - v_1}{v_2}\right)\right),$$

$$k(v) = \frac{1}{1 + \exp\left(-\dfrac{v - v_t}{v_s}\right)},$$

$$\tau_n(v) = \frac{1}{\cosh\left(\dfrac{v - v_3}{2v_4}\right)},$$

where $v_t = 10$, $v_s = 5$, $v_1 = -1.2$, $v_2 = 18$, $v_3 = 2$, and $v_4 = 30$.

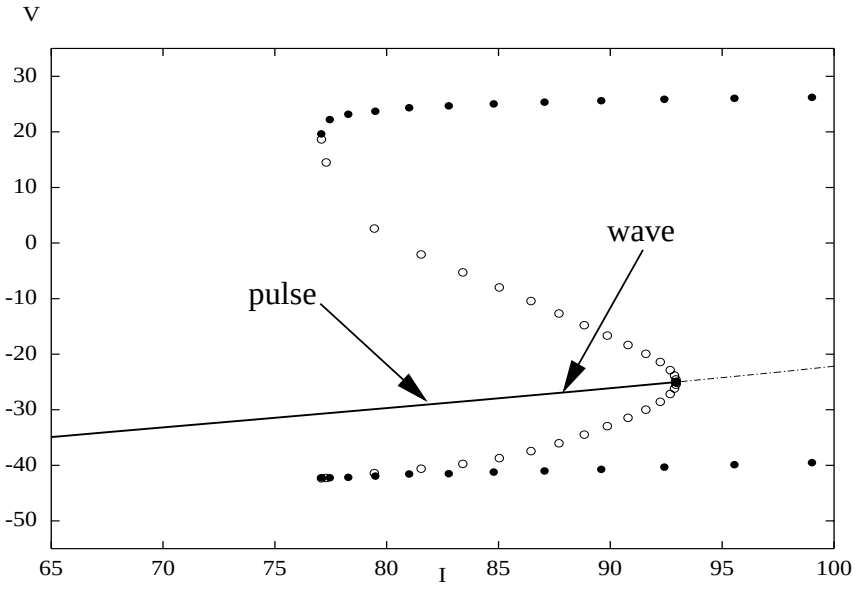


FIG. 8. *Bifurcation diagram obtained by varying I. Arrows denote the values of the current used for the wave and pulse shown in Figure 7.*

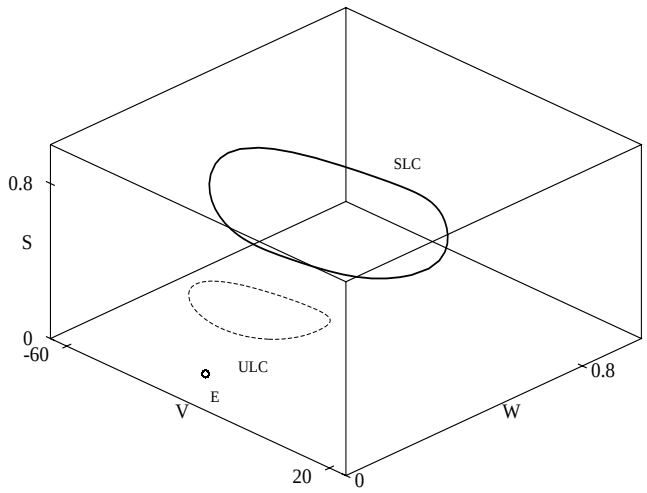


FIG. 9. *In this figure, ULC is the unstable periodic orbit separating the stable fixed point (E) and the stable periodic solution (SLC) in three dimensions (V, w, s) for the ML model.*

The parameters for the ML equations (4.1) for Figure 7 are $\phi = .16$, $g_l = 2$, $g_{ca} = 4.4$, $g_k = 8$, $E_K = -84$, $E_L = -60$, $E_{Ca} = 120$, $E_{syn} = 0$, $\tau_{syn} = 50$, $\alpha = 1$, $g_{syn} = 0.3$, and $C = 5$. With these parameters, the Hopf bifurcation that causes the bistability occurs slightly to the right of $I = 93$. The current used for the traveling waves is $I = 88$ and for the pulses, $I = 82$. These are indicated in the bifurcation diagram. From the bifurcation diagram in Figure 8, at the chosen values for I there

is a stable rest state and a stable periodic, separated by an unstable separatrix. A picture of these orbits is shown in Figure 9. It is not easy in three dimensions to delineate the basins of attraction for the two stable orbits; however, we have found that low values of s and (V, w) near the fixed point are pulled into the rest state, while all other initial data are attracted to the limit cycle.

From the bifurcation diagram, it is possible to estimate the parameters $\lambda = 11.07$, $b = 0.56$, $q = 2.47$ for the model. (Note that we have converted the timescale from milliseconds to seconds, since frequencies are typically measured in Hz.)

REFERENCES

[1] D.G. Aronson, G.B. Ermentrout, and N. Kopell, *Amplitude response of coupled oscillators*, Phys. D, 41 (1990), pp. 403–449.

[2] P.C. Bressloff and S. Coombes, *A dynamical theory of spike train transitions in networks of integrate-and-fire oscillators*, SIAM J. Appl. Math., 60 (2000), pp. 820–841.

[3] M. Camperi and X.-J. Wang, *A model of visuospatial short-term memory in prefrontal cortex: Cellular bistability and recurrent network*, J. Comput. Neurosci., 5 (1998), pp. 383–405.

[4] X. Chen, *Existence, uniqueness, and asymptotic stability of traveling waves in nonlocal evolution equations*, Adv. Differential Equations, 2 (1997), pp. 125–160.

[5] Z. Chen and B. Ermentrout, *Wave propagation mediated by $GABA_B$ synapse and rebound excitation in an inhibitory network: A reduced model approach*, J. Comput. Neurosci., 5 (1998), pp. 53–69.

[6] E. Doedel, *AUTO: A program for the automatic bifurcation analysis of autonomous systems*, Congr. Numer., 30 (1981), pp. 265–284.

[7] G.B. Ermentrout, *Stable small amplitude solutions in reaction-diffusion systems*, Quart. Appl. Math, 39 (1981), pp. 61–86.

[8] B. Ermentrout, *Type* I *membranes, phase resetting curves, and synchrony*, Neural Comput., 8 (1996), pp. 979–1001.

[9] B. Ermentrout, X. Chen, and Z. Chen, *Transition fronts and localized structures in bistable reaction-diffusion equations*, Phys. D, 108 (1997), pp. 147–167.

[10] B. Ermentrout and M. Lewis, *Pattern formation in systems with one spatially distributed species*, Bull. Math. Biol., 59 (1997), pp. 533–549.

[11] D. Golomb and Y. Amitai, *Propagating neuronal discharges in neocortical slices: Computational and experimental study*, J. Neurophysiol., 78 (1997), pp. 1199–1211.

[12] M. Golubitsky and D.G. Schaeffer, *Singularities and Groups in Bifurcation Theory*, Vol. I, Appl. Math. Sci. 51, Springer-Verlag, New York, 1985.

[13] F.C. Hoppensteadt and E.M. Izhikevich, *Weakly Connected Neural Networks*, Springer-Verlag, New York, 1997.

[14] E.M. Izhikevich, *Class* 1 *neural excitability, conventional synapses, weakly connected networks, and mathematical foundations of pulse-coupled models*, IEEE Trans. Neural Networks, 10 (1999), pp. 499–507.

[15] E.M. Izhikevich, *Neural excitability, spiking, and bursting*, Internat. J. Bifur. Chaos, 10 (2000), pp. 1171–1266.

[16] E.M. Izhikevich, *Subcritical elliptic bursting of Bautin type*, SIAM J. Appl. Math., 60 (2000), pp. 503–535.

[17] Yu. A. Kuznetsov, *Elements of Applied Bifurcation Theory*, Springer-Verlag, New York, 1998.

[18] C. Laing and C.C. Chow, *Stationary bumps in networks of spiking neurons*, Neural Comput., 13 (2001), pp. 1473–94.

[19] Y. Loewenstein and H. Sompolinsky, *Oscillations by symmetry breaking in homogeneous networks with electrical coupling*, Phys. Rev. E, 65 (2002), paper 51926.

[20] M.I. Rabinovich, A.B. Ezersky, and P.D. Weidman, *The Dynamics of Patterns*, World Scientific, London, 2000.

[21] J. Rinzel and B. Ermentrout, *Analysis of neural excitability and oscillations*, in Methods in Neuronal Modeling: From Ions to Networks, C. Koch and I. Segev, eds., MIT Press, Cambridge, MA, 1998, Chapter 7, pp. 251–291.

[22] X.-J. Wang, *Synaptic basis of cortical persistent activity: The importance of NMDA receptors to working memory*, J. Neurosci., 19 (1999), pp. 9587–9603.

[23] S. Wiggins, *Introduction to Applied Nonlinear Dynamical Systems and Chaos*, Springer-Verlag, New York, 1990.

SIAM J. APPL. MATH.
Vol. 63, No. 5, pp. 1648–1665

SINGULAR PERTURBATIONS IN OPTION PRICING*

J.-P. FOUQUE†, G. PAPANICOLAOU‡, R. SIRCAR§, AND K. SOLNA¶

Abstract. After the celebrated Black–Scholes formula for pricing call options under constant volatility, the need for more general nonconstant volatility models in financial mathematics motivated numerous works during the 1980s and 1990s. In particular, a lot of attention has been paid to stochastic volatility models in which the volatility is randomly fluctuating driven by an additional Brownian motion. We have shown in [*Derivatives in Financial Markets with Stochastic Volatility*, Cambridge University Press, Cambridge, UK, 2000; *Internat. J. Theoret. Appl. Finance*, 13 (2000), pp. 101–142] that, in the presence of a separation of time scales between the main observed process and the volatility driving process, asymptotic methods are very efficient in capturing the effects of random volatility in simple robust corrections to constant volatility formulas. From the point of view of PDEs, this method corresponds to a singular perturbation analysis. The aim of this paper is to deal with the nonsmoothness of the payoff function inherent to option pricing. We present the case of call options for which the payoff function forms an angle at the strike price. This case is important since these are the typical instruments used in the calibration of pricing models. We establish the pointwise accuracy of the corrected Black–Scholes price by using an appropriate payoff regularization which is removed simultaneously as the asymptotics is performed.

Key words. mathematical finance, option pricing, stochastic volatility, singular perturbations

AMS subject classifications. 60G15, 60G44, 60H15, 60J60, 91B28

DOI. 10.1137/S0036139902401550

1. Introduction.

Stochastic volatility models in financial mathematics can be thought of as a Brownian-type particle (the stock price) moving in an environment where the diffusion coefficient is randomly fluctuating in time according to some ergodic (mean-reverting) diffusion process. We then have two Brownian motions, one driving the motion of the particle and the other driving the fluctuations of the medium. In the context of physics there is no natural correlation between these two Brownian motions since they do not "live" in the same space. In the context of finance they jointly define the dynamics of the stock price under its physical probability measure or an equivalent risk-neutral martingale measure. Correlation between them is perfectly natural. There are economic arguments for a negative correlation or *leverage effect* between stock price and volatility shocks, and from common experience and empirical studies we know that asset prices tend to go down when volatility goes up. The diffusion equation appears as a contingent claim pricing equation, its terminal condition being the payoff of the claim. We refer to [5] or [6] for surveys on stochastic volatility. When volatility is fast mean-reverting, on a timescale smaller than typical maturities, one can perform a singular perturbation analysis of the pricing PDE. As we have shown in [2], this expansion reveals a first correction made of two terms: one is directly associated with the market price of volatility risk, and the other is propor-

*Received by the editors January 28, 2002; accepted for publication (in revised form) January 22, 2003; published electronically July 26, 2003.
http://www.siam.org/journals/siap/63-5/40155.html

†Department of Mathematics, North Carolina State University, Raleigh, NC 27695-8205 (fouque@math.ncsu.edu). The research of this author was partially supported by NSF grant DMS-0071744.

‡Department of Mathematics, Stanford University, Stanford, CA 94305 (papanico@math.stanford.edu).

§Department of Operations Research and Financial Engineering, Princeton University, E-Quad, Princeton, NJ 08544 (sircar@princeton.edu). The research of this author was partially supported by NSF grant DMS-0090067.

¶Department of Mathematics, University of California, Irvine, CA 92697 (ksolna@math.uci.edu).

tional to the correlation coefficient between the two Brownian motions involved. We refer to [2] for a detailed account of evidence of a fast scale in volatility and the use of this asymptotics to parametrize the evolution of the *skew* or the implied volatility surface. We also refer to [4] for a different type of application, namely, variance reduction in Monte Carlo methods.

The present paper deals with the accuracy of such an expansion in the presence of another essential characteristic feature in option pricing, namely, the nonsmoothness of payoff functions. We present the case of call options since these are the liquid instruments used in the calibration of pricing models. By inverting the Black–Scholes formula, the price of a call option is given in terms of its implied volatility, which depends on the strike and the maturity of the option. This set of implied volatilities form the *term structure of implied volatility*. For fixed maturity and across strikes it is known as the *smile* or the *skew* due to the observed asymmetry. These objects and their dynamics are what volatility models are trying to reproduce in order to price and hedge other instruments.

In [2] we have performed an expansion of the price in powers of the characteristic mean-reversion time of volatility, and we have shown that the leading order term corresponds to a Black–Scholes price computed under a constant effective volatility. The first correction involves derivatives of this constant volatility price. When the payoff is smooth, we have shown that the corrected price, leading order term plus first correction, has the expected accuracy; namely, the remainder of the expansion is of the next order. The nonsmoothness of a call payoff which forms an angle at the strike price creates a singularity at the maturity time near the strike price of the option.

This paper is devoted to the proof of the accuracy of the approximation in that case. It is important because this is a natural situation in financial mathematics that one has to deal with. The proof given here relies on a payoff smoothing argument, which can certainly be useful in other contexts.

In section 2 we introduce the class of stochastic volatility models which we consider. They are written directly under the pricing equivalent martingale measure and with a small parameter representing the short timescale of volatility. We recall how option prices are given as expected values of discounted payoffs or as solutions of pricing backward parabolic PDEs with terminal conditions at maturity times. In section 3 we recall the formal asymptotic expansion presented in [2]. In section 4 we introduce the regularization of the payoff and decompose the main result, accuracy of the price approximation, into three lemmas. Section 5 is devoted to the proof of these lemmas. Detailed computations involving derivatives of Black–Scholes prices up to order seven are given in the appendices, where we also recall the properties of the solutions of Poisson equations associated with the infinitesimal generator of the Ornstein-Uhlenbeck process driving the volatility.

2. Class of models and pricing equations. The family of Ornstein–Uhlenbeck (OU) driven stochastic volatility models $(S_t^\varepsilon, Y_t^\varepsilon)$ that we consider can be written, under a risk-neutral probability $\mathbb{P}^\star$, in terms of the small parameter ε,

$$\begin{aligned} dS_t^\varepsilon &= rS_t^\varepsilon dt + f(Y_t^\varepsilon) S_t^\varepsilon dW_t^\star, \\ dY_t^\varepsilon &= \left[\frac{1}{\varepsilon}(m - Y_t^\varepsilon) - \frac{\nu\sqrt{2}}{\sqrt{\varepsilon}}\Lambda(Y_t^\varepsilon)\right] dt + \frac{\nu\sqrt{2}}{\sqrt{\varepsilon}} d\hat{Z}_t^\star, \end{aligned}$$

where the Brownian motions $(W_t^\star, \hat{Z}_t^\star)$ have instantaneous correlation $\rho \in (-1, 1)$,

$$d\langle W^\star, \hat{Z}^\star\rangle_t = \mathbb{E}^\star\{dW_t^\star \, d\hat{Z}_t^\star\} = \rho \, dt$$

and

$$\Lambda(y) = \frac{\rho(\mu - r)}{f(y)} + \gamma(y)\sqrt{1-\rho^2}$$

is a combined market price of risk. It describes the relationship between the physical measure under which the stock price is observed and the risk-neutral measure under which the market prices derivative securities; see [2], for example. The price of the underlying stock is S_t^ε, and the volatility is a function f of the process Y_t^ε. At the leading order $1/\varepsilon$, that is, omitting the Λ-term, Y_t^ε is an OU process that is fast mean-reverting with a normal invariant distribution $\mathcal{N}(m, \nu^2)$. Notice that in this framework the volatility driving process (Y_t^ε) is autonomous in the sense that the coefficients in its defining SDE do not depend on the stock price S_t^ε.

In this *fast mean-reverting* stochastic volatility scenario, the volatility level fluctuates randomly around its mean level, and the epochs of high/low volatility are relatively short. This is the regime that we consider and under which we analyze the price of European derivatives. A derivative is defined by its nonnegative payoff function $H(S)$, which prescribes the value of the contract at its maturity time T when the stock price is S. The payoff function must in general satisfy the integrability condition

$$\mathbb{E}^\star\{H(S_T)^2\} < \infty,$$

with $\mathbb{E}^\star$ denoting expectation with respect to $\mathbb{P}^\star$. Moreover, we *assume* the following:

1. The volatility is positive and bounded: there are constants m_1 and m_2 such that

$$0 < m_1 \le f(y) \le m_2 < \infty \qquad \forall y \in \mathcal{R}.$$

2. The volatility risk-premium is bounded:

$$|\gamma(y)| < l < \infty \qquad \forall y \in \mathcal{R}$$

for some constant l.

It is convenient at this stage to make the change of variable

$$X_t^\varepsilon = \log S_t^\varepsilon, \qquad t \ge 0,$$

and write the problem in terms of the processes $(X_t^\varepsilon, Y_t^\varepsilon)$, which satisfy, by Itô's formula, the stochastic differential equations

$$dX_t^\varepsilon = \left(r - \frac{1}{2} f(Y_t^\varepsilon)^2\right) dt + f(Y_t^\varepsilon)\, dW_t^\star, \tag{2.1}$$

$$dY_t^\varepsilon = \left[\frac{1}{\varepsilon}(m - Y_t^\varepsilon) - \frac{\nu\sqrt{2}}{\sqrt{\varepsilon}}\Lambda(Y_t^\varepsilon)\right] dt + \frac{\nu\sqrt{2}}{\sqrt{\varepsilon}}\, d\hat{Z}_t^\star. \tag{2.2}$$

We also define the payoff function h in terms of the log stock price via

$$H(e^x) = h(x), \qquad x \in \mathcal{R}.$$

The price at time $t < T$ of this derivative is a function of the present value of the stock price, or equivalently the log stock price, $X_t^\varepsilon = x$ and the present value $Y_t^\varepsilon = y$ of the process driving the volatility. We denote this price by $P^\varepsilon(t, x, y)$. It is standard

in finance to assume that the price is given by (2.3), which is the expected discounted payoff under the risk-neutral probability measure $\mathbb{P}^\star$. See [1], for example.

$$P^\varepsilon(t,x,y) = \mathbb{E}^\star\left\{e^{-r(T-t)}h(X_T^\varepsilon)|X_t^\varepsilon = x, Y_t^\varepsilon = y\right\}. \tag{2.3}$$

We shall also write these conditional expectations more compactly as

$$P^\varepsilon(t,x,y) = \mathbb{E}^\star{}_{t,x,y}\left\{e^{-r(T-t)}h(X_T^\varepsilon)\right\}.$$

Under the assumptions on the models considered and the payoff, $P^\varepsilon(t,x,y)$ is the unique classical solution to the associated backward Kolmogorov PDE problem

$$\begin{aligned}\mathcal{L}^\varepsilon P^\varepsilon &= 0,\\ P^\varepsilon(T,x,y) &= h(x)\end{aligned} \tag{2.4}$$

in $t<T$, $x,y\in\mathcal{R}$, where we have defined the operators

$$\mathcal{L}^\varepsilon = \frac{1}{\varepsilon}\mathcal{L}_0 + \frac{1}{\sqrt{\varepsilon}}\mathcal{L}_1 + \mathcal{L}_2,$$

$$\mathcal{L}_0 = \nu^2\frac{\partial^2}{\partial y^2} + (m-y)\frac{\partial}{\partial y}, \tag{2.5}$$

$$\mathcal{L}_1 = \sqrt{2}\rho\nu f(y)\frac{\partial^2}{\partial x\partial y} - \sqrt{2}\nu\Lambda(y)\frac{\partial}{\partial y}, \tag{2.6}$$

$$\mathcal{L}_2 = \frac{\partial}{\partial t} + \frac{1}{2}f(y)^2\frac{\partial^2}{\partial x^2} + \left(r - \frac{1}{2}f(y)^2\right)\frac{\partial}{\partial x} - r\cdot. \tag{2.7}$$

The operator $\mathcal{L}_0$ is the infinitesimal generator of the OU process (Y_t) defined by

$$dY_t = (m - Y_t)\,dt + \nu\sqrt{2}\,d\hat{Z}_t^\star; \tag{2.8}$$

$\mathcal{L}_1$ contains the mixed partial derivative due to the correlation and the derivative due to the market price of risk, and $\mathcal{L}_2$, also denoted by $\mathcal{L}_{BS}(f(y))$, is the Black–Scholes operator in the log variable and with volatility $f(y)$.

3. Price approximation. We present here the formal asymptotic expansion computed as in [2, 3], which leads to a (first-order in $\sqrt{\varepsilon}$) approximation $P^\varepsilon(t,x,y)\approx Q^\varepsilon(t,x)$. In the next section we prove the convergence and accuracy as $\varepsilon\downarrow 0$ of this approximation, which consists of the first two terms of the asymptotic price expansion:

$$Q^\varepsilon(t,x) = P_0(t,x) + \sqrt{\varepsilon}P_1(t,x),$$

which do not depend on y and are derived as follows. We start by writing

$$P^\varepsilon = Q^\varepsilon + \varepsilon Q_2 + \varepsilon^{3/2}Q_3 + \cdots = P_0 + \sqrt{\varepsilon}P_1 + \varepsilon Q_2 + \varepsilon^{3/2}Q_3 + \cdots. \tag{3.1}$$

Substituting (3.1) into (2.4) leads to

$$\begin{aligned}&\frac{1}{\varepsilon}\mathcal{L}_0P_0 + \frac{1}{\sqrt{\varepsilon}}\left(\mathcal{L}_0P_1 + \mathcal{L}_1P_0\right)\\ &+ \left(\mathcal{L}_0Q_2 + \mathcal{L}_1P_1 + \mathcal{L}_2P_0\right) + \sqrt{\varepsilon}\left(\mathcal{L}_0Q_3 + \mathcal{L}_1Q_2 + \mathcal{L}_2P_1\right) + \cdots \;=\; 0.\end{aligned} \tag{3.2}$$

We shall next obtain expressions for P_0 and P_1 by successively equating the four leading order terms in (3.2) to zero. We let $\langle\cdot\rangle$ denote the averaging with respect to the invariant distribution $\mathcal{N}(m,\nu^2)$ of the OU process Y introduced in (2.8):

$$\langle g\rangle = \frac{1}{\nu\sqrt{2\pi}}\int_{\mathcal{R}} g(y)e^{-(m-y)^2/2\nu^2}dy. \tag{3.3}$$

Notice that this averaged quantity does not depend on ε.

Below, we will need to solve the *Poisson equation* associated with $\mathcal{L}_0$,

$$\mathcal{L}_0\chi + g = 0, \tag{3.4}$$

which requires the solvability condition

$$\langle g\rangle = 0 \tag{3.5}$$

in order to admit solutions with reasonable growth at infinity. Properties of this equation and its solutions are recalled in Appendix C.

Consider first the leading order term

$$\mathcal{L}_0 P_0 = 0.$$

Since $\mathcal{L}_0$ takes derivatives with respect to y, any function independent of y satisfies this equation. On the other hand y-dependent solutions exhibit the unreasonable growth $\exp(y^2/2\nu^2)$ at infinity. Therefore we seek solutions which are independent of y: $P_0 = P_0(t,x)$ with the terminal condition $P_0(T,x) = h(x)$.

Consider next

$$\mathcal{L}_0 P_1 + \mathcal{L}_1 P_0 = 0,$$

which corresponds to the second term in (3.2). Since $\mathcal{L}_1$ contains only terms with derivatives in y, it reduces to $\mathcal{L}_0 P_1 = 0$ and, as for P_0, we seek again a function $P_1 = P_1(t,x)$, independent of y, with a zero terminal condition $P_1(T,x) = 0$. Hence, $Q^\varepsilon = P_0 + \sqrt{\varepsilon}P_1$, the leading order approximation, does not depend on the current value of the volatility level.

The next equation

$$\mathcal{L}_0 Q_2 + \mathcal{L}_1 P_1 + \mathcal{L}_2 P_0 = 0,$$

which corresponds to the third term in (3.2), reduces to the Poisson equation

$$\mathcal{L}_0 Q_2 + \mathcal{L}_2 P_0 = 0, \tag{3.6}$$

since $\mathcal{L}_1 P_1 = 0$. Its solvability condition

$$\langle \mathcal{L}_2 P_0\rangle = \langle \mathcal{L}_2\rangle P_0 = 0,$$

is the Black–Scholes PDE (in the log variable) with constant square volatility $\langle f^2\rangle$:

$$\langle \mathcal{L}_2\rangle P_0 = \mathcal{L}_{BS}(\bar\sigma)P_0 = \frac{\partial P_0}{\partial t} + \frac{1}{2}\bar\sigma^2\frac{\partial^2 P_0}{\partial x^2} + \left(r - \frac{1}{2}\bar\sigma^2\right)\frac{\partial P_0}{\partial x} - rP_0 = 0, \tag{3.7}$$

where we define the *effective constant volatility* $\bar\sigma$ by

$$\bar\sigma^2 = \langle f^2\rangle.$$

We choose $P_0(t,x)$ to be the classical Black–Scholes price, solution of (3.7) with the terminal condition $P_0(T,x) = h(x)$.

Observe that $Q_2 = -\mathcal{L}_0^{-1}(\mathcal{L}_2 - \langle\mathcal{L}_2\rangle)P_0$ as a solution of the Poisson equation (3.6). This notation includes an additive constant in y, which will disappear when hit by the operator $\mathcal{L}_1$ below. The fourth term in (3.2) gives the equation

$$\mathcal{L}_0 Q_3 + \mathcal{L}_1 Q_2 + \mathcal{L}_2 P_1 = 0. \tag{3.8}$$

This is a Poisson equation in Q_3, and its solvability condition gives

$$\langle\mathcal{L}_2\rangle P_1 = -\langle\mathcal{L}_1 Q_2\rangle = \langle\mathcal{L}_1\mathcal{L}_0^{-1}(\mathcal{L}_2 - \langle\mathcal{L}_2\rangle)\rangle P_0,$$

which, with its zero terminal condition, determines P_1 as a solution of a Black–Scholes equation with constant square volatility $\langle f^2\rangle$ and a source term. Using the expressions for $\mathcal{L}_i$, one can rewrite the source as

$$\begin{aligned}\langle\mathcal{L}_1\mathcal{L}_0^{-1}(\mathcal{L}_2 - \langle\mathcal{L}_2\rangle)\rangle P_0 &= \left\langle\mathcal{L}_1\mathcal{L}_0^{-1}\left(f(y)^2 - \langle f^2\rangle\right)\right\rangle \frac{1}{2}\left(\frac{\partial^2}{\partial x^2} - \frac{\partial}{\partial x}\right)P_0 \\ &= \left(v_3\frac{\partial^3}{\partial x^3} + (v_2 - 3v_3)\frac{\partial^2}{\partial x^2} + (2v_3 - v_2)\frac{\partial}{\partial x}\right)P_0, \end{aligned} \tag{3.9}$$

where

$$\begin{aligned} v_2 &= \frac{\nu}{\sqrt{2}}(2\rho\langle f\phi'\rangle - \langle\Lambda\phi'\rangle), \\ v_3 &= \frac{\rho\nu}{\sqrt{2}}\langle f\phi'\rangle, \end{aligned} \tag{3.10}$$

and ϕ is a solution of the Poisson equation

$$\mathcal{L}_0\phi(y) = f(y)^2 - \langle f^2\rangle. \tag{3.11}$$

We can therefore conclude the following:

a. The first term P_0 is chosen to be the solution of the "homogenized" PDE problem (3.7). In other words, P_0 is simply the Black–Scholes price of the derivative computed with the effective volatility $\bar\sigma$.

b. The second term, or correction to the Black–Scholes price, is given explicitly as a linear combination of the first three derivatives of P_0, by

$$\sqrt{\varepsilon}\,P_1 = -(T-t)\left(V_3^\varepsilon\frac{\partial^3}{\partial x^3} + (V_2^\varepsilon - 3V_3^\varepsilon)\frac{\partial^2}{\partial x^2} + (2V_3^\varepsilon - V_2^\varepsilon)\frac{\partial}{\partial x}\right)P_0, \tag{3.12}$$

with

$$V_{2,3}^\varepsilon = \sqrt{\varepsilon}\,v_{2,3}, \tag{3.13}$$

since it is easily seen, by using $\langle\mathcal{L}_2\rangle P_0 = 0$, that (3.9) is satisfied and that, on the other hand, the terminal condition $P_1(T,x) = 0$ is satisfied when $\lim_{t\to T}(T-t)\frac{\partial^i P_0}{\partial x^i} = 0$ for $i = 1,2,3$.

Essential instruments in financial markets are put and call options for which the payoff function $H(S)$ is piecewise linear. We shall focus on call options:

$$H(S) = (S-K)^+ \qquad \Rightarrow \qquad h(x) = (e^x - K)^+$$

for some given strike price $K > 0$. Notice that h is only $\mathcal{C}^0$ smooth with a discontinuous first derivative at the kink $x = \log K$ ("at the money" in financial terms). Nonetheless, at $t < T$, the Black–Scholes pricing function $P_0(t, x)$ is smooth and $P_1(t, x)$ is well defined, but second and higher derivatives of P_0 with respect to x blow up as $t \to T$ (at the money).

Our main result on the accuracy of the approximation $Q^\varepsilon = P_0 + \sqrt{\varepsilon}\, P_1$ is as follows.

THEOREM 3.1. *Under the assumptions* 1 *and* 2 *above, at a fixed point* $t < T$, $x, y \in \mathcal{R}$, *the accuracy of the approximation of call prices is given by*

$$\lim_{\varepsilon \downarrow 0} \frac{|P^\varepsilon(t, x, y) - Q^\varepsilon(t, x)|}{\varepsilon |\log \varepsilon|^{1+p}} = 0$$

for any $p > 0$.

Observe that this pointwise approximation is the sense of accuracy needed in finance applications since option prices are computed at given values of (t, x, y).

Before giving in the next section the proof of Theorem 3.1, we comment on the interpretation of the approximation and on the validity of the result for more general payoffs.

Financial interpretation of the approximation. In order to give a meaningful interpretation to the leading order term and the correction in our price approximation it is convenient to return to the variable S, the underlying price. With a slight abuse of notation we denote the call option price approximation by $P_0(t, S) + \sqrt{\varepsilon} P_1(t, S)$. Indeed, the leading order term $P_0(t, S)$ is the standard Black–Scholes price of the call option computed at the effective constant volatility $\bar{\sigma}$. From (3.12), one can easily deduce that

$$\sqrt{\varepsilon}\, P_1(t, S) = -(T - t) \left(V_2^\varepsilon S^2 \frac{\partial^2 P_0}{\partial S^2} + V_3^\varepsilon S^3 \frac{\partial^3 P_0}{\partial S^3} \right), \tag{3.14}$$

which shows that the correction is a combination of the two variables *gamma* and *epsilon*, as introduced in [2]. This correction can alternatively be written in the form

$$\sqrt{\varepsilon}\, P_1(t, S) = -(T - t) \left((V_2^\varepsilon - 2V_3^\varepsilon) S^2 \frac{\partial^2 P_0}{\partial S^2} + V_3^\varepsilon S \frac{\partial}{\partial S} \left(S^2 \frac{\partial^2 P_0}{\partial S^2} \right) \right). \tag{3.15}$$

Using the classical relation between *gamma* and *vega* for Black–Scholes prices of European derivatives

$$\frac{\partial P_0}{\partial \sigma} = (T - t) \sigma S^2 \frac{\partial^2 P_0}{\partial S^2},$$

which is easily obtained by differentiating the Black–Scholes PDE with respect to σ, one can rewrite the correction as

$$\sqrt{\varepsilon}\, P_1(t, S) = -\frac{1}{\bar{\sigma}} \left((V_2^\varepsilon - 2V_3^\varepsilon) \frac{\partial P_0}{\partial \sigma} + V_3^\varepsilon S \frac{\partial}{\partial S} \left(\frac{\partial P_0}{\partial \sigma} \right) \right). \tag{3.16}$$

Therefore the price correction is a combination of the *vega* and the *delta-vega* of the Black–Scholes price. The *vega* term corresponds simply to a volatility level correction. The *delta-vega* term is proportional to the correlation coefficient ρ and captures the main effect of skewness in implied volatility as discussed in detail in [2].

Other payoff functions. The main idea of the proof presented in the next section is a regularization of the payoff, which does not rely on the particular choice of a call option. The only place where we use the explicit Black–Scholes formula for a call option is in the computation (B.1) of the successive derivatives $\partial_x^n P_0^\delta$ carried out in Appendix B. Note that if we had started with a payoff function h which was continuous and piecewise smooth (a call option being a particular case), then P_0^δ, the solution of the parabolic PDE (3.7), would be an integral of the payoff function with respect to a normal density, as in the case of a call option. The first derivative with respect to x can be taken on the payoff function, and the higher-order derivatives can then be taken on the normal density, as detailed in Appendix B for a call option. Therefore Theorem 3.1 remains valid for general European claims with continuous payoffs that have singular behavior in their derivatives.

Numerical illustration. To illustrate the asymptotic approximation, we compare the approximation

$$Q^\varepsilon = P_0 + \sqrt{\varepsilon}\, P_1$$

with a numerical solution of the PDE (2.4) for a particular stochastic volatility model and a call option with strike price $K = 100$ and three months before expiration. (In practice, the asymptotic approximation is not used in this manner because of the difficulties of estimating the volatility parameters precisely; instead the parameters of the approximation V_2^ε and V_3^ε are estimated directly from observed options prices, as described in [2].)

We choose $f(y) = e^y$, where this is understood to stand for a cutoff version of the exponential function, with the cutoffs (above and below) sufficiently large and small, respectively, so as not to affect the calculations within the accuracy of our comparisons. We use the parameter values

$$\varepsilon = \frac{1}{200}, \qquad m = \log 0.1, \qquad \nu = \frac{1}{\sqrt{2}}, \qquad \rho = -0.2,$$

$$\mu = 0.2, \qquad r = 0.04$$

and choose the volatility risk premium $\gamma \equiv 0$. It follows from explicit calculations that the parameters for the asymptotic approximation, are

$$\bar\sigma = 0.165, \qquad V_2^\varepsilon = -3.30 \times 10^{-4}, \qquad V_3^\varepsilon = 8.48 \times 10^{-5}.$$

Figure 1 shows the numerical solution from an implicit finite-difference approximation at two levels of the current volatility e^y, one at the long-run mean-level $\bar\sigma$, and one far above it (0.607). These are compared to the asymptotic approximation, which does not depend on the current volatility level. In the range $0.95 \le K/S \le 1.04$ shown in the graph on the right, the maximum deviation of the asymptotic approximation from the price with the higher volatility is 9% of the latter price, and the maximum deviation of the asymptotic approximation from the price with the lower volatility is 2.1% of this price.

4. Derivation of the accuracy of the price approximation. In order to prove Theorem 3.1, we introduce in the next section the regularized price $P^{\varepsilon,\delta}$, the price of a slightly smoothed call option, with δ being the (small) smoothing parameter. We denote the associated price approximation by $Q^{\varepsilon,\delta}$. The proof then involves showing that (i) $P^\varepsilon \approx P^{\varepsilon,\delta}$, (ii) $Q^{\varepsilon,\delta} \approx Q^\varepsilon$, (iii) $P^{\varepsilon,\delta} \approx Q^{\varepsilon,\delta}$ and controlling the accuracy in these approximations by choosing δ appropriately.

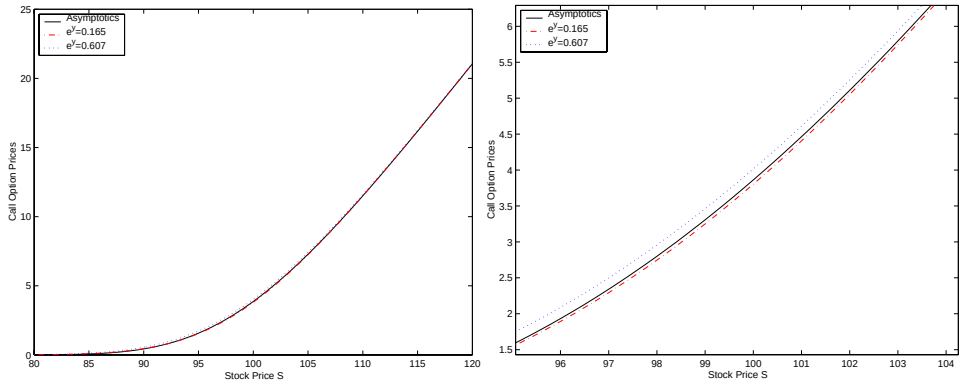


FIG. 1. *Call option prices three months from maturity as a function of the current stock price S. The strike price is $K = 100$, and the graph on the right focuses on the region "around the money."*

4.1. Regularization. We begin by regularizing the payoff, which is a *call option*, by replacing it with the Black–Scholes price of a call with volatility $\bar{\sigma}$ and time to maturity δ. We define

$$h^\delta(x) := C_{BS}(T-\delta, x; K, T; \bar{\sigma}),$$

where $C_{BS}(t, x; K, T; \bar{\sigma})$ denotes the Black–Scholes call option price as a function of current time t, log stock price x, strike price K, expiration date T, and volatility $\bar{\sigma}$. It is given by

$$\begin{aligned} C_{BS}(t,x;K,T;\bar{\sigma}) = P_0(t,x;K,T;\bar{\sigma}) &= e^x N(d_1) - K e^{-r\frac{\tau^2}{\bar{\sigma}^2}} N(d_2), \\ N(x) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2}\, dy, \\ d_1 &= \frac{x - \log K}{\tau} + b\tau, \\ d_2 &= d_1 - \tau, \end{aligned} \tag{4.1}$$

where we define

$$\tau = \bar{\sigma}\sqrt{T-t}, \qquad b = \frac{r}{\bar{\sigma}^2} + \frac{1}{2}.$$

For $\delta > 0$, this new payoff is $\mathcal{C}^\infty$. The price $P^{\varepsilon,\delta}(t,x,y)$ of the option with the regularized payoff solves

$$\begin{aligned} \mathcal{L}^\varepsilon P^{\varepsilon,\delta} &= 0, \\ P^{\varepsilon,\delta}(T,x,y) &= h^\delta(x). \end{aligned}$$

4.2. Main convergence result. Let $Q^{\varepsilon,\delta}(t,x)$ denote the first-order approximation to the regularized option price:

$$P^{\varepsilon,\delta} \approx Q^{\varepsilon,\delta} \equiv P_0^\delta + \sqrt{\varepsilon} P_1^\delta,$$

where

$$P_0^\delta(t,x) = C_{BS}(t-\delta, x; K, T; \bar{\sigma}), \tag{4.2}$$

$$\sqrt{\varepsilon} P_1^\delta = -(T-t)\left(V_3^\varepsilon \frac{\partial^3}{\partial x^3} + (V_2^\varepsilon - 3V_3^\varepsilon)\frac{\partial^2}{\partial x^2} + (2V_3^\varepsilon - V_2^\varepsilon)\frac{\partial}{\partial x} \right) P_0^\delta. \tag{4.3}$$

We establish the following pathway to proving Theorem 3.1, where constants may depend on (t, T, x, y) but not on (ε, δ).

LEMMA 4.1. *Fix the point (t, x, y), where $t < T$. There exist constants $\bar{\delta}_1 > 0$, $\bar{\varepsilon}_1 > 0$, and $c_1 > 0$ such that*

$$|P^{\varepsilon}(t, x, y) - P^{\varepsilon,\delta}(t, x, y)| \leq c_1 \delta$$

for all $0 < \delta < \bar{\delta}_1$ and $0 < \varepsilon < \bar{\varepsilon}_1$.

This establishes that the solutions to the regularized and unregularized problems are close.

LEMMA 4.2. *Fix the point (t, x, y), where $t < T$. There exist constants $\bar{\delta}_2 > 0$, $\bar{\varepsilon}_2 > 0$, and $c_2 > 0$ such that*

$$|Q^{\varepsilon}(t, x) - Q^{\varepsilon,\delta}(t, x)| \leq c_2 \delta$$

for all $0 < \delta < \bar{\delta}_2$ and $0 < \varepsilon < \bar{\varepsilon}_2$.

This establishes that the first-order asymptotic approximations to the regularized and unregularized problems are close.

LEMMA 4.3. *Fix the point (t, x, y), where $t < T$. There exist constants $\bar{\delta}_3 > 0$, $\bar{\varepsilon}_3 > 0$, and $c_3 > 0$ such that*

$$|P^{\varepsilon,\delta}(t, x, y) - Q^{\varepsilon,\delta}(t, x)| \leq c_3 \left(\varepsilon |\log \delta| + \varepsilon \sqrt{\frac{\varepsilon}{\delta}} + \varepsilon \right)$$

for all $0 < \delta < \bar{\delta}_3$ and $0 < \varepsilon < \bar{\varepsilon}_3$.

This establishes that for fixed δ the approximation to the regularized problem converges to the regularized price as $\varepsilon \downarrow 0$.

The convergence result proceeds from these lemmas as follows.

Proof of Theorem 3.1. Take $\bar{\delta} = \min(\bar{\delta}_1, \bar{\delta}_2, \bar{\delta}_3)$ and $\bar{\varepsilon} = \min(\bar{\varepsilon}_1, \bar{\varepsilon}_2, \bar{\varepsilon}_3)$. Then using Lemmas 4.1, 4.2, and 4.3, we obtain

$$\begin{aligned} |P^{\varepsilon} - Q^{\varepsilon}| &\leq |P^{\varepsilon} - P^{\varepsilon,\delta}| + |P^{\varepsilon,\delta} - Q^{\varepsilon,\delta}| + |Q^{\varepsilon,\delta} - Q^{\varepsilon}| \\ &\leq 2 \max(c_1, c_2) \delta + c_3 \left(\varepsilon |\log \delta| + \varepsilon \sqrt{\frac{\varepsilon}{\delta}} + \varepsilon \right) \end{aligned}$$

for $0 < \delta < \bar{\delta}$ and $0 < \varepsilon < \bar{\varepsilon}$, where the functions are evaluated at the fixed (t, x, y). Taking $\delta = \varepsilon$, we have

$$|P^{\varepsilon} - Q^{\varepsilon}| \leq c_5 (\varepsilon + \varepsilon |\log \varepsilon|)$$

for some fixed $c_5 > 0$, and Theorem 3.1 follows.

A general conclusion to our work is given in section 6 after the proofs of Lemmas 4.1, 4.2, and 4.3 given in the following section.

5. Proof of lemmas.

5.1. Proof of Lemma 4.1. We use the probabilistic representation of the price given as the expected discounted payoff with respect to the risk-neutral pricing equivalent martingale measure $\mathbb{P}^{\star}$:

$$P^{\varepsilon,\delta}(t, x, y) = \mathbb{E}^{\star}{}_{t,x,y} \left\{ e^{-r(T-t)} h^{\delta}(X_T^{\varepsilon}) \right\}.$$

We define the new process $(\tilde{X}_t^\varepsilon)$ by

$$d\tilde{X}_t^\varepsilon = \left(r - \frac{1}{2}\tilde{f}(t, Y_t^\varepsilon)^2\right) dt + \tilde{f}(t, Y_t^\varepsilon)\left(\sqrt{1-\rho^2}\, d\hat{W}_t^\star + \rho\, d\hat{Z}_t^\star\right),$$

where $(\hat{W}_t^\star)$ is a Brownian motion independent of $(\hat{Z}_t^\star)$, (Y_t^ε) is still a solution of (2.2), and

$$\tilde{f}(t,y) = \begin{cases} f(y) & \text{for } t \le T, \\ \bar{\sigma} & \text{for } t > T. \end{cases}$$

Then we can write

$$P^{\varepsilon,\delta}(t,x,y) = \mathbb{E}^\star{}_{t,x,y}\left\{e^{-r(T-t+\delta)} h(\tilde{X}^\varepsilon_{T+\delta})\right\}$$

and

$$P^{\varepsilon}(t,x,y) = \mathbb{E}^\star{}_{t,x,y}\left\{e^{-r(T-t)} h(\tilde{X}^\varepsilon_{T})\right\}.$$

Next we use the iterated expectations formula

$$\begin{aligned} &P^{\varepsilon,\delta}(t,x,y) - P^{\varepsilon}(t,x,y) \\ &\quad = \mathbb{E}^\star{}_{t,x,y}\big\{\mathbb{E}^\star\big\{e^{-r(T-t+\delta)} h(\tilde{X}^\varepsilon_{T+\delta}) - e^{-r(T-t)} h(\tilde{X}^\varepsilon_{T}) \mid (\hat{Z}^\star_s)_{t\le s\le T}\big\}\big\} \end{aligned}$$

to obtain a representation of this price difference in terms of the Black–Scholes function P_0, which is smooth away from the terminal date T. In the uncorrelated case it corresponds to the Hull–White formula [7]. In the correlated case, as considered here, this formula is in [8], and can be found in [2, (2.8.3)]. It is simple to compute explicitly the conditional distribution $\mathcal{D}(\tilde{X}^\varepsilon_T | (\hat{Z}^\star_s)_{t\le s\le T}, \tilde{X}^\varepsilon_t)$ of $\tilde{X}^\varepsilon_T$, given the path of the second Brownian motion $(\hat{Z}^\star_s)_{t\le s\le T}$. One obtains

$$\mathcal{D}(\tilde{X}^\varepsilon_T | (\hat{Z}^\star_s)_{t\le s\le T}, \tilde{X}^\varepsilon_t = x) = \mathcal{N}(m_1^\varepsilon, v_1^\varepsilon),$$

where the mean and variance are given by

$$\begin{aligned} m_1^\varepsilon &= x + \xi_{t,T} + \left(r - \frac{1}{2}\bar{\sigma}_\rho^2\right)(T-t), \\ v_1^\varepsilon &= \bar{\sigma}_\rho^2 (T-t) \end{aligned}$$

and we define

$$\begin{aligned} \xi_{t,T} &= \rho \int_t^T \tilde{f}(s, Y_s^\varepsilon)\, d\hat{Z}_s^\star - \frac{1}{2}\rho^2 \int_t^T \tilde{f}(s, Y_s^\varepsilon)^2 ds, \\ \bar{\sigma}_\rho^2 &= \frac{1-\rho^2}{T-t} \int_t^T \tilde{f}(s, Y_s^\varepsilon)^2 ds. \end{aligned} \tag{5.1}$$

It follows from the calculation that leads to the Black–Scholes formula that

$$\mathbb{E}^\star{}_{t,x,y}\{e^{-r(T-t)} h(\tilde{X}^\varepsilon_T) \mid (\hat{Z}^\star_s)_{t\le s\le T}\} = P_0(t, \tilde{X}^\varepsilon_t + \xi_{t,T}; K, T; \bar{\sigma}_\rho).$$

Similarly, we compute

$$\mathcal{D}(\tilde{X}^\varepsilon_{T+\delta} \mid (\hat{Z}^\star_s)_{t\le s\le T}, \tilde{X}^\varepsilon_t = x) = \mathcal{N}(m_2^\varepsilon, v_2^\varepsilon),$$

where the mean and variance are given by

$$m_2^\varepsilon = x + \xi_{t,T} + r\delta + \left(r - \frac{1}{2}\tilde{\sigma}^2_{\rho,\delta}\right)(T-t),$$
$$v_2^\varepsilon = \tilde{\sigma}^2_{\rho,\delta}(T-t)$$

and we define

$$\tilde{\sigma}^2_{\rho,\delta} = \bar{\sigma}^2_\rho + \frac{\delta\bar{\sigma}^2}{T-t}.$$

Therefore

$$\mathbb{E}^\star{}_{t,x,y}\{e^{-r(T-t+\delta)}h(\tilde{X}^\varepsilon_{T+\delta}) \mid (\hat{Z}^\star_s)_{t\le s\le T}\} = P_0(t, \tilde{X}^\varepsilon_t + \xi_{t,T} + r\delta; K, T; \tilde{\sigma}_{\rho,\delta}),$$

and we can write

$$\begin{aligned} &P^{\varepsilon,\delta}(t,x,y) - P^\varepsilon(t,x,y) \\ &\quad = \mathbb{E}^\star{}_{t,x,y}\left\{P_0(t, x + \xi_{t,T} + r\delta; K, T; \tilde{\sigma}_{\rho,\delta}) - P_0(t, x + \xi_{t,T}; K, T; \bar{\sigma}_\rho)\right\}. \end{aligned}$$

Using the explicit representation (4.1) and that $\bar{\sigma}_\rho$ is bounded above and below as $f(y)$ is, we find

$$|P_0(t, x + \xi_{t,T} + r\delta; K, T; \tilde{\sigma}_{\rho,\delta}) - P_0(t, x + \xi_{t,T}; K, T; \bar{\sigma}_\rho)| \le \delta c_1(e^{\xi_{t,T}}\left[|\xi_{t,T}| + 1\right] + 1)$$

for some c_1 and for δ small enough. Using the definition (5.1) of $\xi_{t,T}$ and the existence of its exponential moments, we thus find that

$$|P^\varepsilon(t,x,y) - P^{\varepsilon,\delta}(t,x,y)| \le c_2\delta$$

for some c_2 and for δ small enough.

5.2. Proof of Lemma 4.2. From the definition (3.12) of the correction $\sqrt{\varepsilon}P_1$ and the corresponding definition (4.3) of the correction $\sqrt{\varepsilon}P_1^\delta$ we deduce

$$Q^{\varepsilon,\delta} - Q^\varepsilon = \left(1 - (T-t)\left(V_3^\varepsilon\frac{\partial^3}{\partial x^3} + (V_2^\varepsilon - 3V_3^\varepsilon)\frac{\partial^2}{\partial x^2} + (2V_3^\varepsilon - V_2^\varepsilon)\frac{\partial}{\partial x}\right)\right)(P_0^\delta - P_0).$$

From the definition (3.10) of the v_i's, the definition (3.13) of the V_i's, and the bounds on the solution of the Poisson equation (3.11) given in Appendix C, it follows that

$$\max(|V_2^\varepsilon|, |V_3^\varepsilon|) \le c_1\sqrt{\varepsilon}$$

for some constant $c_1 > 0$. Notice that we can write

$$P_0^\delta(t,x) = P_0(t-\delta, x).$$

Using the explicit formula (4.1), it is easily seen that P_0 and its successive derivatives with respect to x are differentiable in t at any $t < T$. Therefore we conclude that for (t,x,y) fixed with $t < T$

$$|Q^\varepsilon(t,x) - Q^{\varepsilon,\delta}(t,x)| \le c_2\delta$$

for some $c_2 > 0$ and δ small enough.

5.3. Proof of Lemma 4.3. We first introduce some additional notation. Define the error $Z^{\varepsilon,\delta}$ in the approximation for the regularized problem by

$$P^{\varepsilon,\delta} = P_0^\delta + \sqrt{\varepsilon} P_1^\delta + \varepsilon Q_2^\delta + \varepsilon^{3/2} Q_3^\delta - Z^{\varepsilon,\delta}$$

for Q_2^δ and Q_3^δ stated below in (5.3) and (5.4). Setting

$$\mathcal{L}^\varepsilon = \frac{1}{\varepsilon}\mathcal{L}_0 + \frac{1}{\sqrt{\varepsilon}}\mathcal{L}_1 + \mathcal{L}_2,$$

one can write

$$\begin{aligned}
\mathcal{L}^\varepsilon Z^{\varepsilon,\delta} &= \mathcal{L}^\varepsilon \left(P_0^\delta + \sqrt{\varepsilon} P_1^\delta + \varepsilon Q_2^\delta + \varepsilon^{3/2} Q_3^\delta - P^{\varepsilon,\delta}\right) \\
&= \frac{1}{\varepsilon}\mathcal{L}_0 P_0^\delta + \frac{1}{\sqrt{\varepsilon}}(\mathcal{L}_0 P_1^\delta + \mathcal{L}_1 P_0^\delta) \\
&\quad + (\mathcal{L}_0 Q_2^\delta + \mathcal{L}_1 P_1^\delta + \mathcal{L}_2 P_0^\delta) + \sqrt{\varepsilon}\left(\mathcal{L}_0 Q_3^\delta + \mathcal{L}_1 Q_2^\delta + \mathcal{L}_2 P_1^\delta\right) \\
&\quad + \varepsilon\left(\mathcal{L}_1 Q_3^\delta + \mathcal{L}_2 Q_2^\delta + \sqrt{\varepsilon}\mathcal{L}_2 Q_3^\delta\right) \\
&= \varepsilon\left(\mathcal{L}_1 Q_3^\delta + \mathcal{L}_2 Q_2^\delta\right) + \varepsilon^{3/2}\mathcal{L}_2 Q_3^\delta \equiv G^{\varepsilon,\delta}
\end{aligned} \tag{5.2}$$

because $P^{\varepsilon,\delta}$ solves the original equation $\mathcal{L}^\varepsilon P^{\varepsilon,\delta} = 0$, and we choose P_0^δ, P_1^δ, Q_2^δ, and Q_3^δ to cancel the first four terms. In particular, we choose

$$Q_2^\delta(t,x,y) = -\frac{1}{2}\phi(y)\left(\frac{\partial^2 P_0^\delta}{\partial x^2} - \frac{\partial P_0^\delta}{\partial x}\right), \tag{5.3}$$

so that

$$\mathcal{L}_0 Q_2^\delta = -\mathcal{L}_2 P_0^\delta$$

(with an "integration constant" arbitrarily set to zero), whereas Q_3^δ is a solution of the Poisson equation

$$\mathcal{L}_0 Q_3^\delta = -(\mathcal{L}_1 Q_2^\delta + \mathcal{L}_2 P_1^\delta), \tag{5.4}$$

where the centering condition is ensured by our choice of P_1^δ.
At the terminal time T we have

$$Z^{\varepsilon,\delta}(T,x,y) = \varepsilon\left(Q_2^\delta(T,x,y) + \sqrt{\varepsilon} Q_3^\delta(T,x,y)\right) \equiv H^{\varepsilon,\delta}(x,y), \tag{5.5}$$

where we have used the terminal conditions $P^{\varepsilon,\delta}(T,x,y) = P_0^\delta(T,x) = h^\delta(x)$ and $P_1^\delta(T,x) = 0$. This assumes smooth derivatives of P_0^δ in the domain $t \le T$, which is the case because h^δ is smooth. It is shown in Appendix A that the source term $G^{\varepsilon,\delta}(t,x,y)$ on the right-hand side of (5.2) can be written in the form

$$\begin{aligned}
G^{\varepsilon,\delta} &= \varepsilon\left(\sum_{i=1}^4 g_i^{(1)}(y)\frac{\partial^i}{\partial x^i}P_0^\delta + (T-t)\sum_{i=1}^6 g_i^{(2)}(y)\frac{\partial^i}{\partial x^i}P_0^\delta\right) \\
&\quad + \varepsilon^{3/2}\left(\sum_{i=1}^5 g_i^{(3)}(y)\frac{\partial^i}{\partial x^i}P_0^\delta + (T-t)\sum_{i=1}^7 g_i^{(4)}(y)\frac{\partial^i}{\partial x^i}P_0^\delta\right).
\end{aligned} \tag{5.6}$$

In Appendix A we also show that the terminal condition $H^{\varepsilon,\delta}(x,y)$ in (5.5) can be written

$$H^{\varepsilon,\delta}(x,y) = \varepsilon\left(\sum_{i=1}^{2} h_i^{(1)}(y)\frac{\partial^i}{\partial x^i}P_0^\delta(T,x)\right) + \varepsilon^{3/2}\left(\sum_{i=1}^{3} h_i^{(2)}(y)\frac{\partial^i}{\partial x^i}P_0^\delta(T,x)\right). \tag{5.7}$$

To bound the contributions from the source term and terminal conditions we need the following two lemmas, which are derived in Appendices C and B, respectively.

LEMMA 5.1. *Let $\chi = g_i^{(j)}$ or $\chi = h_i^{(j)}$ with the functions $g_i^{(j)}$ and $h_i^{(j)}$ being defined in* (5.6) *and* (5.7). *Then there exists a constant $c > 0$ (which may depend on y) such that $\mathbb{E}^\star\{|\chi(Y_s^\varepsilon)|Y_t^\varepsilon = y\} \le c < \infty$ for $t \le s \le T$.*

LEMMA 5.2. *Assume $T - t > \Delta > 0$ and $\mathbb{E}^\star\{|\chi(Y_s^\varepsilon)|Y_t^\varepsilon = y\} \le c_1 < \infty$ for some constant c_1; then there exist constants $c_2 > 0$ and $\bar\delta > 0$ such that for $\delta < \bar\delta$ and $t \le s \le T$*

$$\left|\mathbb{E}^\star{}_{t,x,y}\left\{\sum_{i=1}^{n}\chi(Y_s^\varepsilon)\frac{\partial^i}{\partial x^i}P_0^\delta(s,X_s^\varepsilon)\right\}\right| \le c_2[T+\delta-s]^{\min[0,1-n/2]}, \tag{5.8}$$

and consequently

$$\begin{aligned}&\left|\mathbb{E}^\star{}_{t,x,y}\left\{\int_t^T (T-s)^p\sum_{i=1}^{n} e^{-r(s-t)}\chi(Y_s^\varepsilon)\frac{\partial^i}{\partial x^i}P_0^\delta(s,X_s^\varepsilon)\ ds\right\}\right| \\ &\le \begin{cases} c_2\ |\log(\delta)| & \text{for } n = 4+2p, \\ c_2\ \delta^{\min[0,p+(4-n)/2]} & \text{otherwise.}\end{cases}\end{aligned} \tag{5.9}$$

Proof of Lemma 4.3. We use the probabilistic representation of (5.2), $\mathcal{L}^\varepsilon Z^{\varepsilon,\delta} = G^{\varepsilon,\delta}$ with terminal condition $H^{\varepsilon,\delta}$:

$$Z^{\varepsilon,\delta}(t,x,y) = \mathbb{E}^\star{}_{t,x,y}\left\{e^{-r(T-t)}H^{\varepsilon,\delta}(X_T^\varepsilon,Y_T^\varepsilon) - \int_t^T e^{-r(s-t)}G^{\varepsilon,\delta}(s,X_s^\varepsilon,Y_s^\varepsilon)ds\right\}.$$

From Lemma 5.2 it follows that there exists a constant $c > 0$ such that

$$\left|\mathbb{E}^\star{}_{t,x,y}\left\{\int_t^T e^{-r(s-t)}G^{\varepsilon,\delta}(X_s^\varepsilon,Y_s^\varepsilon)ds\right\}\right| \le c\left\{\varepsilon + \varepsilon|\log(\delta)| + \varepsilon\sqrt{\varepsilon/\delta}\right\}, \tag{5.10}$$

$$|\mathbb{E}^\star{}_{t,x,y}\left\{H^{\varepsilon,\delta}(X_T^\varepsilon,Y_T^\varepsilon)\right\}| \le c\left\{\varepsilon + \varepsilon\sqrt{\varepsilon/\delta}\right\}, \tag{5.11}$$

and therefore also for (t,x,y) fixed with $t < T$

$$\begin{aligned}|P^{\varepsilon,\delta} - Q^{\varepsilon,\delta}| &= |\varepsilon Q_2^\delta + \varepsilon^{3/2}Q_3^\delta - Z^{\varepsilon,\delta}| \\ &\le c\left\{\varepsilon + \varepsilon|\log(\delta)| + \varepsilon\sqrt{\varepsilon/\delta}\right\},\end{aligned} \tag{5.12}$$

since Q_2^δ and Q_3^δ evaluated for $t < T$ can also be bounded using (5.3) and (A.5).

6. Conclusion. We have shown that the singular perturbation analysis of fast mean-reverting stochastic volatility pricing PDEs can be rigorously carried out for call options. We found that the leading order term and the first correction in the formal expansion are correct. The accuracy is pointwise in time, stock price, and volatility

level. It is precisely given in Theorem 3.1. The first correction involves higher-order derivatives of the Black–Scholes price, which blow up at maturity time and at the strike price. To overcome this difficulty we have used a payoff smoothing method, and we have exploited the fact that the perturbation is around the Black–Scholes price, for which there is an explicit formula. The case of call options is particularly important, since the calibration of models is based on these instruments. The case of other types of singularities is open. With some work one can certainly generalize the method presented here to other European derivatives such as binary options. The case of path-dependent derivatives such as barrier options is more difficult due to the lack of an explicit formula for the correction. The situation with American contracts such as the simplest one, the American put, is much more involved due to the singularities at the exercise boundary.

Appendix A. Expressions for source term and terminal condition. From (5.2), the source term in the equation for the error $Z^{\varepsilon,\delta}$ is

$$G^{\varepsilon,\delta} = \varepsilon\left(\mathcal{L}_1 Q_3^\delta + \mathcal{L}_2 Q_2^\delta\right) + \varepsilon^{3/2}\mathcal{L}_2 Q_3^\delta. \tag{A.1}$$

To obtain an explicit form for this source term, we consider the three terms separately. We first introduce the convenient notation

$$\mathcal{D} \equiv \frac{\partial}{\partial x},$$
$$\mathcal{D}_2 \equiv \frac{\partial^2}{\partial x^2} - \frac{\partial}{\partial x}.$$

Consider the term $\mathcal{L}_2 Q_2^\delta$ in (A.1). Using that

$$\mathcal{L}_2 = \mathcal{L}_{BS}(f(y)) = \mathcal{L}_{BS}(\bar\sigma) + \frac{1}{2}\left(f(y)^2 - \bar\sigma^2\right)\mathcal{D}_2, \tag{A.2}$$
$$\mathcal{L}_{BS}(\bar\sigma)\mathcal{D}_2 P_0^\delta = 0,$$

and (5.3), one deduces

$$\mathcal{L}_2 Q_2^\delta = -\frac{1}{4}\left(f(y)^2 - \bar\sigma^2\right)\phi(y)\mathcal{D}_2\mathcal{D}_2 P_0^\delta.$$

Consider next the term $\mathcal{L}_1 Q_3^\delta$ in (A.1). Using (3.8), we have

$$\begin{aligned} Q_3^\delta &= -\mathcal{L}_0^{-1}\left(\mathcal{L}_1 Q_2^\delta + \mathcal{L}_2 P_1^\delta - \langle \mathcal{L}_1 Q_2^\delta + \mathcal{L}_2 P_1^\delta\rangle\right) \\ &= -\mathcal{L}_0^{-1}\left(\mathcal{L}_1 Q_2^\delta - \langle \mathcal{L}_1 Q_2^\delta\rangle + \left(\mathcal{L}_2 - \langle\mathcal{L}_2\rangle\right)P_1^\delta\right). \end{aligned} \tag{A.3}$$

It follows from (5.3) that

$$\begin{aligned} \mathcal{L}_1 Q_2^\delta &= \left(\sqrt{2}\nu\rho f(y)\frac{\partial^2}{\partial x\partial y} - \sqrt{2}\nu\Lambda(y)\frac{\partial}{\partial y}\right)\left(-\frac{1}{2}\phi(y)\mathcal{D}_2 P_0^\delta\right) \\ &= -\frac{1}{\sqrt{2}}\nu\rho f(y)\phi'(y)\mathcal{D}\mathcal{D}_2 P_0^\delta + \frac{1}{\sqrt{2}}\nu\Lambda(y)\phi'(y)\mathcal{D}_2 P_0^\delta. \end{aligned}$$

Now let ψ_1 and ψ_2 be solutions of

$$\begin{aligned} \mathcal{L}_0\psi_1 &= f(y)\phi'(y) - \langle f\phi'\rangle, \\ \mathcal{L}_0\psi_2 &= \Lambda(y)\phi'(y) - \langle\Lambda\phi'\rangle; \end{aligned} \tag{A.4}$$

then we find, using (3.11) and (A.2), that Q_3^δ can be written

$$Q_3^\delta = \left(\frac{\nu\rho}{\sqrt{2}}\psi_1(y)\mathcal{D}\mathcal{D}_2P_0^\delta - \frac{\nu}{\sqrt{2}}\psi_2(y)\mathcal{D}_2P_0^\delta\right) - \frac{1}{2}\left(\phi(y)\mathcal{D}_2P_1^\delta\right). \tag{A.5}$$

Substituting for $\mathcal{L}_1$ and expanding gives

$$\begin{aligned}\mathcal{L}_1Q_3^\delta = {}& \nu^2\rho^2 f(y)\psi_1'(y)\mathcal{D}\mathcal{D}\mathcal{D}_2P_0^\delta - \nu^2\rho f(y)\psi_2'(y)\mathcal{D}\mathcal{D}_2P_0^\delta\\ & - \nu^2\rho\Lambda(y)\psi_1'(y)\mathcal{D}\mathcal{D}_2P_0^\delta + \nu^2\Lambda(y)\psi_2'(y)\mathcal{D}_2P_0^\delta\\ & - \frac{\nu}{\sqrt{2}}\left(\rho f(y)\phi'(y)\mathcal{D}\mathcal{D}_2P_1^\delta - \Lambda(y)\phi'(y)\mathcal{D}_2P_1^\delta\right).\end{aligned}$$

Consider finally the term $\mathcal{L}_2Q_3^\delta$ in (A.1); we find, using (A.2) and (A.5),

$$\begin{aligned}\mathcal{L}_2Q_3^\delta = {}& \frac{1}{2}(f(y)^2-\bar\sigma^2)\left[\frac{\rho\nu}{\sqrt{2}}\psi_1(y)\mathcal{D}_2\mathcal{D}\mathcal{D}_2P_0^\delta - \frac{\nu}{\sqrt{2}}\psi_2(y)\mathcal{D}_2\mathcal{D}_2P_0^\delta - \frac{1}{2}\phi(y)\mathcal{D}_2\mathcal{D}_2P_1^\delta\right]\\ & - \frac{1}{2}\phi(y)\mathcal{D}_2\left(v_3\mathcal{D}_3P_0^\delta + v_2\mathcal{D}_2P_0^\delta\right),\end{aligned}$$

with

$$\mathcal{D}_3 = \frac{\partial^3}{\partial x^3} - 3\frac{\partial^2}{\partial x^2} + 2\frac{\partial}{\partial x}$$

and $v_{2,3}$ defined in (3.10).

To summarize, the source term is given by

$$\begin{aligned}G^{\varepsilon,\delta} = {}& \varepsilon\left\{\nu^2\rho^2 f(y)\psi_1'(y)\mathcal{D}\mathcal{D}\mathcal{D}_2P_0^\delta - \nu^2\rho f(y)\psi_2'(y)\mathcal{D}\mathcal{D}_2P_0^\delta\right.\\ & - \nu^2\rho\Lambda(y)\psi_1'(y)\mathcal{D}\mathcal{D}_2P_0^\delta + \nu^2\Lambda(y)\psi_2'(y)\mathcal{D}_2P_0^\delta\\ & - \frac{\nu}{\sqrt{2}}\left(\rho f(y)\phi'(y)\mathcal{D}\mathcal{D}_2P_1^\delta - \Lambda(y)\phi'(y)\mathcal{D}_2P_1^\delta\right)\\ & \left. - \frac{1}{4}\left(f(y)^2-\bar\sigma^2\right)\phi(y)\mathcal{D}_2\mathcal{D}_2P_0^\delta\right\}\\ & + \varepsilon^{3/2}\left\{\frac{1}{2}(f(y)^2-\bar\sigma^2)\left[\frac{\rho\nu}{\sqrt{2}}\psi_1(y)\mathcal{D}_2\mathcal{D}\mathcal{D}_2P_0^\delta - \frac{\nu}{\sqrt{2}}\psi_2(y)\mathcal{D}_2\mathcal{D}_2P_0^\delta - \frac{1}{2}\phi(y)\mathcal{D}_2\mathcal{D}_2P_1^\delta\right]\right.\\ & \left. - \frac{1}{2}\phi(y)\mathcal{D}_2(v_3\mathcal{D}_3P_0^\delta + v_2\mathcal{D}_2P_0^\delta)\right\}.\end{aligned}$$

By inspection, this can be written in the form (5.6).

From (5.3) and (A.5) we can also see that the terminal condition $H^{\varepsilon,\delta}$ in (5.5) can be written in the form (5.7).

Appendix B. Proof of Lemma 5.2. To prove Lemma 5.2 notice first that a calculation based on the analytic expression for the Black–Scholes price in the standard constant volatility case gives

$$\partial_x^n P_0^\delta(s,x) = \begin{cases} e^x N(u/\tau + b\tau) & \text{for } n=1,\\ e^x N(u/\tau + b\tau) + \sum_{i=0}^{n-2}\frac{b_i^{(n)}}{\tau}e^u\partial_u^i e^{-(u/\tau+b\tau)^2/2} & \text{for } n\geq 2\end{cases} \tag{B.1}$$

for some constants b_i and with

$$\tau \equiv \bar{\sigma}\sqrt{T+\delta-s},$$
$$u \equiv x - \log(K),$$
$$b \equiv \left(\frac{r}{\bar{\sigma}^2} + \frac{1}{2}\right).$$

Assume first that $T - s \geq (T-t)/2 > 0$, so that $\tau \geq \bar{\sigma}\sqrt{(T-t)/2}$. Since $\partial_x^i P_0^\delta(s,x)$ is uniformly bounded in δ, it follows that

$$|\mathbb{E}^\star{}_{t,x,y}\left\{\chi(Y_s^\varepsilon)\partial_x^i P_0^\delta(s,X_s^\varepsilon)\right\}| \leq c\mathbb{E}^\star{}_{t,x,y}\left\{|\chi(Y_s^\varepsilon)|\right\} \tag{B.2}$$

for some constant c that may depend on x.

Consider next the case $0 < T - s < (T-t)/2$; then

$$\left|\mathbb{E}^\star{}_{t,x,y}\left\{\chi(Y_s^\varepsilon)\partial_x^i P_0^\delta(s,X_s^\varepsilon)\right\}\right| = \left|\mathbb{E}^\star{}_{t,x,y}\left\{\chi(Y_s^\varepsilon)\mathbb{E}^\star{}_{t,x,y}\left\{\partial_x^i P_0^\delta(s,X_s^\varepsilon) \mid \hat{Z}_v^\star; t \leq v \leq s\right\}\right\}\right|$$

and

$$\begin{aligned}
&\left|\mathbb{E}^\star{}_{t,x,y}\left\{\frac{1}{\tau}e^u\partial_u^i e^{-(u/\tau+b\tau)^2/2} \;\middle|\; \hat{Z}_v^\star; t \leq v \leq s\right\}\right| \\
&= \frac{1}{\tau}\left|\int e^u \partial_u^i e^{-(u/\tau+b\tau)^2/2} p(u)du\right| \\
&= \frac{1}{\tau^i}\left|\int e^{\tau u}\partial_u^i e^{-(u+b\tau)^2/2}p(\tau u)du\right| \;\leq\; \frac{c}{\tau^i},
\end{aligned} \tag{B.3}$$

where p is the conditional distribution of $u \equiv X_s^\varepsilon - \log(K)$, which is the Gaussian distribution with variance at least $(T-t)(1-\rho^2)m_1^2/2$. The bound (5.8) follows readily from (B.1), (B.2), and (B.3). The bound (5.9) is a direct consequence of (5.8), and Lemma 5.2 is established.

Appendix C. On the solution of the Poisson equation. Let χ solve

$$\mathcal{L}_0\chi + g = 0,$$

with $\mathcal{L}_0$ defined as in (2.5) and with g satisfying the centering condition

$$\langle g \rangle = 0,$$

where the averaging is done with respect to the invariant distribution associated with the infinitesimal generator $\mathcal{L}_0$ (see (3.3) for an explicit formula). Using the explicit form of the differential operator $\mathcal{L}_0$, one can easily deduce that

$$\Phi(y)\chi'(y) = \frac{-1}{\nu^2}\int_{-\infty}^{y} g(z)\Phi(z)\,dz = \frac{1}{\nu^2}\int_{y}^{\infty} g(z)\Phi(z)\,dz,$$

with Φ being the probability density of the invariant distribution $\mathcal{N}(m,\nu^2)$ associated with $\mathcal{L}_0$. From this it follows that if g is bounded,

$$|\chi'(y)| \leq c_1,$$
$$|\chi(y)| \leq c_2(1 + \log(1+|y|)).$$

Notice that χ in Lemma 5.1 satisfies

$$|\chi(y)| \leq c \max(|\phi(y)|, |\phi'(y)|, |\psi_{1,2}(y)|, |\psi'_{1,2}(y)|)$$

for some constant c and with ϕ and $\psi_{1,2}$ defined in (3.11) and (A.4), respectively. These functions are solutions of Poisson equations with $g = f^2 - \langle f^2 \rangle$ or $g = f\phi' - \langle f\phi' \rangle$ or $g = \Lambda\phi' - \langle \Lambda\phi' \rangle$, which are bounded. Therefore $\chi(y)$ is at most logarithmically growing at infinity. The bound in Lemma 5.1 now follows from classical a priori estimates on the moments of the process Y_t^ε, which are uniform in ε. In the case $\Lambda = 0$ this can easily be seen by a simple time change $t = \varepsilon t'$ in (2.2). The case $\Lambda \neq 0$ follows by a Girsanov change of measure argument.

REFERENCES

[1] D. Duffie, *Dynamic Asset Pricing Theory*, 2nd ed., Princeton University Press, Princeton, NJ, 1996.

[2] J.-P. Fouque, G. Papanicolaou, and K.R. Sircar, *Derivatives in Financial Markets with Stochastic Volatility*, Cambridge University Press, Cambridge, UK, 2000.

[3] J.-P. Fouque, G. Papanicolaou, and K.R. Sircar, *Mean-reverting stochastic volatility*, Internat. J. Theoret. Appl. Finance, 13 (2000), pp. 101–142.

[4] J.-P. Fouque and T. Tullie, *Variance reduction for Monte Carlo simulation in a stochastic volatility environment,* Quantitative Finance, 2 (2002), pp. 24–30.

[5] R. Frey, *Derivative asset analysis in models with level-dependent and stochastic volatility,* CWI Quarterly, 10 (1996), pp. 1–34.

[6] E. Ghysels, A. Harvey, and E. Renault, *Stochastic volatility,* in Statistical Methods in Finance, G. Maddala and C. Rao, eds., Handbook of Statist. 14, North–Holland, Amsterdam, 1996, pp. 119–191.

[7] J. Hull and A. White, *The pricing of options on assets with stochastic volatilities,* J. Finance, 42 (1987), pp. 281–300.

[8] G. Willard, *Calculating Prices and Sensitivities for Path-Independent Derivative Securities in Multifactor Models,* Ph.D. thesis, Washington University in St. Louis, St. Louis, MO, 1996.

SIAM J. APPL. MATH.
Vol. 63, No. 5, pp. 1666–1681

CHEMOTACTIC CELLULAR MIGRATION: SMOOTH AND DISCONTINUOUS TRAVELLING WAVE SOLUTIONS*

K. A. LANDMAN†, G. J. PETTET‡, AND D. F. NEWGREEN§

Abstract. A simple model of chemotactic cell migration gives rise to travelling wave solutions. By varying the cellular growth rate and chemoattractant production rate, travelling waves with both smooth and discontinuous fronts are found using phase plane analysis. The phase plane exhibits a curve of singularities whose position relative to the equilibrium points in the phase plane determines the nature of the heteroclinic orbits, where they exist. Smooth solutions have trajectories connecting the steady states lying to one side of the singular curve. Travelling shock waves arise by connecting trajectories passing through a special point in the singular curve and recrossing the singular curve, by way of a discontinuity. Hyperbolic partial differential equation theory gives the necessary shock condition. Conditions on the parameter values determine when the solutions are smooth travelling waves versus discontinuous travelling wave solutions. These conditions provide bounds on the travelling wave speeds, corresponding to bounds on the chemotactic velocity or bounds on cellular growth rate. This analysis gives rise to the possibility of representing sharp fronts to waves of invading cells through a simple chemotactic term, without introducing a nonlinear diffusion term. This is more appropriate when cell populations are sufficiently dense.

Key words. migration, chemotaxis, travelling wave, phase plane, shock

AMS subject classifications. 34A34, 35L40, 35L67, 92C17

DOI. 10.1137/S0036139902404694

1. Introduction.

The active migration of cells is a significant feature of numerous biological phenomena ranging from wound healing, scar tissue formation, and tumor invasion to embryo implantation and organogenesis.

Despite having been the focus of much research, a comprehensive understanding of the processes of receptor activation, cell-cell signalling, and intracellular organization associated with cell migration in various contexts eludes us. In part this is due to the numerous and complicated mechanisms that are involved and, perhaps more importantly, to the cooperative or antagonistic interactions between such processes.

Mathematical modelling of biological phenomena provides a timely and efficient theoretical tool for considering the interaction of various cell migration mechanisms, and the emergent behavior that arises from these interactions. To date, much of the mathematical modelling of cell migration has been predicated on the phenomena of diffusion, chemotaxis, and haptotaxis, either singly or in combination. Such models typically support travelling wave solutions, which are taken to represent the invading fronts of populations of migrating cells. An exemplar of such a model based on the process of diffusion is Fisher's equation on a one-dimensional spatial domain [15]. Chemotaxis-based models employ a gradient of a diffusible signaling chemical to determine the velocity of cell migration [1], [14], [18], whilst haptotaxis-based models

*Received by the editors March 28, 2002; accepted for publication (in revised form) January 15, 2003; published electronically July 26, 2003.
http://www.siam.org/journals/siap/63-5/40469.html
†Department of Mathematics and Statistics, University of Melbourne, Victoria 3010, Australia (k.landman@ms.unimelb.edu.au).
‡Centre in Statistical Science and Industrial Mathematics, School of Mathematical Sciences, Queensland University of Technology, Brisbane, GPO Box 2434, Queensland 4001, Australia (gpettet@fsc.qut.edu.au).
§The Embryology Laboratory, The Murdoch Children's Research Institute, Royal Children's Hospital, Parkville, Victoria 3052, Australia (newgreen@cryptic.rch.unimelb.edu.au).

employ a gradient of extracellular matrix or ligand density [17] to this end.

The extensive literature developed over the last few decades concerning the theoretical modelling of chemotaxis is testament to the perceived importance of the role it plays in cell migration. Many of these mathematical models can identify their origins in the work of Keller and Segel [9] describing the motion of bacteria as a chemotactic response. Significant contributions to this body of knowledge include those by Tranquillo [22], Tranquillo and Alt [23], Hillen [8], and Othmer and Stevens [16], while others may be identified in a recent review by Ford and Cummings [6].

The analysis of such models can lead to an understanding of the relative contributions by the mechanisms modelled to the speed of the invading front of cells, and, by implication, potential strategies for effecting changes to the speed of the invading front can be hypothesized and investigated.

Typically, invasive phenomena in the context of migrating populations of cells are characterized by a well-defined boundary. This feature is difficult to reproduce when using a diffusive flux to represent the migration process. Mathematical models involving a linear diffusive flux give rise to smooth-fronted solutions, with the solution being nonzero everywhere (albeit small). A nonlinear diffusive flux, where the diffusivity is density dependent (and equal to zero when the density is zero), gives rise to solutions with distinct interfaces beyond which the density equals zero. Such models have been explored in describing population pressures and moisture infiltration [5], [7], [15], [20], [24], [25]. Even though such solutions have compact support, they are smooth when the density is positive and do not exhibit shocks. Solutions with shocks are not realizable with a diffusive mechanism.

For a limited number of specific and very simple models based on chemotaxis, recent research by Pettet, McElwain, and Norbury [19] and others [17], [12] has shown the potential for chemotaxis-based models to exhibit travelling wave solutions with shock fronts. Such sharp-fronted solutions may be viewed as being more indicative of invading cell populations with a well-defined front or margin than those described above.

In this article we consider a simple model of chemotactic cell migration, where there is no contribution to the cell velocity from a diffusion-like term. We explore this theoretical model of cell migration numerically and analytically to show that, for various parameter regimes, smooth-fronted or shock-fronted travelling wave solutions can be supported.

A coupled system of partial differential equations for cell density and chemoattractant concentration is considered. We introduce a travelling wave coordinate system with an unknown wave speed to convert the system into a coupled system of ordinary differential equations. This is explored using phase plane analysis, giving rise to a rich variety of possible solutions. Consideration of the original system as a hyperbolic system allows shock conditions to be specified uniquely. Some asymptotic analysis for large wave speed is also examined.

2. Chemotactic cell migration in a fixed spatial domain of one dimension. We begin by describing a simple system of equations designed to describe the chemotactic migration of cells in a fixed domain. We use a coordinate x fixed in space (i.e., a Eulerian system). The cell density per unit length and the chemoattractant concentration are denoted by $n(x,t)$ and $g(x,t)$, respectively. The usual conservation-of-mass argument for a generic chemotaxis problem gives

$$\frac{\partial n}{\partial t} = -\chi \frac{\partial}{\partial x}\left(n\frac{\partial g}{\partial x}\right) + f(n,g), \tag{2.1}$$

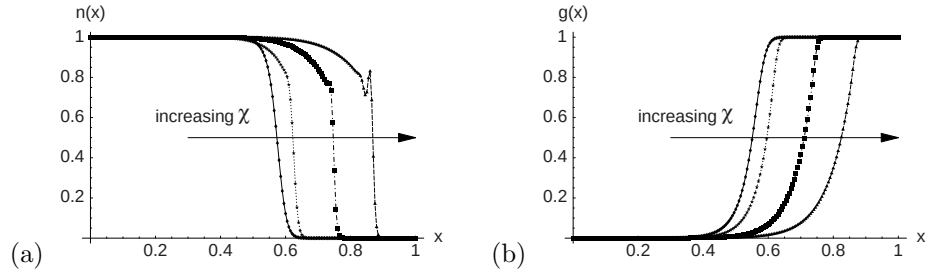


FIG. 2.1. *Chemotactic cell migration at different wave speeds. Numerical solutions of equations* (2.1)–(2.4) *on* $[0,1]$ *with the inclusion of Fickian diffusion* ($D_n \frac{\partial^2 n}{\partial x^2}$) *in* (2.1). *Here* $D_n = 0.00001$, $\lambda_1 = 2$, $k_1 = 1$, $\lambda_2 = 0.25$, $\lambda_3 = 1$, $k_2 = 1$ *and* $\chi = 0.0001, 0.001, 0.002$ *and* 0.003. *Initial conditions are* $n(x,0) = e^{-500x^2}$ *and* $g(x,0) = 1$ *with zero flux boundary conditions at* $x = 0, 1$. (a) *Advancing front (moving left to right) of migrating cells at dimensioned time* $t = 17.5$. (b) *Retreating front (moving left to right) of chemoattractant at dimensioned time* $t = 17.5$.

$$\frac{\partial g}{\partial t} = h(n,g), \tag{2.2}$$

where the chemotactic factor χ is assumed to be a constant and f and h represent the kinetic terms. This type of model system, where only chemotaxis contributes to the cell migration, has been studied by several authors [2], [19], [11]. For this reason we have excluded diffusivity from the model equations (2.1)–(2.2).

Our problem concerns cells n, which proliferate by mitosis and may die or differentiate into another cell type. These two effects can be incorporated into a logistic-type term for f. The chemoattractant g is produced uniformly throughout the domain and has a maximum value. Furthermore, the cells consume the chemoattractant, which creates a gradient of g and produces a migration velocity. These effects are reflected in our choice of f and h as

$$f = \lambda_1 n(k_1 - n), \tag{2.3}$$

$$h = \lambda_2 g(k_2 - g) - \lambda_3 ng. \tag{2.4}$$

Numerical solutions to such a system on a finite domain exhibit travelling wave solutions, as illustrated in Figure 2.1. Here we have included a small amount of diffusion in n in order to stabilize the system, allowing the use of the Numerical Algorithms Group (NAG) parabolic partial differential equations package DO3PCF. In Landman, Pettet, and Newgreen [10] we explicitly introduce two migration mechanisms, namely, chemotaxis and diffusion. We look at the effect of decreasing diffusivity, when the diffusion coefficient is small in comparison to the chemotactic sensitivity coefficient, and show that the solutions look almost identical as the diffusion coefficient is reduced by several orders of magnitude. Since our interest is in systems of invasion, for which chemotaxis is the dominant cell migration mechanism, the arguments considered in [10] allow us to neglect any effect attributable to the small diffusivity introduced for these numerical results.

We observe that the front of n steepens as the chemoattractant coefficient χ increases. If the parameter is pushed too far, the solution appears to develop a numerical instability which may be interpreted as the evolution of a jump discontinuity. These numerical simulations initiate questions about the nature of such solutions and whether or not smooth and discontinuous travelling wave solutions can be determined analytically.

An appropriate dimensionalization is carried out with the following transformation, where we introduce a length scale L and scaled parameters a and b as shown:

$$n = k_1 n^*, \qquad g = k_2 g^*, \qquad t = Tt^*, \qquad x = Lx^*, \tag{2.5}$$

$$T = \frac{1}{\lambda_3 k_1}, \qquad L^2 = \frac{\chi k_2}{\lambda_3 k_1}, \qquad a = \frac{\lambda_1}{\lambda_3}, \qquad b = \frac{\lambda_2 k_2}{\lambda_3 k_1}. \tag{2.6}$$

Omitting the asterisk notation, the dimensionless system is

$$\frac{\partial n}{\partial t} = -\frac{\partial}{\partial x}\left(n\frac{\partial g}{\partial x}\right) + an(1-n), \tag{2.7}$$

$$\frac{\partial g}{\partial t} = bg(1-g) - ng. \tag{2.8}$$

There are many scalings we could have chosen. This scaling focuses on the two most important terms in the problem, namely, chemotactic migration and the interaction terms between n and g.

Now, making the traveling wave coordinate transformation $z = x - ct$, we obtain

$$\frac{dn}{dz} = \frac{1}{c}\frac{d}{dz}\left(n\frac{dg}{dz}\right) - \frac{an}{c}(1-n), \tag{2.9}$$

$$\frac{dg}{dz} = -\frac{1}{c}\left[bg(1-g) - ng\right], \tag{2.10}$$

which, after appropriate substitutions from (2.10) into (2.9), may be written as

$$\left[1 + \frac{g}{c^2}(b(1-g) - 2n)\right]\frac{dn}{dz} = \frac{ng}{c^3}\left[b(1-2g) - n\right]\left[b(1-g) - n\right] - \frac{an}{c}(1-n), \tag{2.11}$$

$$\frac{dg}{dz} = -\frac{1}{c}\left[bg(1-g) - ng\right]. \tag{2.12}$$

We will be considering (2.11)–(2.12) in the phase plane, and we will plot trajectories in the (g, n) plane. For $b > 0$, the steady states of the system are $(g, n) = (0,0), (1,0), (0,1)$ and $(1-\frac{1}{b}, 1)$. Since we are interested only in solutions where n and g are nonnegative, the last state exists only for $b > 1$. When $b = 0$, the steady state $(0,0)$ is replaced by a continuum of steady states $(\bar{g}, 0)$. In this paper, we concentrate on the case $b > 0$ and briefly comment on the degenerate case $b = 0$ in section 5.3.

We seek traveling wave solutions connecting $(0,1)$ or $(1-\frac{1}{b}, 1)$ and $(1,0)$. Linearization about the steady states yields the nature of these states. This information, together with the eigenvalues and eigenvectors, is listed in Table 1.

3. Phase plane analysis. We will investigate the positive quadrant of the (g, n) phase plane; this is made interesting by the position of the curve, where the function premultiplying $\frac{dn}{dz}$ in (2.11) is identically equal to zero. Pettet, McElwain, and Norbury [19] defined such a curve as a "wall-of-singularities." Here the wall-of-singularities can be written as

$$n = \frac{1}{2}\left(\frac{c^2}{g} + b(1-g)\right). \tag{3.1}$$

This wall is asymptotic to the n-axis, cutting the positive g-axis at

$$g = \frac{1}{2}\left(1 + \sqrt{1 + \frac{4c^2}{b}}\right),$$

TABLE 1
Nature of equilibrium points.

(g, n)	Type	Eigenvalues	Corresponding eigenvectors (g, n)
$(0, 0)$	Stable node	$-\frac{a}{c}, -\frac{b}{c}$	$(0, 1)$, $(1, 0)$
$(1, 0)$	Saddle	$-\frac{a}{c}, \frac{b}{c}$	$(1, -a - b)$, $(1, 0)$
$(0, 1)$	$b < 1$ Unstable node $b > 1$ Saddle	$\frac{a}{c}, \frac{1-b}{c}$	$(0, 1)$, $\left(1, -\frac{(b-1)^2}{c^2(a+b-1)}\right)$
$(1 - \frac{1}{b}, 1)$ for $b > 1$	$c^2 > 1 - \frac{1}{b}$ Unstable node $c^2 < 1 - \frac{1}{b}$ Saddle	$\frac{(a+b-1)c \pm \sqrt{\beta+\gamma^2}}{2(c^2-1+\frac{1}{b})}$, where $\beta = \frac{4a(b-1)^2}{b}$, $\gamma = (a - b + 1)c$	Complicated form

to the right of the steady state $(1, 0)$. Hence all the steady states are to the left of the wall when $0 < b < 1$. If $b > 1$, the new steady state $(1 - \frac{1}{b}, 1)$ is below (above) the wall if $c^2 > 1 - \frac{1}{b}$ ($c^2 < 1 - \frac{1}{b}$). From Table 1, we can see that the nature of this steady state changes according to the same inequality. The wall gets closer to the origin as c decreases.

Pettet, McElwain, and Norbury [19] showed that solutions approaching a wall-of-singularities could not cross the wall unless it passed through special points called gates or holes in the wall. These points are defined by both the function premultiplying $\frac{dn}{dz}$ and the right-hand side in (2.11) being equal to zero simultaneously. Thus, travelling wave solutions joining two steady states (one unstable and the other stable) on the same side of the wall-of-singularities could under some circumstance be shown not to exist when the wall-of-singularities interfered with the trajectory emanating from the unstable steady state.

Pettet, McElwain, and Norbury concerned themselves only with smooth-fronted travelling waves. They presumed that any trajectory exiting the unstable steady state of interest that passed through a hole in the wall could then not recross the wall in order to connect with the stable steady state. However, Marchant, Norbury, and Perumpanani [12] showed that for a very simple system of equations (in the class of (2.1)–(2.2)) a trajectory in the phase plane could indeed follow such a path, recrossing the wall by way of a jump discontinuity.

In the system we describe here, there is always a hole at the intersection of the wall with the g-axis. However, it is necessary for the holes to lie within the positive (g, n) quadrant if any trajectory passing through the hole is to remain in that quadrant. Depending on the parameter values, for our system there can be no, one, or two holes contained within the positive quadrant.

The interaction of the trajectories and the wall-of-singularities is extremely interesting. We start by giving some examples.

Consider first the case $0 < b < 1$. We seek a trajectory connecting the unstable node $(0, 1)$ to the saddle $(1, 0)$. By considering the vector field associated with (2.11) and (2.12), it can be shown that such a trajectory certainly exists if the wall is sufficiently far from the axes. For example, we fix the wall position (fix b and c) and vary a, the effective cellular growth rate or mitotic index of n, as illustrated in Figure 3.1. For sufficiently small values of a there is a unique trajectory to the left of the

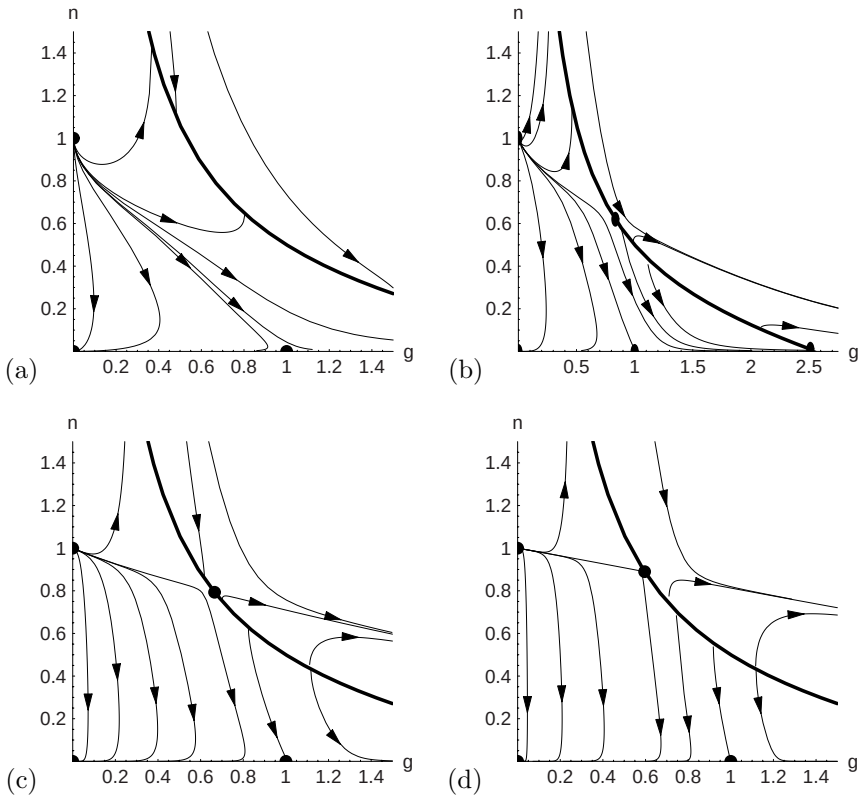


FIG. 3.1. *Phase plane for (g, n) for increasing values of cellular growth rate a. Here $b = 0.25, c = 1$. The wall is indicated with a dark line; the holes in the wall and the steady states are marked with a* ●. (a) $a = 0.5$, *no holes.* (b) $a = 1.0$, *two holes (both with $g > 0$).* (c) $a = 2.0$, *one hole.* (d) $a = 4.0$, *one hole.*

wall, connecting the two states (shown here with $a = 0.5$ and 1); this gives a smooth travelling wave. However, for large enough values of a, no such trajectory can be found below the wall, as shown here with $a = 2$ and 4. In fact, there appears to be a trajectory from $(0, 1)$ travelling towards the hole in the wall.

Alternatively, we can consider the effect of fixing the two rates a and b and decreasing the wave speed c. This translates the wall closer to the axes, as shown in Figure 3.2. When $c = 1.5$, there is a trajectory lying below the wall joining the steady states. However, for $c = 1$ and 0.5, no such trajectory exists, but again there is one trajectory from $(0, 1)$ heading towards the hole in the wall.

Since trajectories cannot cross each other, or cross the wall at any point other than a hole in the wall, we must determine how it is possible for a trajectory passing through a hole to recross the wall and connect to the other steady state. Marchant [11] investigated this scenario for his system, and his arguments apply equally to our system of equations. No smooth connection between the two states can be made; however, there is the possibility for the solution to be nonsmooth, by containing a jump discontinuity.

It is appropriate to apply Marchant's approach here to our system. We do this now, and then return to the phase plane analysis in section 5.

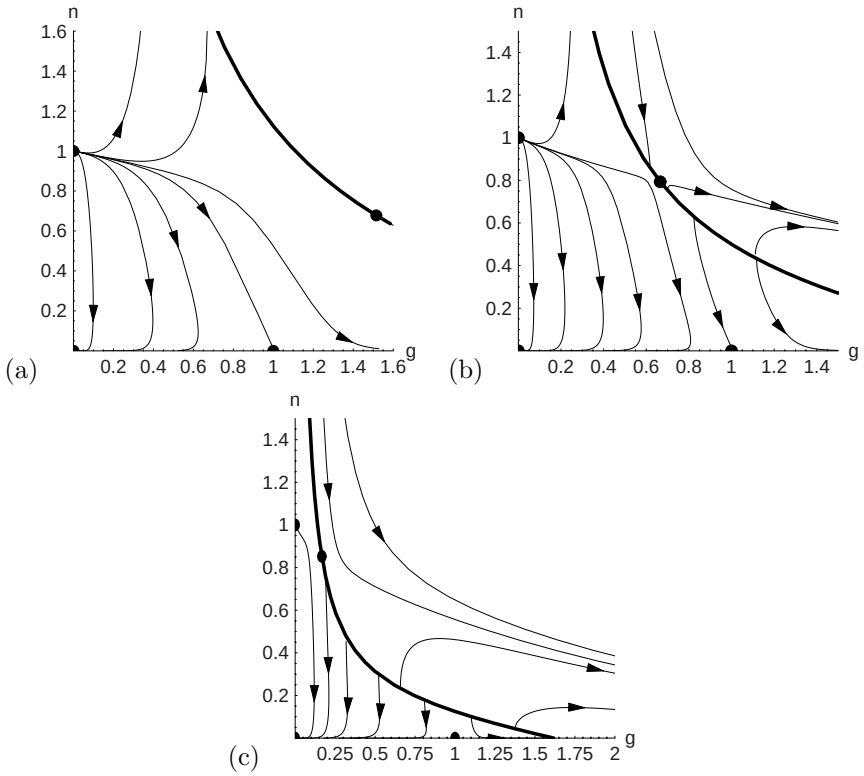


FIG. 3.2. *Phase plane for (g, n) for decreasing values of wave speed c. Here $a = 2.0$, $b = 0.25$. Wall, holes, and steady states marked as indicated previously. Each example has one hole in the positive quadrant.* (a) $c = 1.5$. (b) $c = 1.0$. (c) $c = 0.5$.

4. Hyperbolic PDE theory: Shocks and discontinuities. Introducing a third variable $u = \frac{\partial g}{\partial x}$, consider the scaled generic chemotaxis problem (2.7) and (2.8) as a hyperbolic system, namely,

$$\frac{\partial n}{\partial t} = -\frac{\partial}{\partial x}(nu) + f(n,g) = -u\frac{\partial n}{\partial x} - n\frac{\partial u}{\partial x} + f(n,g), \tag{4.1}$$

$$\frac{\partial u}{\partial t} = h_n\frac{\partial n}{\partial x} + h_g\frac{\partial g}{\partial x}, \tag{4.2}$$

$$\frac{\partial g}{\partial t} = h(n,g), \tag{4.3}$$

which in matrix form becomes

$$\frac{\partial}{\partial t}\begin{bmatrix} n \\ u \\ g \end{bmatrix} + \begin{bmatrix} u & n & 0 \\ -h_n & 0 & -h_g \\ 0 & 0 & 0 \end{bmatrix}\frac{\partial}{\partial x}\begin{bmatrix} n \\ u \\ g \end{bmatrix} = \begin{bmatrix} f \\ 0 \\ h \end{bmatrix}. \tag{4.4}$$

The characteristic slopes are determined from the eigenvalues of the 3×3 matrix in (4.4). These are solutions of

$$\lambda\left[(u-\lambda)(\lambda) - nh_n\right] = 0. \tag{4.5}$$

There are three distinct solutions for $h_n < 0$ and $n > 0$. This confirms that the

system is strictly hyperbolic [21] with

$$\lambda_1 = \frac{1}{2}\left[u - \sqrt{u^2 - 4nh_n}\right] \le \lambda_2 = 0 \le \lambda_3 = \frac{1}{2}\left[u + \sqrt{u^2 - 4nh_n}\right]. \tag{4.6}$$

The characteristics have gradient $dx/dt = \lambda_i$, which is never infinite, so the line $t = 0$ is nowhere tangent to a characteristic. Hence if initial data for n, u, g is given along the line $t = 0$, the resulting Cauchy problem is well posed. By considering the matrix of right eigenvectors, which correspond to each λ_i, the λ_2 field is always linearly degenerate, and the λ_1 and λ_3 fields are genuinely nonlinear characteristic fields for (n, u, g) in the positive quadrant.

A shock (i.e., a curve separating intersecting characteristics defining a discontinuity in at least one of the variables on either side of the curve) may exist in either of the two genuinely nonlinear fields. We are looking for a shock that propagates with the travelling wave speed c, since *all* the information on a travelling wave moves with this speed. Following Marchant [11], [12], the Lax entropy condition ensures that the shocks are physically relevant [4]; hence, since the wave speed $c > 0$, only the λ_3 field is relevant.

4.1. Shock conditions. We write the system (4.4) in conservation form,

$$\frac{\partial P}{\partial t} + \frac{\partial Q}{\partial x} = S, \tag{4.7}$$

where

$$P = \begin{bmatrix} n \\ u \\ g \end{bmatrix}, \quad Q = \begin{bmatrix} nu \\ -h \\ 0 \end{bmatrix}, \quad S = \begin{bmatrix} f \\ 0 \\ h \end{bmatrix}. \tag{4.8}$$

The Rankine–Hugoniot jump condition [4] defining the shock moving with velocity c in the third field is

$$[P]\, c = [Q]\,, \tag{4.9}$$

where $[q]$ denotes the jump in the quantity q. For our system this gives

$$[n]\, c = [nu]\,, \tag{4.10}$$

$$[u]\, c = [-h]\,, \tag{4.11}$$

$$[g]\, c = 0. \tag{4.12}$$

Since $u = \frac{\partial g}{\partial x} = -\frac{1}{c}h$, the second equation always holds, while the third equation says that there is no discontinuity in g. Using the definition of u and our particular choice of (dimensionless) kinetic term $h = bg(1-g) - ng$, the first equation gives

$$\begin{aligned} [n]\, c = [nu] &= \left[-\frac{1}{c}nh\right] \\ &= -\frac{1}{c}[bng(1-g) - n^2 g] \\ &= -\frac{1}{c}bg(1-g)[n] + \frac{1}{c}g[n^2]. \end{aligned} \tag{4.13}$$

This simplifies to

$$\left(c^2 + bg(1-g)\right)[n] = g\left[n^2\right]. \tag{4.14}$$

Using the definition $[n] = n_L - n_R$, where n_L and n_R are the values of n on either side of the shock, (4.14) and (4.12) reduce to

$$n_L + n_R = \frac{1}{g}\left(c^2 + bg(1-g)\right), \quad g_L = g_R. \tag{4.15}$$

Recall that the wall-of-singularities satisfies (3.1). Hence, the geometric center of the jump $\frac{1}{2}(n_L + n_R)$ lies exactly on the wall-of-singularities, and therefore any jump takes the trajectory to the other side of the wall. Of course, a jump is only allowable if $n_R > 0$. Note that the Lax entropy condition [4] for the λ_3 field is satisfied only if $n_L > n_R$.

4.2. Power series. The trajectories needed to construct a travelling shock wave can be determined by power series solutions. Equations (2.11) and (2.12) can be written in the form

$$\frac{dn}{dg} = -\frac{\frac{ng}{c^2}\left[b(1-2g)-n\right]\left[b(1-g)-n\right] - an(1-n)}{\left[1+\frac{g}{c^2}(b(1-g)-2n)\right]\left[bg(1-g)-ng\right]}. \tag{4.16}$$

Solutions $n(g)$ can be found by expanding in powers about special points. Such points are steady-state solutions to the system and holes in the wall. Let (g_s, n_s) be such a point, and then write

$$g = g_s + f, \tag{4.17}$$

$$n(g) = n_s + \alpha_1 f + \alpha_2 f^2 + \alpha_3 f^3 + \cdots. \tag{4.18}$$

The coefficients α_i are determined sequentially by substituting into (4.16) and then matching powers of f. The resulting power series has a radius of convergence defined by the analyticity of the right-hand side of (4.16). This term is not analytic along the wall, and the lines $g = 0$ and $n = bg(1-g)$. Below we will be generating a power series about a hole in the wall and around (1, 0).

5. Phase plane revisited.

5.1. Production rate of g satisfies $0 < b < 1$. We now show how to construct a travelling shock wave to the example in Figure 3.1(c), where there is one hole in the wall in the positive quadrant. This is illustrated in Figure 5.1(a). We first determine the power series about this hole and find that there are two possible values of α_1. Each of these values provides unique values of the other α_i; hence we obtain two trajectories through the hole in the wall. One of these passes through the n-axis at $n = 1$, and this is the one of interest. Points on this curve, to the right of the hole, are possible values of n_L. We next determine the power series through $(1, 0)$ and determine its intersection with the wall (the limit of its convergence). To the right of this, we determine the curve which lies below the wall, which marks the outer envelope of the possible points for n_R, such that the midpoint of the shock lies on the wall. There is a unique value of g that satisfies the jump conditions (4.15). Hence we obtain two trajectories, one allowing passage through the hole, and, by recrossing the wall with a jump discontinuity, the other connecting to the steady state on the g-axis. The corresponding $n(z)$ and $g(z)$ are shown in Figure 5.1(b) (where we have arbitrarily placed the shock at $z = 0$).

This example illustrates two interesting facts about the solutions for n. For wave speed sufficiently large, a unique smooth travelling solution exists between the steady states $(0, 1)$ and $(1, 0)$. Furthermore, there exists a sufficiently large cellular growth

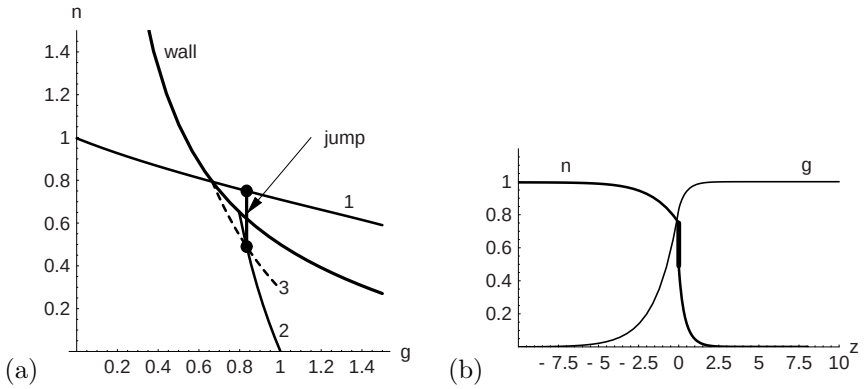


FIG. 5.1. *Travelling shock wave. Here* $a = 2.0$, $b = 0.25$, $c = 1.0$. (a) *Construction of the shock solution. Curve* 1 *is the power series solution around the hole in the wall;* n_L *must lie on this curve to the right of the hole. Curve* 2 *is the power series solution around* $(1, 0)$. *Curve* 3 *is equidistant below the wall as curve* 1 *is above the wall, defined to the right of the hole. The jump is located at the point of intersection of curves* 2 *and* 3. (b) *The solution profiles for* n *and* g *versus* z*; the discontinuity in* n *at* $z = 0$ *is highlighted with a thicker line.*

rate such that smooth solutions no longer exist and a travelling wave with a shock exists. Alternatively, for fixed a and b, there is a minimum wave speed such that smooth travelling wave solutions exist for $c > c_{crit}$. We have also found that if c is decreased further, travelling shock wave solutions exist for $c_{min} < c < c_{crit}$. The c_{min} is the value which determines the trajectory which jumps directly to the steady state $(1, 0)$. For $0 < c < c_{min}$, no smooth or nonsmooth trajectories exist, since to the right of the hole the distance between the wall and the trajectory through the hole is greater than the distance between the wall and the g-axis.

5.2. Production rate of g satisfies $b > 1$. We now turn to increasing b. Within the range $0 < b < 1$, the qualitative behavior of the phase plane is the same as outlined here with $b = 0.25$. It remains qualitatively similar when $b = 1$, although now all trajectories (except the one along the n-axis) leave the point $(0, 1)$ horizontally. In Figure 5.2(a) there is a smooth trajectory corresponding to a travelling wave, and in Figure 5.2(b) there will be a trajectory with a jump corresponding to a travelling shock wave. However, as b increases through unity, the steady state $(0, 1)$ changes from an unstable node to a saddle. The only outgoing trajectory emanating from this point is along the n-axis; hence there is no trajectory joining this point to $(1, 0)$. However, at the same time a new steady state moves into the positive quadrant, namely $(1 - \frac{1}{b}, 1)$, which is an unstable node if it lies below the wall (i.e., if $c^2 > 1 - \frac{1}{b}$), and is a saddle if it lies above the wall (i.e., if $c^2 < 1 - \frac{1}{b}$). We wish to determine whether a trajectory joining $(1 - \frac{1}{b}, 1)$ and $(1, 0)$ exists and whether it corresponds to smooth travelling wave solutions.

5.2.1. Sufficiently large wave speed: $c^2 > 1 - \frac{1}{b}$. In this case, both steady states are below the wall. Figure 5.3(a) shows that there is a trajectory connecting these two states. Again, when a is increased as illustrated in Figure 5.3(b), these two states can be connected by a trajectory passing through a hole in the wall, allowing a jump discontinuity in n with the wall lying at the midpoint of the jump. Hence again there is a transition from smooth travelling solutions to travelling solutions with a discontinuity.

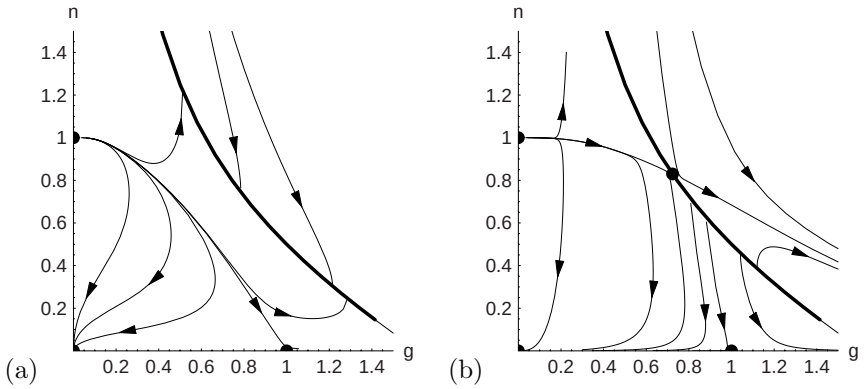


FIG. 5.2. *Phase plane for* (g, n) *for increasing values of cellular growth rate* a*. Here* $b = 1.0$*,* $c = 1.0$*. Wall, holes, and steady states marked as indicated previously.* (a) $a = 0.5$*, no holes.* (b) $a = 3.0$*, one hole.*

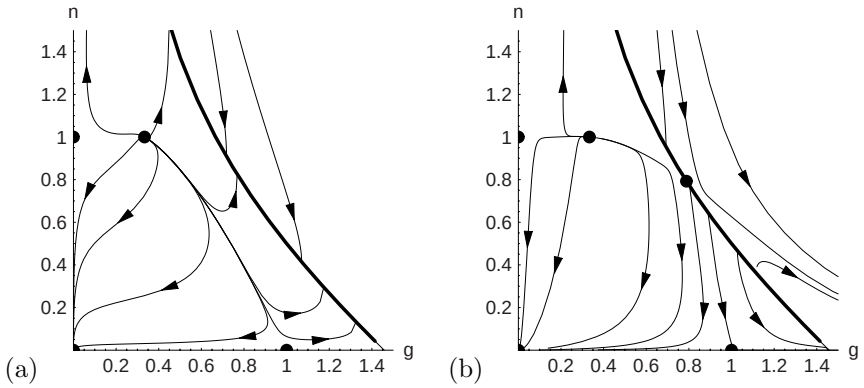


FIG. 5.3. *Phase plane for* (g, n) *for sufficiently large wave speed and increasing values of cellular growth rate* a*. Here* $b = 1.5$*,* $c = 1.0$*, and so* $c^2 > 1 - \frac{1}{b}$*. Wall, holes, and steady states marked as indicated previously.* (a) $a = 0.5$*, no holes.* (b) $a = 3.0$*, one hole.*

5.2.2. Sufficiently small wave speed: $c^2 < 1 - \frac{1}{b}$. In this case, both steady states are separated by the wall. The only possible way to connect them would be with a trajectory passing through a hole in the wall. In the two examples shown in Figure 5.4 (and in others we have tried), there appears to be no connection between these states. To understand the phase plane figures, it is useful to use a coordinate transformation similar to that employed by Pettet, McElwain, and Norbury [19], namely,

$$\frac{d}{dZ} = \left[1 + \frac{g}{c^2}(b(1-g) - 2n)\right] \frac{d}{dz}$$

when $[1 + \frac{g}{c^2}(b(1-g) - 2n)] \neq 0$. Now the holes in the wall become new steady states, and the wall becomes a g-nullcline. Figure 5.5 gives the transformed phase plane corresponding to the examples in Figure 5.4. In Figure 5.5(a), the hole becomes a stable spiral, and the one on the g-axis is a stable node. It appears that there is a limit cycle in this example. Since there is no physical or biological interpretation of Z, unlike $z = x - ct$, it is not fruitful to pursue the limit cycle analysis here. In

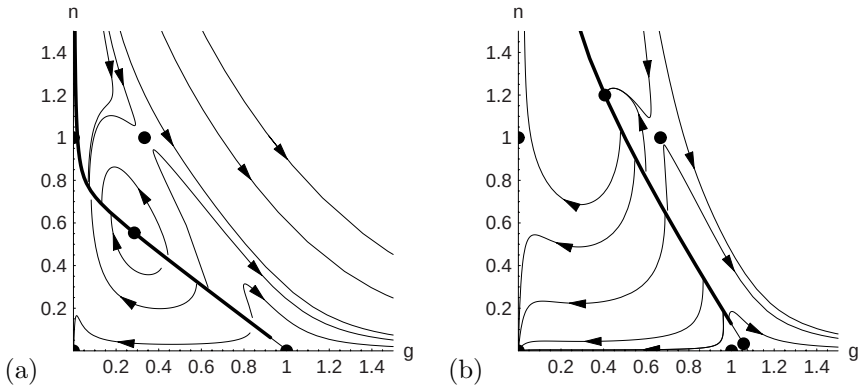


FIG. 5.4. *Phase plane for (g, n) for sufficiently small wave speed and $b > 1$. Here $a = 3.0$, and both examples satisfy $c^2 < 1 - \frac{1}{b}$. Wall, holes, and steady states marked as indicated previously.* (a) *$b = 1.5$, $c = 0.1$, one hole.* (b) *$b = 3.0$, $c = 0.5$, two holes.*

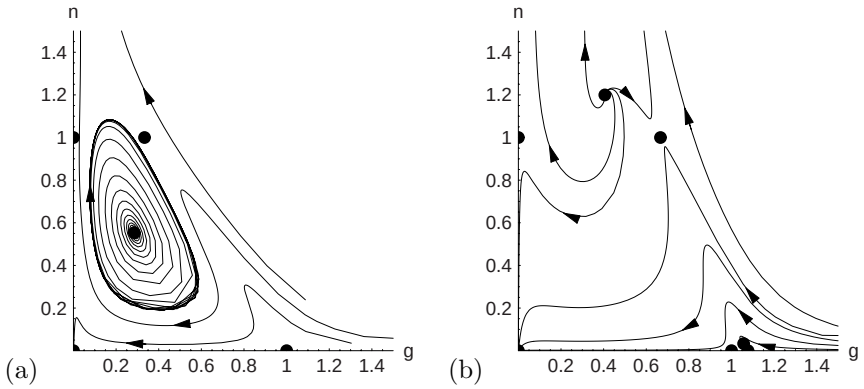


FIG. 5.5. *Z phase plane for (g, n) for sufficiently small wave speed and $b > 1$. Here $a = 3.0$, and both examples satisfy $c^2 < 1 - \frac{1}{b}$. Steady states marked as indicated previously. The parameter values are as for Figure* 5.4. (a) *$b = 1.5$, $c = 0.1$.* (b) *$b = 3.0$, $c = 0.5$.*

Figure 5.5(b), one hole becomes an unstable spiral and the other a stable node, while the one on the g-axis is a saddle.

5.3. Production rate of g satisfies $b = 0$. In this case, the chemoattractant has no production term. Since this degenerate case is similar to other recent work [11], [12], [17], we just summarize the results. When $b = 0$, the steady state $(0, 0)$ is replaced by a continuum of steady states $(\bar{g}, 0)$. Smooth travelling waves exist between $(0, 1)$ and $(\bar{g}, 0)$ for $\bar{g} < \bar{g}_{crit}$, as shown in Figure 5.6. Using the power series method about the hole, we find that there are two values of α_1, resulting in two trajectories through the hole in the wall. One of these passes through the n-axis at $n = 1$, and the other passes through the g-axis at $g = \bar{g}$, thus defining $\bar{g}_{crit}$. For $\bar{g}_{crit} < \bar{g} < \bar{g}_{max}$, the connecting trajectory passes through a hole in the wall and has a jump in it, satisfying the jump condition. The maximum value $\bar{g}_{max}$ is defined as the value of g when the trajectory through the hole jumps directly to a point on the g-axis, that is, $n_R = 0$.

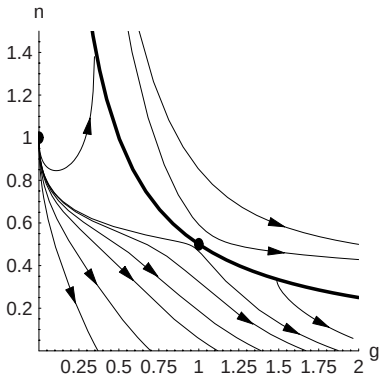


FIG. 5.6. *Phase plane for* (g, n) *for degenerate case* $b = 0$. *Here* $a = 0.5$ *and* $c = 1$. *Wall, holes, and steady states marked as indicated previously.*

6. Asymptotic analysis for large wave speed *c*. We have demonstrated that there is a minimum wave speed for a smooth travelling wave to exist. The phase plane analysis does not actually allow for a calculation of the actual solution. It can be found numerically on a finite domain, but the numerical solution will always tend to the *minimum* wave speed (see Figure 2.1). Here we investigate the analytic form of the solution for large wave speed c, following Canosa [3]. We introduce a new space variable as $z = c\xi$ and $\epsilon = \frac{1}{c^2}$ into (2.11)–(2.12) and obtain

$$[1 + \epsilon g(b(1-g) - 2n)] \frac{dn}{d\xi} = \epsilon n g \left[b(1-2g) - n\right] \left[b(1-g) - n\right] - an(1-n), \tag{6.1}$$

$$\frac{dg}{d\xi} = -(bg(1-g) - ng). \tag{6.2}$$

For small ϵ we look for an asymptotic expansion of the solution in terms of ϵ as

$$n = n_0 + \epsilon n_1 + \epsilon^2 n_2 + \cdots, \tag{6.3}$$

$$g = g_0 + \epsilon g_1 + \epsilon^2 g_2 + \cdots. \tag{6.4}$$

Here we investigate the lowest order terms n_0 and g_0 of the solution. These satisfy

$$\frac{dn_0}{d\xi} = -an_0(1 - n_0), \tag{6.5}$$

$$\frac{dg_0}{d\xi} = -bg_0(1 - g_0) + n_0 g_0. \tag{6.6}$$

The n_0 equation is decoupled and can be solved as

$$n_0 = (1 + e^{a\xi})^{-1}, \tag{6.7}$$

where the integration constant has been chosen without any loss of generality so that $n_0(0) = \frac{1}{2}$. Equation (6.6) can be solved analytically, but it is in terms of hypergeometric functions, which is not very helpful. For $b = 0$, the solution is simply

$$g_0 = (1 + e^{-a\xi})^{-1/a}, \tag{6.8}$$

so that $n_0 = 1 - g_0^a$. We see in Figure 6.1 that the numerical solution for g_0 when $0 < b < 1$ differs only slightly from the solution when $b = 0$, and all solutions for

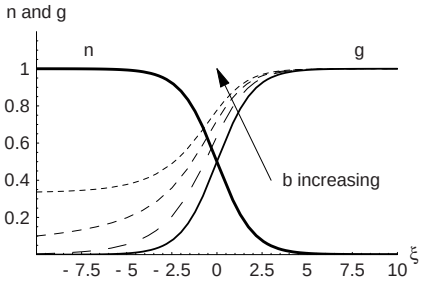


FIG. 6.1. *The lowest order solution profiles for n and g versus $\xi = z/c$ for large wave speed c. Here $a = 1.0$. The g curves depend on its production rate b; here $b = 0, 0.5, 1.0,$ and 1.5.*

g_0 tend to zero as $\xi \to -\infty$. For $b > 1$, g_0 tends to $1 - \frac{1}{b}$. Increasing the cellular growth rate a steepens the gradient of the front for n_0, as expected from our previous analysis. In particular, the slope $-\frac{\partial n}{\partial \xi}(0) = a$, and so it increases linearly with a and is independent of b to lowest order.

7. Conclusions. In this article we have considered a simple mathematical model of cell migration, where the dominant mechanism driving the migration is the phenomenon of chemotaxis. Such models have been explored in a number of contexts, generally with the view to seeking travelling wave solutions that may in some way characterize the front of invasion.

We have explored the possibility of the existence of both smooth-fronted and shock-fronted travelling wave solutions to a general model of chemotactic cell migration. Not surprisingly the mathematical model supports a rich variety of solutions exhibiting a family of identifiable characteristic behaviors such as shock-fronted travelling wave solutions with lower wave speeds than the smooth-fronted waves.

We have shown that for the model of chemotactic migration considered, for fixed a and $0 < b < 1$, there is a minimum wave speed such that smooth travelling wave solutions exist between the steady states $(0, 1)$ and $(1, 0)$ for $c > c_{crit}$. We have also found that if c is decreased further, travelling shock wave solutions exist for $c_{min} < c < c_{crit}$. The c_{min} is the value which gives the trajectory that jumps directly to the steady state $(1, 0)$. For $0 < c < c_{min}$, no smooth or nonsmooth trajectories exist, since to the right of the hole in the wall, the distance between the wall and the trajectory through the hole is greater than the distance between the wall and the g-axis. A similar argument holds for the case of a sufficiently large wave speed c, such that for fixed b and c it is possible to increase the cellular growth rate a through a critical value so that there is a transition from the existence of a smooth-fronted to a sharp-fronted travelling wave.

When $b > 1$, similar critical wave speeds can be found defining smooth and discontinuous travelling wave solutions between the steady states $(1 - \frac{1}{b}, 1)$ and $(1, 0)$, but only if the steady states are both on the same side of the wall. If the wall separates these two steady states, then no travelling wave solutions appear to exist.

The wave speeds described above are dimensionless. In terms of dimensioned variables they correspond to $\frac{Tv}{L} = \frac{v}{\sqrt{\chi \lambda_3 k_1 k_2}}$, using (2.5)–(2.6) and where v is dimensioned wave speed. Hence, converting bounds on c to bounds on χ, smooth travelling wave solutions exist for $0 < \chi < \chi_{crit}$, and travelling shock solutions exist for $\chi_{crit} < \chi < \chi_{max}$. For $\chi > \chi_{max}$, no smooth or nonsmooth trajectories exist.

These results have been obtained by considering a phase plane analysis for a coupled system of equations (2.11)–(2.12), along with power series solutions around some special points. It should be noted that, near the wall, this system is a singularly perturbed system, since then the coefficient multiplying the derivative in (2.11) is small and vanishes identically on the wall. Hence at the wall the system reduces to a differential-algebraic system. The dynamics near the wall could be decomposed into fast and slow times. Solutions consist of outer segments away from the wall and inner segments near the wall; matching occurs at holes in the wall. Such an approach could be used to establish our results theoretically. This is beyond the scope of this current paper, which is concerned with establishing the possibility of both smooth-fronted and shock-fronted travelling wave solutions to a general model of chemotactic cell migration.

Typically, migrating cell populations in invasive phenomena are characterized by a well-defined boundary. A nonlinear diffusive flux can capture this feature, but does not allow for shock solutions. Of particular interest here, though, is the possibility of representing sharp fronts to waves of invading cells by the simple chemotactic term, without the need for nonlinear diffusion. As presented here, the chemotactic cell migration model (2.1)–(2.2) supports jump discontinuities when $c_{min} < c < c_{crit}$. However, solutions with compact support exist only for $c = c_{min}$, whereas for $c_{min} < c < c_{crit}$ there is a smooth leading segment of the front ahead of the discontinuity. We anticipate that the minimum invasion speed solution (with c_{min}) will evolve as the stable solution, using a hyperbolic numerical solver, as indicated by Marchant and Norbury [13]. Furthermore, the numerical solutions in Figure 2.1 suggest that this may be the case, since as χ increases, corresponding to c decreasing, the leading edge contracts. We are presently tackling these issues.

When cell population numbers are sufficiently dense to imply frequent and significant cell-cell interactions, the suitability of a diffusive flux term comes into question. In such circumstances, mathematical models of cell migration in which a chemotactic flux dominates over the diffusive flux are more appropriate. Being essentially hyperbolic in nature, these models have the potential to support shock-fronted solutions, which may be seen as a new paradigm for the representation of cell migration.

Acknowledgments. We would like to thank Dr. Mark McGuinness, at Victoria University of Wellington, for initial discussions on the model. We also thank the reviewers for their helpful comments and suggestions.

REFERENCES

[1] H. M. Byrne, M. A. J. Chaplain, G. J. Pettet, and D. L. S. McElwain, *A mathematical model of trophoblast invasion,* J. Theoret. Med., 1 (1999), pp. 275–286.

[2] H. M. Byrne, M. A. J. Chaplain, G. J. Pettet, and D. L. S. McElwain, *An analysis of a mathematical model of trophoblast invasion*, Appl. Math. Lett., 14 (2001), pp. 1005–1010.

[3] J. Canosa, *On a nonlinear diffusion equation describing population growth,* IBM J. Res. Dev., 17 (1973), pp. 307–313.

[4] R. Courant and D. Hilbert, *Methods of Mathematical Physics. Vol.* II, 2nd ed., Interscience, New York, 1964.

[5] C. M. Elliott, M. A. Herrero, J. R. King, and J. R. Ockendon, *The mesa problem—Diffusion of patterns for* $u_t = \nabla \cdot (u^m \nabla u)$ *as* $m \to \infty$, IMA J. Appl. Math., 37 (1986), pp. 147–154.

[6] R. M. Ford and P. T. Cummings, *Mathematical models of bacterial chemotaxis,* in Mathematical Modeling in Microbial Ecology, A. L. Koch, J. A. Robinson, and G. A. Milliken, eds., Chapman and Hall, New York, 1998, pp. 228–269.

[7] P. Grindrod, *Patterns and Waves*, Clarendon Press, Oxford, UK, 1991.

[8] T. Hillen, *Hyperbolic models for chemosensitive movement,* Math. Models Methods Appl. Sci., 12 (2002), pp. 1007–1034.
[9] E. F. Keller and L. A. Segel, *Travelling bands of chemotactic bacteria: A theoretical analysis,* J. Theoret. Biol., 30 (1971), pp. 235–248.
[10] K. A. Landman, G. J. Pettet, and D. F. Newgreen, *Mathematical models of cell colonisation of uniformly growing domains,* Bull. Math. Biol., 65 (2003), pp. 235–262.
[11] B. P. Marchant, *Modelling Cell Invasion*, Ph.D. thesis, University of Oxford, Oxford, UK, 1999.
[12] B. P. Marchant, J. Norbury, and A. J. Perumpanani, *Travelling shock waves arising in a model of malignant invasion,* SIAM J. Appl. Math., 60 (2000), pp. 463–476.
[13] B. P. Marchant and J. Norbury, *Discontinuous travelling wave solutions for certain hyperbolic systems,* IMA J. Appl. Math., 67 (2002), pp. 201–224.
[14] J. D. Murray and M. R. Myerscough, *Pigmentation pattern formation on snakes,* J. Theoret. Biol., 149 (1991), pp. 339–360.
[15] J. D. Murray, *Mathematical Biology,* 2nd ed., Springer-Verlag, Heidelberg, 1993.
[16] H. G. Othmer and A. Stevens, *Aggregation, blowup, and collapse: The ABC's of taxis in reinforced random walks,* SIAM J. Appl. Math., 57 (1997), pp. 1044–1081.
[17] A. J. Perumpanani, J. A. Sherratt, J. Norbury, and H. M. Byrne, *A two parameter family of travelling waves with a singular barrier arising from the modelling of extracellular matrix mediated cell invasion,* Phys. D., 126 (1999), pp. 145–159.
[18] G. J. Pettet, H. M. Byrne, D. L. S. McElwain, and J. Norbury, *A model of wound-healing angiogenesis in soft tissue,* Math. Biosci., 136 (1996), pp. 35–63.
[19] G. J. Pettet, D. L. S. McElwain, and J. Norbury, *Lotka–Volterra equations with chemotaxis: Walls, barriers and travelling waves,* IMA J. Math. Appl. Med. Biol., 17 (2000), pp. 395–413.
[20] F. Sanchez-Garduno and P. K. Maini, *Existence and uniqueness of a sharp front travelling wave in degenerate nonlinear diffusion Fisher–KPP equations,* J. Math. Biol., 33 (1994), pp. 163–192.
[21] J. Smoller, *Shock waves and reaction-diffusion systems,* 2nd ed., Springer-Verlag, New York, 1994.
[22] R. T. Tranquillo, *Perspectives and models of gradient perception,* in Biology of the Chemotactic Response, J. P. Armitage and J. M. Lackie, eds., Cambridge University Press, Cambridge, UK, 1991, pp. 35–75.
[23] R. T. Tranquillo and W. Alt, *Receptor-mediated models for leukocyte chemotaxis,* in Dynamics of Cell and Tissue Motion, W. Alt, A. Deutsch, and G. Dunn, eds., Birkhäuser, Berlin, 1997, pp. 141–147.
[24] T. P. Witelski, *Stopping and merging problems for the porous media equation,* IMA J. Appl. Math., 54 (1995), pp. 227–243.
[25] T. P. Witelski, *The structure of internal layers for unstable nonlinear diffusion equations,* Stud. Appl. Math., 97 (1996), pp. 277–300.

SIAM J. APPL. MATH.
Vol. 63, No. 5, pp. 1682–1717

FORMAL ASYMPTOTICS OF BUBBLING IN THE HARMONIC MAP HEAT FLOW*

JAN BOUWE VAN DEN BERG†, JOSEPHUS HULSHOF†, AND JOHN R. KING‡

Abstract. The harmonic map heat flow is a model for nematic liquid crystals and also has origins in geometry. We present an analysis of the asymptotic behavior of singularities arising in this flow for a special class of solutions which generalizes a known (radially symmetric) reduction. Specifically, the rate at which blowup occurs is investigated in settings with certain symmetries, using the method of matched asymptotic expansions. We identify a range of blowup scenarios in both finite and infinite time, including degenerate cases.

Key words. harmonic maps, blowup, bubbling, matched asymptotics, nematic liquid crystal

AMS subject classifications. 35B35, 35B40, 35B60, 58E20, 74H35, 76A15

DOI. 10.1137/S0036139902408874

1. Introduction.

We consider the equation

$$\theta_t = \theta_{rr} + \frac{1}{r}\theta_r - \frac{\sin 2\theta}{2r^2}, \qquad 0 < r < 1, \tag{1.1}$$

with boundary conditions $\theta(t,0) \in \pi\mathbb{Z}$ and $\theta(t,1) = \theta_1 \in \mathbb{R}$; the reason for this type of boundary condition at $r = 0$ will become clear shortly. Solutions of (1.1) may develop a singularity. In this paper we analyze this blowup behavior using formal matched asymptotics.

Equation (1.1) is a special case of the harmonic map heat flow

$$\frac{\partial u}{\partial t} = \Delta u + |\nabla u|^2 u, \tag{1.2}$$

where $u(t,\cdot) : \Omega \to S^2$; i.e., $u(t,x)$ denotes a unit vector in $\mathbb{R}^3$, $\Omega \subset \mathbb{R}^N$ (in most physical models $N = 3$), and $|\nabla u|^2 = \sum_{j=1}^{N} \sum_{i=1}^{3} (\frac{\partial u_i}{\partial x_j})^2$. Stationary solutions of (1.2) are harmonic maps from Ω to S^2.

Observe that we are dealing with blowup of the derivative ∇u while u remains bounded (in fact, $|u(t,x)| = 1$ for all t and x), and similarly θ_r blows up while θ remains bounded. (As we shall see, θ can make finite jumps.) This stands in contrast to many widely studied blowup problems, such as the reaction-diffusion equation $u_t = \Delta u + u^p$ with $p > 1$, where u itself blows up (see, e.g., [8] and references therein).

Equation (1.2) may be reduced to (1.1) if Ω is a disk in $\mathbb{R}^2$: assuming the solution to be radially symmetric and using polar coordinates (r,ϕ) on the unit disk, a special type of solution of (1.2) is given by

$$u(t,\cdot) : (r,\phi) \to \begin{pmatrix} \cos\phi \, \sin\theta(t,r) \\ \sin\phi \, \sin\theta(t,r) \\ \cos\theta(t,r) \end{pmatrix}, \tag{1.3}$$

*Received by the editors June 3, 2002; accepted for publication (in revised form) February 28, 2003; published electronically July 26, 2003. This work was partly supported by the EPSRC and the TMR network Fronts-Singularities.
http://www.siam.org/journals/siap/63-5/40887.html
†Department of Mathematical Analysis, Vrije Universiteit Amsterdam, De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands (janbouwe@cs.vu.nl, jhulshof@cs.vu.nl).
‡Theoretical Mechanics, Division of Applied Mathematics, University of Nottingham, Nottingham NG7 2RD, UK (John.King@nottingham.ac.uk).

where $\theta(t,r)$ satisfies (1.1). Similarly, when Ω is a cylinder, then the problem of finding solutions of (1.2) that are both radially symmetric and uniform in the axial direction may be reduced to (1.1). This is the configuration studied in [11] as a model for aligned nematic liquid crystals, with the motivation coming from applications in fiber spinning. Beside the context of liquid crystals (see, e.g., [14]), another application in which (1.2) appears is the theory of ferromagnetic materials (e.g., [6, 3]). In geometry, (1.2) is studied in the construction of harmonic maps of certain homotopy types (see, e.g., [15]), where Ω is generally (a subset of) an N-manifold. (The target manifold may also differ from S^2, but if it is not a sphere, (1.2) is altered.) The formation of singularities in the flow of (1.2) has been extensively studied; we refer to [15, 16, 9]. Singularities occur due to topological obstructions, a situation which is comparable to closely related problems in the Ginzburg–Landau equation (see, e.g., [4]). In the present context, the issue is that while all solutions eventually converge to equilibria, we can choose initial data in a topological class that does not contain any equilibria. The solution must then "jump" to another topological class. The different topological classes are (for fixed $\theta(t,1)=\theta_1$) characterized by the value of θ at the origin, which has to be a multiple of π for the solution to have finite energy (see below).

We recall some important results, where in view of (1.1) we concentrate on domains $\Omega \subset \mathbb{R}^2$. Equation (1.2) is the gradient flow associated with the energy $\mathcal{E} = \frac{1}{2}\int_\Omega |\nabla u|^2\,dx$. It is well known that a weak solution of (1.2) exists globally. There may be many weak solutions, but there is a unique one in the class of energy-decreasing solutions; see [7, 15]. Weak solutions are in $H^1(\Omega, S^2)$ for almost all $t>0$, and, when the initial data are smooth, the solution is locally a classical solution. In fact, the solution is smooth everywhere except for at most a finite number of space-time points [15]. Moreover, there are smooth initial conditions for which a singularity occurs in finite time [5, 2]. As $t\to\infty$ the solution converges weakly to a stationary solution and does so smoothly away from at most a finite number of points. At a singularity, either in finite time or at $t=\infty$, a sphere is said to bubble off: an appropriate blowup near the singularity converges to a harmonic map on the sphere $S^2 \cong \overline{\mathbb{R}^2}$ (see [15]). In this paper we investigate the rate at which these spheres bubble off in the symmetric setting of (1.1). In the context of liquid crystals this bubbling means that quanta of energy (i.e., a multiple of 4π) are stored in a singularity (a region smaller than that captured by the model).

Let us now concentrate on the implications for (1.1). The (weak) solution $\theta(t,\cdot)$ is continuous on $[0,1]$ and $\theta(t,0)\in\pi\mathbb{Z}$ for all $t>0$. The requirement that $\theta(t,0)\in\pi\mathbb{Z}$ is necessary for solutions to have finite energy. Singularities can develop only at the origin. At a singularity the energy

$$\mathcal{E}(t) = \pi\int_0^1 \left(r\,\theta_r^2 + \frac{\sin^2\theta}{r}\right) dr \tag{1.4}$$

decreases (jumps) by 4π or a multiple thereof. (Of course, away from such singularities, the energy $\mathcal{E}(t)$ decreases continuously throughout the evolution, since (1.1) is the gradient flow associated with (1.4).) Notice that the stationary solutions of (1.1) with finite energy $\mathcal{E}$ are given by

$$\theta(r) = m\pi + 2\arctan qr \qquad \text{for any } q\in\mathbb{R} \text{ and } m\in\mathbb{Z},$$

and their energy tends to 4π as $q\to\infty$. The solutions $\theta(r)=(m+\frac{1}{2})\pi$, $m\in\mathbb{Z}$, have infinite energy and can be disregarded.

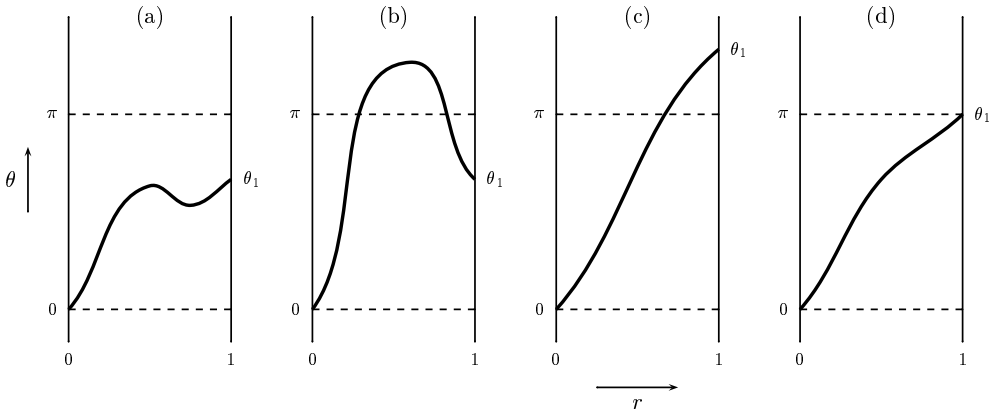


FIG. 1.1. *Several initial conditions* $\theta(0,r)$*:* (a) *no blowup will occur;* (b) *blowup may occur;* (c) *blowup must occur;* (d) *the degenerate case* $\theta_1 = \pi$*, which leads to infinite time blowup, as opposed to finite time blowup for* $\theta_1 > \pi$*.*

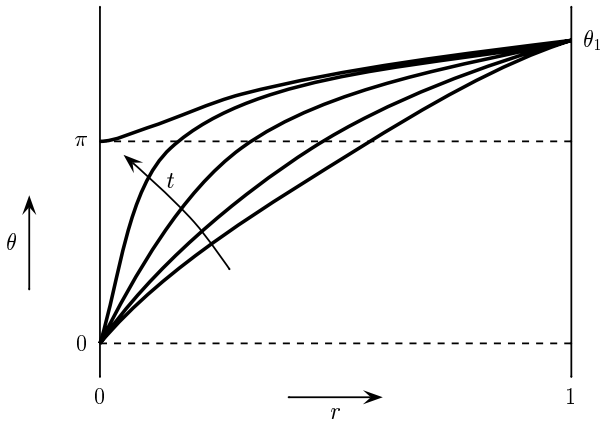


FIG. 1.2. *The profile at several times leading up to blowup at* $t = T$*. The slope* $\theta_r(0)$ *goes to infinity, and the solution jumps from* 0 *to* π *at the origin.*

Whether or not singularities occur will depend on the initial and boundary data (see also Figure 1.1). Consider initial conditions such that $|\theta(0,0)-\theta_1| < \pi$. Then the solution may converge to one of the stationary states as $t \to \infty$ without forming a singularity. This is indeed what happens when $\theta(0,r) \in C^1$ and $\|\theta(0,r) - \theta(0,0)\|_\infty \leq \pi$. On the other hand, it has been proved in [2] that blowup may occur for initial data with $|\theta(0,0) - \theta_1| < \pi$ but $|\theta(0,0) - \theta(0,r)| > \pi$ for some $r \in (0,1)$. A more dramatic situation occurs when $|\theta(0,0) - \theta_1| \geq \pi$: no stationary solution is available that obeys both boundary conditions (i.e., $\theta(0) = \theta(0,0)$ and $\theta(1) = \theta_1$). Therefore, for the solution to approach any of the stationary solutions a jump at the origin must necessarily occur. This is depicted in Figure 1.2. We focus on this case, and after shifting θ by a multiple of π, we may restrict our attention to initial/boundary data $\theta(0,0) = 0$ and $\theta(t,1) = \theta_1 \geq \pi$. Without loss of generality we will analyze the first instance of blowup.

As mentioned before, when blowup occurs, appropriately zooming in on the singularity will reveal a harmonic map, i.e., one of the stationary solutions. We first focus on boundary data $\theta_1 > \pi$ and consider the special case $\theta_1 = \pi$ later. Assuming the jump of $\theta(t,0)$ to be upwards (without loss of generality), we select a zooming function $R(t) > 0$ by requiring that

$$R(t)\theta_r(t,0) = 2 \qquad \text{for all } t \text{ up to the blowup time } T \in (0,\infty].$$

We choose the constant in the right-hand side to be 2 in order to keep the subsequent algebra as simple as possible. When blowup occurs at $t = T$, we will thus have

$$\lim_{t\uparrow T} \theta(t, \rho R(t)) = 2\arctan\rho \qquad \text{for all fixed } \rho > 0.$$

The main objective of this paper is to determine the asymptotic form of $R(t)$.

At first sight one might think that the blowup rate simply corresponds to self-similar variables, i.e., $R(t) \sim \kappa\sqrt{T-t}$. However, no suitable self-similar solution exists. (We postpone justification of this statement until the end of this section.) Indeed, we find that the blowup rate is not the self-similar one. Using formal asymptotics, we match an inner layer near $r = 0$ to an outer region near $\theta = \pi$ (where the solution is approximately self-similar), which in turn matches into the remote region where $r = O(1)$. We find that generically

$$R(t) \sim \kappa \frac{T-t}{|\ln(T-t)|^2} \qquad \text{as } t \uparrow T \tag{1.5}$$

for some blowup time $T > 0$ and some constant $\kappa > 0$. That there is an unknown constant κ is a consequence of the fact that the profile in the remote region plays a subdued role. Therefore, in spite of the finiteness of the domain, the scaling invariance $(t,r) \mapsto (\mu^2 t, \mu r)$ with $\mu > 0$ of (1.1) causes an indeterminacy.

The point $S(t)$, the smallest intersection of $\theta(t,r)$ with π, behaves as

$$S(t) \sim 2\sqrt{\frac{T-t}{|\ln(T-t)|}} \qquad \text{as } t \uparrow T.$$

In particular, in this scenario the solution always intersects π close to blowup. There is no undetermined constant in this asymptotic expression for $S(t)$ because it is (to leading order) invariant under the scaling invariance. The limit profile at $t = T$ for small r becomes

$$\theta(T,r) \sim \pi + \frac{1}{4}\kappa \frac{r}{|\ln r|} \qquad \text{for small } r,$$

with the same constant $\kappa > 0$ as in (1.5). We remark that there may be additional blowup times $T' > T$, for example when $|\theta(0,0) - \theta_1| > 2\pi$, and that $\theta(t,0)$ can jump only by $\pm\pi$ at a time (and thus the energy by 4π); see [12].

A nongeneric case arises when $\theta_1 = \pi$ and, for example, $-\pi < \theta(0,r) < \pi$ for $r \in (0,1)$. In that case the asymptotics indicate blowup in infinite time:

$$R(t) \sim e^{-2\sqrt{t}-5/4} \qquad \text{as } t \to \infty.$$

Notice that the blowup is now in infinite time as opposed to finite time for $\theta_1 > \pi$. Besides, there is no undetermined constant in the leading order term for $R(t)$; the

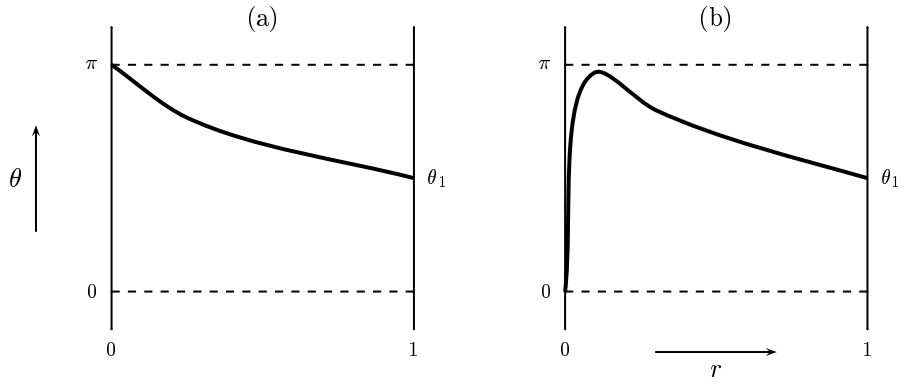


FIG. 1.3. *In a reverse jump the energy of the solution increases:* (a) *just before* t_0, (b) *just after* t_0.

length of the interval has a direct bearing on the analysis so that the scaling invariance is lifted. We remark that when $\theta(0,r) > \pi$ for some $r \in (0,1)$, then it can (but does not necessarily) happen that blowup occurs in finite time via the scenario described before. In that case the solution intersects π just before blowup. On the other hand, for initial profiles with $-\pi < \theta(0,r) < \pi$ a comparison argument shows that this is not possible.

A different situation in which we can easily see that nongeneric behavior must occur is when $\theta_1 < \pi$, and the initial data are roughly as depicted in Figure 1.1(b). When the initial profile has a sufficiently large bump above π, the solution will blow up in finite time via the scenario described above. On the other hand, when the bump is small (e.g., stays below π), no blowup occurs. In between these generic (codimension 0) possibilities there needs to be at least one borderline (nongeneric) scenario, and such degenerate cases are also discussed below. Regarding the large time behavior of these solutions, in the latter generic case (when no blowup occurs) the limit profile as $t \to \infty$ is $\theta_\infty(r) = 2\arctan\bigl(r\tan\frac{\theta_1}{2}\bigr)$, while in the former case the stationary state $\theta_\infty(r) = \pi - 2\arctan\bigl(r\tan\frac{\pi-\theta_1}{2}\bigr)$ is selected (provided that no additional jump back to 0 occurs), and it turns out this last stationary state is also the limit profile in the degenerate scenario (if no additional jumps occur).

There is another issue related to these singularities. As explained in [2, 18, 5], weak solutions have the possibility of releasing the energy formerly lost in a singularity, thereby causing a sudden *increase* in the energy $\mathcal{E}(t)$. The physical interpretation is that this released energy was stored in a region of smaller scale than that captured by the model. We consider the situation where θ makes such a "reverse" jump at $t = t_0$; see also Figure 1.3. When θ jumps from π to 0 at $t = t_0$, we define as before $R(t) = \frac{2}{\theta_r(t,0)}$ for $t > t_0$. One finds that generically

$$R(t) \sim \kappa \frac{t - t_0}{|\ln(t - t_0)|} \qquad \text{as } t \downarrow t_0$$

for some arbitrary constant $\kappa > 0$. Notice the slight difference with (1.5).

Reexamining the reduction of (1.2) to (1.1), the physical meaning of the ansatz (1.3) is that the direction field $u(t,\cdot)$ at the boundary of the cylinder is axisymmetric and the in-plane component points in the radial direction. In fact, the solution class

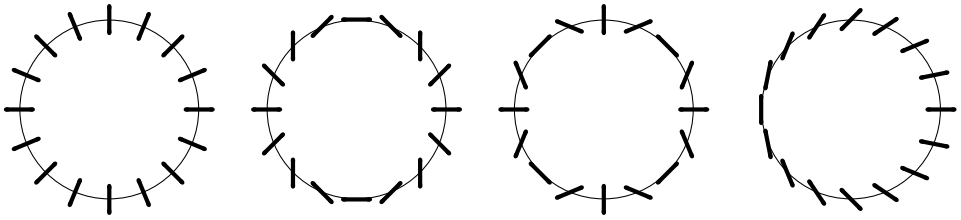


FIG. 1.4. *Top view of the behavior of the vector u at the boundary of the domain for $n = 1$, $n = 2$, $n = 3$, and $n = \frac{1}{2}$, the last one necessarily leading to configurations with infinite energy.*

defined by (1.3) belongs to a family of solution classes given by

$$u(t,\cdot) : (r,\phi) \to \begin{pmatrix} \cos n\phi \, \sin\theta(t,r) \\ \sin n\phi \, \sin\theta(t,r) \\ \cos\theta(t,r) \end{pmatrix}. \tag{1.6}$$

The equation for θ now becomes

$$\theta_t = \theta_{rr} + \frac{1}{r}\theta_r - n^2 \frac{\sin 2\theta}{2r^2}, \qquad 0 \le r \le 1. \tag{1.7}$$

From a mathematical point of view the constant $n > 0$ (ignoring the trivial case $n = 0$ throughout) can be considered as a continuous parameter in (1.7), and it can be used to unravel the delicate analysis of blowup for $n = 1$ (which is a borderline case, so that the asymptotic analysis is particularly delicate). From a physical point of view, only the values $n = 1, 2, 3, \ldots$ make sense. In Figure 1.4 the configurations for $n = 1$, 2, and 3 are depicted. We note that for $n = \frac{1}{2}$ (and odd multiples of $\frac{1}{2}$) the view from the top (see Figure 1.4) gives the impression of smoothness (because the molecules in a nematic liquid crystal are invariant under inversion, or in other words, because in (1.2) the function $u(t,\cdot)$ maps from Ω to the projective plane instead of to the sphere). However, on closer inspection, one observes that in fact a line singularity with infinite energy is unavoidable, and hence such cases fall outside the scope of the present paper.

The stationary solutions of (1.7) with finite energy are

$$\theta(r) = m\pi + 2\arctan(qr^n), \qquad m \in \mathbb{Z},\ q \in \mathbb{R}.$$

We define $R(t)$ such that

$$R(t)^n \theta(t,r) \sim 2r^n \quad \text{as } r \downarrow 0 \qquad \text{for all } t \text{ up to the blowup time.}$$

After rescaling with this blowup rate, the profile tends to a harmonic map (a stationary state) as t approaches the blowup time T:

$$\lim_{t \uparrow T} \theta(t, \rho R(t)) = 2\arctan\rho^n \qquad \text{for all fixed } \rho > 0.$$

The results of our asymptotic analysis for $R(t)$ give the following as the generic blowup

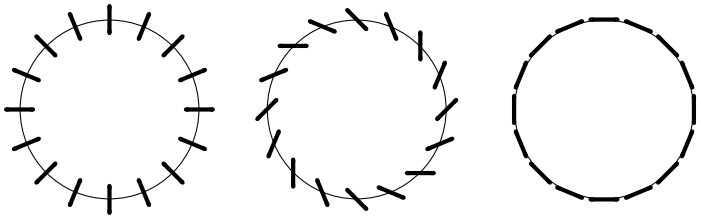


FIG. 1.5. *Top view of the behavior of the vector u at the boundary of the domain for $n = 1$ with $\phi_0 = 0$, $\phi_0 = \frac{\pi}{4}$, and $\phi_0 = \frac{\pi}{2}$.*

behavior:

$$\begin{aligned}
n < 1: &\qquad R \sim \kappa\,(T-t)^{1/n} &&\text{as } t \uparrow T,\\
n = 1: &\qquad R \sim \kappa\,\frac{T-t}{|\ln(T-t)|^2} &&\text{as } t \uparrow T,\\
1 < n < 2: &\qquad R \sim \kappa\,(T-t)^{1/(2-n)} &&\text{as } t \uparrow T,\\
n = 2: &\qquad R \sim \kappa\, e^{-\frac{\alpha_0}{E_2}t} &&\text{as } t \to \infty,\\
n > 2: &\qquad R \sim \left(\tfrac{(n-2)\alpha_0}{E_n}t\right)^{-1/(n-2)} &&\text{as } t \to \infty.
\end{aligned}$$

Here $\kappa > 0$ is an arbitrary constant, $E_n = \frac{\pi}{2n^2 \sin(\frac{\pi}{n})}$, and $\alpha_0 = \tan(\frac{\theta_1 - \pi}{2})$ for $\theta_1 \in (\pi, 2\pi)$. The above represent the generic behavior (e.g., $\theta_1 = \pi$ needs to be considered separately), and for $n \geq 2$ it does not apply to boundary conditions with $\theta_1 \geq 2\pi$; the analysis is more involved in that case (see section 3.6). Notice that the blowup is in finite time for $n < 2$, versus infinite time blowup for $n \geq 2$. Furthermore, it is remarkable that the borderline case $n = 1$ has the fastest blowup rate. Finally, there is no unknown constant for $n > 2$ since the boundary condition on the right has direct influence on the asymptotics (and thus there is no scaling invariance). This is equally true for $n = 2$, and the translation invariance in time rather than the scaling invariance can be considered responsible for the indeterminacy here. On the other hand, $n = 2$ marks the transition from finite to infinite time blowup, and subtle behavior can be expected at such a critical value.

There is an additional symmetry which needs to be noted. In the right-hand side of (1.6) one may replace ϕ by $\phi + \phi_0$, which again leads to (1.7). For $n = 1$ this presents us with a family of geometrically different solutions, while for $n \neq 1$ all these solutions are equivalent by rotation of the domain. In Figure 1.5 we have depicted the situation occurring for $\phi_0 = \frac{\pi}{4}$ and $\phi_0 = \frac{\pi}{2}$ (and $n = 1$), which may be compared to $\phi_0 = 0$ to see the difference in geometry. All these cases are covered by (1.7).

In order to prevent cumbersome bookkeeping and to be able to clarify the crucial points, we will first analyze the special case $n = 1$ in section 2. Degenerate cases, including the special boundary condition $\theta_1 = \pi$, and reverse jumps are treated in sections 2.5 to 2.7. In section 3 we analyze the general case (1.7), and the special role of $n = 1$ will become apparent. We also discuss in section 3.6 the multiscale blowup associated with boundary conditions $\theta_1 \geq 2\pi$ for $n \geq 2$; in section 3.8 we deal with the case of an unbounded domain. Finally, we present an overview of our results and draw conclusions in section 4.

It remains a challenge to find proofs for the formal asymptotic results in this paper. We refer to [1] for some tentative results in which the comparison principle and

lap number theorem for parabolic equations are exploited. These methods circumvent nondegeneracy conditions so that, while avoiding the problems associated with degenerate cases, they fail to uncover the full range of the generic behavior. Another open problem is what happens in a nonsymmetric situation, both in two and three dimensions, and what role is played by symmetric solutions in that context. The fact that $n = 1$ is such an exceptional case may suggest that it plays a special role.

To finish this section we show why blowup is not governed by self-similar variables. (This was also observed in [1].) A self-similar solution for finite time blowup is of the form $\theta(t, r) = \Theta(r/\sqrt{T-t})$. For the function $\Theta(y)$ one obtains the equation

$$\Theta_{yy} + \left(\frac{1}{y} - \frac{y}{2}\right)\Theta_y - n^2\frac{\sin 2\Theta}{2y^2} = 0.$$

Since $\theta(t, 0)$ jumps from 0 to π, the boundary conditions are $\Theta(0) = 0$, and at infinity $\lim_{y\to\infty}\Theta(y) = \pi$. We now change coordinates to $z = \ln y$ and obtain for $\tilde{\Theta}(z) = \Theta(y)$

$$\tilde{\Theta}_{zz} - \frac{1}{2}e^{2z}\tilde{\Theta}_z - \frac{n^2}{2}\sin 2\tilde{\Theta} = 0, \tag{1.8}$$

with boundary conditions $\lim_{z\to-\infty}\tilde{\Theta}(z) = 0$ and $\lim_{z\to\infty}\tilde{\Theta}(z) = \pi$. There is no solution to this problem since, on the one hand, $G(z) \stackrel{\text{def}}{=} \frac{1}{8}\tilde{\Theta}_z^2 + n^2\cos 2\tilde{\Theta}$ is monotonically increasing, while, on the other hand, $\lim_{z\to-\infty}G(z) = \lim_{z\to+\infty}G(z) = n^2$. This contradiction shows that there is no such solution, and hence this self-similar scenario for blowup cannot occur. In fact, $\tilde{\Theta} = \pi$ is the only solution satisfying the condition as $z \to \infty$; this in part explains why $\theta \sim \pi$ necessarily holds in the outer region described below. In this argument it is crucial that $\int_0^\pi \sin 2\theta\, d\theta = 0$. For nonlinearities that do not have zero average, a self-similar blowup rate may be expected. It could thus be interesting to study a problem in which the average of the nonlinearity approaches zero as a parameter is varied.

2. The case $n = 1$.

2.1. Preamble. There will be three scales: the inner, the outer, and the remote region (see Figure 2.1). The inner is a small region near the origin in which the blowup is concentrated. It is of order $r = O(R(t))$, where $R(t)$ is an unknown function (in fact, the main goal is to determine the asymptotic behavior of $R(t)$), and in this region the profile near blowup is $2\arctan(r/R(t))$. The outer region is a region with θ near π in which the equation can be linearized. The typical scale in this region is $r = O(\sqrt{T-t})$, where T is the blowup time. An obvious requirement for self-consistency is that $R(t) \ll \sqrt{T-t}$. The remote region is the region where $r = O(1)$, and at the time of blowup the profile in this region is unknown, but the limit profile as the origin is approached will come out of the matching procedure.

Throughout the paper the constants C, $\tilde{C}$, and C_i ($i \in \mathbb{N}$) will vary from subsection to subsection.

2.2. The inner approximation. We analyze the boundary layer near $r = 0$. As explained in section 1, it is known on general grounds that, when we zoom in appropriately, we should see $2\arctan(r/R(t))$, with $R(t) \to 0$ as $t \uparrow T$. Recall that the definition of R is $R(t) = \frac{2}{\theta_r(t,0)}$. We introduce a new variable $\xi = r/R(t)$ and obtain for $v(t, \xi) = \theta(t, r)$

$$R^2 v_t - R'R\,\xi v_\xi = v_{\xi\xi} + \frac{1}{\xi}v_\xi - \frac{\sin 2v}{2\xi^2}.$$

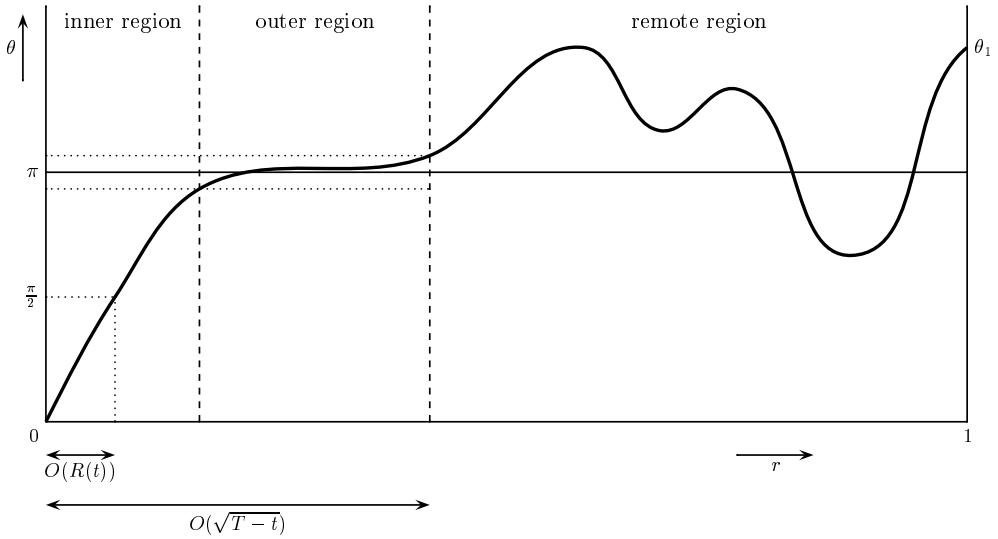


FIG. 2.1. *The three different scales.*

Since $R(t)$ becomes small for t close to blowup, we formally expand the solution in powers of $R'R$:

$$v(t,\xi) \sim \Phi_0(\xi) + R'(t)R(t)\Phi_1(\xi) + (R'(t)R(t))^2\Phi_2(\xi). \tag{2.1}$$

A motivation for this expansion is that we anticipate that $R \ll R'$ as $t \to T$. To these orders, $R^2 v_t$ does not contribute, since the rescaling is chosen such that the leading order solution is stationary. Of course we need that $R'R \to 0$ as $t \to T$.

One finds that

$$\Phi_0(\xi) = 2\arctan\xi,$$

and $\Phi_1(\xi)$ satisfies

$$\Phi_{1\xi\xi} + \frac{1}{\xi}\Phi_{1\xi} - \frac{\cos 2\Phi_0}{\xi^2}\Phi_1 = -\xi\Phi_{0\xi}.$$

It turns out that Φ_1 is already the interesting term, so that we could have restricted our analysis to linearizing around Φ_0. The equation for Φ_2 is

$$\Phi_{2\xi\xi} + \frac{1}{\xi}\Phi_{2\xi} - \frac{\cos 4\arctan\xi}{\xi^2}\Phi_2 = \big(1+K\big)\Phi_1 - \xi\Phi_{1\xi} - \frac{\sin(4\arctan\xi)}{\xi^2}\Phi_1^2, \tag{2.2}$$

where $K = \lim_{t\to T}\frac{R''R}{R'^2}$. Here we see that we need $R''R = O(R'^2)$ as $t \to T$ in order for (2.1)–(2.2) to be self-consistent. (It includes, for example, $R \sim (T-t)^a$ and $R \sim t^{-a}$ when $T = \infty$ for any $a > 0$.) This will turn out to be the case with $K = 0$.

The nonuniqueness of Φ_i is resolved by requiring Φ_i to be regular near $\xi = 0$, i.e., $\Phi_i(0) = 0$, and $\Phi_i'(0) = 0$ in view of the definition of $R(t)$. One finds

$$\Phi_1 = \frac{\xi}{1+\xi^2}\int_0^\xi \frac{s(s^4+4s^2\ln s - 1)}{(1+s^2)^2}ds - \frac{\xi^4+4\xi^2\ln\xi - 1}{\xi(1+\xi^2)}\int_0^\xi \frac{s^3}{(1+s^2)^2}ds,$$

which, after a quite tedious calculation, can be rewritten as

$$\Phi_1 = \frac{(1-\xi^4)\ln(1+\xi^2) + 2\xi^4 - \xi^2 - 4\xi^2 \int_0^\xi \frac{\ln(1+s^2)}{s} ds}{2\xi(1+\xi^2)}.$$

For large ξ the inner approximation thus satisfies

$$v(t,\xi) \sim \pi - 2\,\xi^{-1} + R'(t)R(t)(-\xi \ln \xi + \xi) \qquad \text{for } t \text{ close to } T \text{ and large } \xi. \tag{2.3}$$

Here, and in other asymptotic expansions to come, we include all those terms which might be necessary for performing the matching analysis. (It is not a priori clear whether all of them will be needed.) Let us briefly comment on the inclusion of the term $R'(t)R(t)\xi$ in this expansion. Although it is dominated by $R'(t)R(t)\xi \ln \xi$, it is well known that terms that differ only by orders of $\ln \xi$ can play a role in the matching. We remark that the leading order approximation $v \sim \pi - 2\xi^{-1} - R'(t)R(t)\xi \ln \xi$ is valid if $\xi^4 \gg \frac{1}{|R'R|}$ and $\xi^2 \ll \frac{1}{|R'R|}$, because the next terms are of order $O(\xi^{-3})$ and $O(R'^2R^2\xi^3 \ln \xi)$; this last term comes from the large ξ behavior for the solution of (2.2).

We remark that we could for most purposes have restricted our attention to the asymptotic equation for Φ_1,

$$\Phi_{1\xi\xi} + \frac{1}{\xi}\Phi_{1\xi} - \frac{1}{\xi^2}\Phi_1 \sim -\frac{2}{\xi},$$

from which we obtain $\Phi_1 \sim -\xi \ln \xi + C\xi$ for large ξ. The value of C can be determined only by solving the full problem for Φ_1 (with boundary conditions at $\xi = 0$) as performed above (i.e., $C = 1$).

2.3. The outer solution. To analyze the outer solution we convert to self-similar coordinates

$$\tau = \ln(T-t)^{-1}, \qquad y = e^{\tau/2} r,$$

where T is the time of blowup. When we set $\zeta(\tau, y) = \theta(t, r)$, ζ satisfies the equation

$$\zeta_\tau = \zeta_{yy} + \left(\frac{1}{y} - \frac{y}{2}\right)\zeta_y - \frac{\sin 2\zeta}{2y^2}.$$

Linearizing around $\zeta = \pi$, one obtains the linear equation

$$\eta_\tau = \mathcal{L}_0 \eta \stackrel{\text{def}}{=} \eta_{yy} + \left(\frac{1}{y} - \frac{y}{2}\right)\eta_y - \frac{1}{y^2}\eta. \tag{2.4}$$

We require that $\eta(\tau, y)$ grow less than exponentially for large y, since otherwise it is impossible to match the outer to the remote region. As a boundary condition for small y we take $\eta(\tau, 0) = 0$, at least to leading order. The reason for this choice is not transparent at the moment since it is actually part of the matching process. We will clarify this point in the next section. For now we just stress that this boundary condition is not forced by regularity requirements but turns out to be required to get consistent matching.

We look for separable solutions of (2.4) obeying these two "boundary" conditions (one boundary condition and one growth condition, really). The solution of this type that decays most slowly with τ is $\eta = e^{-\tau/2} y$. From a different perspective this means

that, under the above boundary conditions, $\lambda = -\frac{1}{2}$ is the smallest eigenvalue of $\mathcal{L}_0$. (By standard arguments the eigenfunctions are polynomials.) Hence one expects η to approach π essentially at rate $e^{-\tau/2}$, and thus we set

$$\eta \sim \sigma(\tau) e^{-\tau/2}\, y.$$

Since we still have to match to the inner solution (i.e., the true boundary condition is not $\eta(\tau, 0) = 0$), we need to let the coefficient σ depend on τ, but in such a way that σ is not exponential in τ. We thus use "almost" separable solutions or, in the original variables, "almost" self-similar solutions. For the solution of (2.4) we now put forward the ansatz

$$\eta \sim e^{-\tau/2} \sum_{i=0}^{\infty} \frac{d^i \sigma(\tau)}{d\tau^i} \Omega_i(y), \tag{2.5}$$

where we anticipate σ to be algebraically decaying. When (with the linear operator $\mathcal{L}_0$ defined in (2.4))

$$\left(\mathcal{L}_0 + \tfrac{1}{2}\right) \Omega_0 = 0 \qquad \text{and} \qquad \left(\mathcal{L}_0 + \tfrac{1}{2}\right) \Omega_i = \Omega_{i-1}, \qquad i = 1, 2, \ldots,$$

then (2.5) is formally a solution of (2.4) for arbitrary $\sigma(\tau)$. In some sense the sequence Ω_i forms a Jordan sequence of the linear operator $\mathcal{L}_0$ at eigenvalue $-\frac{1}{2}$, except that the left boundary condition need no longer be satisfied. (In fact, since $-\frac{1}{2}$ is a simple eigenvalue, the left boundary condition cannot be satisfied.) This shortage of boundary conditions causes nonuniqueness for Ω_i, but this is not an issue. The left boundary condition was not a true boundary condition anyway, but merely induced by matching requirements. One obtains

$$\begin{aligned} \Omega_0 &= C_0 y, \\ \Omega_1 &= C_0(4y^{-1} - 2y \ln y) + C_1 y, \end{aligned}$$

where the values of the constants are of no significance because they can be absorbed in σ, and without loss of information we set $C_0 = 1$ and $C_1 = 0$. Thus for large τ the outer approximation is

$$\zeta(\tau, y) \sim \pi + e^{-\tau/2}\left[\sigma(\tau) y + \sigma'(\tau)(4y^{-1} - 2y \ln y)\right] \qquad \text{for large } \tau. \tag{2.6}$$

The function $\sigma(\tau)$ will have to be determined by matching to the inner solution. The outer approximation is valid, provided that $y^2 (\ln y)^2 \ll |\frac{\sigma'}{\sigma''}|$ and $y^2 |\ln y| \gg |\frac{\sigma''}{\sigma'}|$ on the side of large and small y, respectively, because the next terms are of order $O(\sigma'' y (\ln y)^2)$ for large y and $O(\sigma'' y^{-1})$ for small y. Here we should keep in mind that σ will turn out to be algebraically decaying (in which case $|\frac{\sigma''}{\sigma'}| = O(\tau^{-1})$).

2.4. Matching. To match the outer to the inner solution we rewrite (2.3) in terms of the self-similar variables:

$$v(t, \xi) = \tilde{v}(\tau, y) \sim \pi - 2e^{\tau/2} \tilde{R}\, y^{-1} + e^{\tau/2} \tilde{R}'\, y \left(-\ln y + \ln \tilde{R} + \frac{\tau}{2} + 1\right),$$

where $\tilde{R}(\tau) = R(t)$ and hence $R'(t) = e^{\tau} \tilde{R}'(\tau)$. Comparing this to (2.6), we obtain

$$O(y^{-1}): \qquad 4e^{-\tau/2}\sigma' \sim -2\tilde{R} e^{\tau/2}, \tag{2.7}$$

$$O(y \ln y): \qquad -2e^{-\tau/2}\sigma' \sim -\tilde{R}' e^{\tau/2}, \tag{2.8}$$

$$O(y): \qquad e^{-\tau/2}\sigma \sim \tilde{R}' e^{\tau/2} \left(\ln \tilde{R} + \frac{\tau}{2} + 1\right). \tag{2.9}$$

Equations (2.7) and (2.8) suggest that $\tilde{R}(\tau) = e^{-\tau}\rho(\tau)$, where $\rho(\tau)$ is algebraic for large τ. Substituting this into (2.7) and (2.9), we obtain the following relations for ρ and σ,

$$\rho \sim -2\sigma' \quad \text{and} \quad \frac{\tau}{2}\rho \sim \sigma \qquad \text{as } \tau \to \infty,$$

from which we conclude that

$$\sigma(\tau) \sim \frac{\kappa}{2\tau} \quad \text{and} \quad \rho(\tau) \sim \frac{\kappa}{\tau^2} \qquad \text{as } \tau \to \infty$$

for some $\kappa > 0$. We thus find

$$R(t) \sim \kappa \frac{T-t}{|\ln(T-t)|^2} \qquad \text{as } t \uparrow T. \tag{2.10}$$

We emphasize that this asymptotic behavior of $R(t)$ is clearly completely different from the self-similar rate $(T-t)^{1/2}$. It is now possible to check that the regions of validity for the inner and outer approximations do indeed overlap: the region of overlap is $(\tau \ln \tau)^{-1/2} \ll y \ll 1$. We note that the matching conditions (2.7) and (2.8) convey the same information, so that (2.8) is in some sense redundant, though it provides additional confidence in the matching.

The smallest intersection of $\theta(t,r)$ with π, denoted by $r = S(t)$, can be calculated from (2.6) using the asymptotic form of σ, which leads to

$$S(t) \sim 2\sqrt{\frac{T-t}{|\ln(T-t)|}} \qquad \text{as } t \uparrow T.$$

Notice that there is no undetermined constant in this leading order formula.

The asymptotic behavior for small r of the limit profile at $t = T$ is computed by matching the outer approximation to the remote solution

$$\theta(t,r) \sim \Theta(r) + (t-T)\left[\Theta_{rr} + \frac{1}{r}\Theta_r - \frac{\sin 2\Theta}{2r^2}\right] \qquad \text{as } t \to T,$$

with $\Theta(r)$ being the limit profile $\theta(T,r)$. This approximation in the remote region is valid for $r \gg (T-t)^{1/2}$ (at the least). The regions of validity of the outer and remote approximation overlap. Matching the remote solution to (2.6), we find

$$\theta(T,r) \sim \pi + \frac{1}{4}\kappa \frac{r}{|\ln r|} \qquad \text{for small } r, \tag{2.11}$$

with the same constant $\kappa > 0$ as in (2.10). Notice that $\theta(T,r) > \pi$ for small r. A quick way to obtain this behavior from (2.6) is by letting τ tend to infinity for fixed y.

We remark that immediately after blowup a different type of inner layer near $r = 0$ appears, which describes how analyticity is recovered. In this layer almost self-similar behavior occurs of the form

$$\theta(t,r) \sim \pi + \frac{\sqrt{t-T}}{|\ln(t-T)|} f_1\left(\frac{r}{\sqrt{t-T}}\right) + \frac{\sqrt{t-T}}{|\ln(t-T)|^2} f_2\left(\frac{r}{\sqrt{t-T}}\right)$$

as $t \downarrow T$ for small r. Substituting this into the equation for θ, one finds that $f_1(z) = Dz$ for some $D \in \mathbb{R}$ to be determined, and $f_2(z)$ is a solution of

$$f_2'' + \left(\frac{1}{z} + \frac{z}{2}\right) f_2' - \left(\frac{1}{z^2} + \frac{1}{2}\right) f_2 = f_1,$$

with the property that $f_2(z) = O(z)$ as $z \to 0$. One may solve for f_2 by reduction of order, but it is sufficient to remark that it follows from the limiting equation for large z that $f_2(z) \sim 2Dz \ln z$ as $z \to \infty$. Matching this with (2.11) implies that $D = \frac{1}{8}\kappa$, so that the result is

$$\theta(t,r) \sim \pi + \frac{1}{8}\kappa \frac{r}{|\ln(t-T)|} + \frac{\sqrt{t-T}}{|\ln(t-T)|^2} f_2\left(\frac{r}{\sqrt{t-T}}\right) \qquad \text{as } t \downarrow T \text{ for small } r.$$

Let us now come back to the point of choosing the boundary condition for the outer solution. In section 2.3 it was put forward that the left boundary condition for the outer solution is $\eta(\tau, 0) = 0$ but that this is in fact already part of the matching. Here we explain the reasoning behind this. There is a family of separable solutions to (2.4):

$$\eta = e^{\lambda\tau} f_\lambda(y),$$

where $f = f_\lambda$ obeys

$$\lambda f = \mathcal{L}_0 f \stackrel{\text{def}}{=} f_{yy} + \left(\frac{1}{y} - \frac{y}{2}\right) f_y - \frac{1}{y^2} f. \tag{2.12}$$

Since we are looking for solutions that do not blow up as $\tau \to \infty$, we restrict ourselves to $\lambda \leq 0$. (On the other hand, this restriction will also be a consequence of what follows.)

Now suppose that λ is not in the set $\{-\frac{1}{2}, -\frac{3}{2}, -\frac{5}{2}, \dots\}$. We will show that this leads to inconsistent matching conditions. We find two linearly independent solutions $g(y)$ and $h(y)$ of (2.12) with the following properties. The asymptotic behavior of $g(y)$ is $g(y) \sim y$ as $y \downarrow 0$, and it grows faster than exponentially as $y \to \infty$; it can therefore be ruled out. The other, linearly independent, solution $h(y)$ grows less than exponentially as $y \to \infty$, and for small y it behaves as

$$h(y) \sim y^{-1} + \frac{2\lambda - 1}{4} y \ln y + C_\lambda y$$

for some constant $C_\lambda \in \mathbb{R}$, the value of which is not relevant except for $\lambda = 0$: $C_0 = \frac{1}{8}(4\ln 2 + 1 - \gamma)$, where γ is Euler's constant. We note that $h(y)$ is closely related to a Kummer-U function.

Still assuming that $\lambda \notin \{-\frac{1}{2}, -\frac{3}{2}, -\frac{5}{2}, \dots\}$, we are lead to matching conditions of the form (assuming throughout that $\sigma' \ll \sigma$)

$$\begin{aligned} O(y^{-1}): &\qquad e^{\lambda\tau}\sigma \sim -2\tilde{R}e^{\tau/2}, \\ O(y\ln y): &\quad \frac{2\lambda-1}{4} e^{\lambda\tau}\sigma \sim -\tilde{R}' e^{\tau/2}, \\ O(y): &\qquad C_\lambda e^{\lambda\tau}\sigma \sim \tilde{R}' e^{\tau/2}\left(\ln\tilde{R} + \frac{\tau}{2} + 1\right). \end{aligned}$$

Now $\tilde{R} = e^{(\lambda - 1/2)\tau}\rho(\tau)$, where $\rho(\tau)$ is not exponential in τ. Since we require that $R \ll (T-t)^{1/2} = e^{-\tau/2}$ for self-consistency, we find that $\lambda \leq 0$. For $\rho(\tau)$ one obtains the relations (if $\lambda < 0$)

$$\sigma(\tau) \sim -2\rho(\tau) \qquad \text{and} \qquad C_\lambda \sigma(\tau) \sim \lambda\left(\lambda - \frac{1}{2}\right)\tau\rho(\tau),$$

which immediately leads to a contradiction. When $\lambda = 0$ (which would correspond to an almost self-similar blowup rate) we obtain

$$\sigma(\tau) \sim -2\rho(\tau) \qquad \text{and} \qquad C_0\sigma(\tau) \sim -\frac{1}{2}\rho(\tau),$$

and since $C_0 \neq \frac{1}{4}$, this leads to a contradiction as well. A more intuitive explanation for the matching failure is that it is impossible to match a leading order term y^{-1} from the outer expansion with a correction term from the inner one.

If $\lambda \in \{-\frac{1}{2}, -\frac{3}{2}, -\frac{5}{2}, \dots\}$, then the solution obeying the growth condition on the right is regular near $y = 0$; i.e., there are no terms of order y^{-1} (or $y \ln y$). The set $\{-\frac{1}{2}, -\frac{3}{2}, -\frac{5}{2}, \dots\}$ thus consists of the eigenvalues of problem (2.12) with boundary condition $f(0) = 0$ and the growth condition for $y \to \infty$. This explains the choice of boundary conditions in section 2.3. The case $\lambda = -\frac{1}{2}$ was dealt with in the previous sections. In the next section we consider the remaining possibilities.

2.5. Degenerate (nongeneric) cases. There is a whole family of separable solutions of (2.4) which obey the growth condition (less than exponential) on the right and the boundary condition on the left (regular near $y = 0$). The solution $\eta = e^{-\tau/2}y$ is the first one, i.e., the least rapidly decaying one. In degenerate cases it may, however, happen that the coefficient σ in front of this term in (2.6) vanishes. In that case a degenerate situation occurs with codimension 1. The outer solution in that case becomes

$$\zeta \sim \pi + e^{-3\tau/2}\sigma(\tau)\left(y - \frac{1}{8}y^3\right),$$

since $\lambda = -\frac{3}{2}$ is the second smallest eigenvalue of $\mathcal{L}_0$, with eigenfunction $y - \frac{1}{8}y^3$. Following the calculation in section 2.3, the outer approximation now becomes

(2.13)
$$\zeta \sim \pi + e^{-3\tau/2}\left[\sigma(\tau)\left(y - \frac{1}{8}y^3\right) + \sigma'(\tau)\left(2y^{-1} - 2y\ln y + \frac{1}{4}y^3\ln y - \frac{3}{2}y\right)\right].$$

Notice that the profile is nonmonotone for times close to blowup. The matching conditions become

$$\begin{aligned} O(y^{-1}): &\quad 2e^{-3\tau/2}\sigma' \sim -2\tilde{R}e^{\tau/2}, \\ O(y\ln y): &\quad -2e^{-3\tau/2}\sigma' \sim -\tilde{R}'e^{\tau/2}, \\ O(y): &\quad e^{-3\tau/2}\sigma \sim \tilde{R}'e^{\tau/2}\left(\ln\tilde{R} + \frac{\tau}{2} + 1\right), \end{aligned}$$

so that $\tilde{R}(\tau) = e^{-2\tau}\rho(\tau)$ and $\rho(\tau) \sim C\tau^{-4/3}$; hence

$$R(t) \sim \kappa\frac{(T-t)^2}{|\ln(T-t)|^{4/3}} \qquad \text{as } t \uparrow T$$

for some $\kappa > 0$.

To calculate the limit profile we have to match the outer solution into the remote solution $\theta(t, r) \sim \Theta(r) = \theta(T, r)$. To match (2.13) to the remote region, one needs to take into account the highest order terms in τ and y only. For the limit profile one finds

$$\theta(T, r) \sim \pi - \frac{3}{8}\kappa\frac{r^3}{|2\ln r|^{1/3}} \qquad \text{for small } r,$$

with the same constant $\kappa > 0$ as above. Notice that $\theta(T, r) < \pi$ for small r (in contrast to the nondegenerate case). The first two intersections of $\theta(t, r)$ with π, denoted by $S_1(t)$ and $S_2(t)$, behave asymptotically as

$$S_1 \sim \sqrt{\frac{2}{3}\frac{T-t}{|\ln(T-t)|}} \qquad \text{and} \qquad S_2 \sim \sqrt{8(T-t)}\,.$$

The first intersection comes from the balance between Ω_0 and Ω_1 (i.e., occurs for small y), while the second intersection depends only on Ω_0 (having $y = O(1)$).

In a similar vein, for degenerate cases occurring with codimension k (with $k = 0, 1, 2, \ldots$; the generic case is embedded in this), the $(k+1)$th eigenvalue of $\mathcal{L}_0$ is $\lambda = -k - \frac{1}{2}$, with as eigenfunction a $(2k+1)$th order polynomial with only odd terms, say $\Omega_0^k(y)$, which we normalize so that $\Omega_0^k(y) \sim y$ as $y \to 0$. We note that $\Omega_0^k(y)$ can be expressed in terms of a generalized Laguerre polynomial. Then

$$\Omega_0^k(y) = \sum_{i=0}^{k} a_i y^{2i+1} \quad \text{with} \quad a_i = (-1)^i \frac{k!}{2^{2i}(k-i)!(i+1)!i!} \quad \text{for } i = 0, 1, \ldots, k.$$

To calculate the next term in the expansion Ω_1^k we have to solve

$$\left(\mathcal{L}_0 + k + \frac{1}{2}\right)\Omega_1^k = \Omega_0^k.$$

After a bit of calculation one finds that

$$\Omega_1^k(y) = \frac{4}{k+1} y^{-1} - 2\Omega_0^k(y)\ln y + h_k(y),$$

where h_k is some odd polynomial of degree $2k-1$. The outer approximation becomes

$$\zeta \sim \pi + e^{-(k+1/2)\tau}[\sigma(\tau)\Omega_0^k(y) + \sigma'(\tau)\Omega_1^k(y)],$$

and thus for small y

$$\zeta \sim \pi + e^{-(k+1/2)\tau}\left[\sigma(\tau)\, y + \sigma'(\tau)\,\left(\frac{4}{k+1} y^{-1} - 2y\ln y\right)\right].$$

The matching condition for the codimension k degeneracy become

$$\begin{aligned} O(y^{-1}): &\quad \frac{4}{k+1} e^{-(k+1/2)\tau}\sigma' \sim -2\tilde{R}e^{\tau/2}, \\ O(y\ln y): &\quad -2e^{-(k+1/2)\tau}\sigma' \sim -\tilde{R}'e^{\tau/2}, \\ O(y): &\quad e^{-(k+1/2)\tau}\sigma \sim \tilde{R}'e^{\tau/2}\left(\ln\tilde{R} + \frac{\tau}{2} + 1\right). \end{aligned}$$

Hence $\tilde{R} \sim \kappa e^{-(k+1)\tau}\tau^{-(2k+2)/(2k+1)}$, or in the original variables,

$$R(t) \sim \kappa \frac{(T-t)^{k+1}}{|\ln(T-t)|^{(2k+2)/(2k+1)}} \qquad \text{as } t \uparrow T$$

for some $\kappa > 0$. Since $\Omega_0^k \sim a_k y^{2k+1}$ as $y \to \infty$, the limit profile is

$$\theta(T, r) \sim \pi + (-1)^k \kappa \frac{2k+1}{2^{2k+1}k!} \frac{r^{2k+1}}{|2\ln r|^{1/(2k+1)}} \qquad \text{as } t \uparrow T.$$

This limit profile is again obtained by matching to the outer solution or, alternatively, by taking the limit $\tau \to \infty$ for fixed y and subsequently $y \to \infty$. One can analyze the short time behavior just after blowup in a similar way to that in section 2.4. Finally, the asymptotic behavior of the first intersection of $\theta(t,r)$ with π, denoted by $S_1(t)$, is

$$S_1(t) \sim 2\sqrt{\frac{1}{(k+1)(2k+1)}\frac{T-t}{|\ln(T-t)|}} \qquad \text{as } t \uparrow T.$$

We have thus found a countable family of nongeneric blowup scenarios. The codimension 1 situation, for example, occurs at the borderline between the generic blowup scenario and the case of no blowup (as explained in section 1); it is characterized by the fact that *two* intersections of $\theta(t,r)$ with π approach the origin simultaneously as $t \uparrow T$ (though at different rates). More generally, in the codimension k scenario the profile $\theta(t,r)$ just before blowup has $k+1$ intersections with π that approach the origin as $t \uparrow T$; in other words, it corresponds to a nongeneric scenario in which the disappearance of sign changes in $\theta - \pi$ coincides with the blowup time (cf. [13] for a detailed discussion of sign change solutions in a different second order parabolic problem). The appearance of degenerate cases indicates that it might be hard to obtain a proof of the formal result that near blowup $R(t) \sim \kappa \frac{T-t}{|\ln(T-t)|^2}$ holds *generically*. Restricting our analysis to certain classes of monotone initial data excludes the degenerate possibilities, because one can show that the solution then has to remain monotone for all $t > 0$ (see [1]), and all the degenerate blowup scenarios have nonmonotone profiles just before blowup. Our analysis strongly suggests, however, that even without such restrictions on the initial data, the generic blowup rate is $R(t) \sim \kappa \frac{T-t}{|\ln(T-t)|^2}$.

2.6. Boundary condition $\theta_1 = \pi$. When the boundary condition is $\theta(t,1) = \pi$, there is, besides the finite time blowup scenarios described above, an additional possibility, namely that blowup occurs in infinite time. This infinite time blowup is a codimension 0 scenario (we analyze degenerate cases as well). In this case there is no urge to change to self-similar coordinates. Close to blowup the profile is assumed to be near π in the whole of the remote region (and the limit profile is identically equal to π). The linearized equation around π is

$$w_t = \mathcal{L}_1 w \stackrel{\text{def}}{=} w_{rr} + \frac{1}{r} w_r - \frac{1}{r^2} w. \tag{2.14}$$

We substitute a formal series

$$w \sim \pi + \sum_{i=0}^{\infty} \frac{d^i s(t)}{dt^i} W_i(r), \tag{2.15}$$

where $s(t)$ is an arbitrary function and

$$\mathcal{L}_1 W_0 = 0 \qquad \text{and} \qquad \mathcal{L}_1 W_i = W_{i-1}, \qquad i = 1, 2, \ldots.$$

Now (2.15) is again a formal solution of the linearized differential equation for any $s(t)$, and when we require that $W_i(1) = 0$, then the right boundary condition is satisfied. There is no a priori left boundary condition. We obtain

$$\begin{aligned} W_0 &= C_0(r^{-1} - r), \\ W_1 &= C_0\left(\frac{1}{2} r \ln r + \frac{1}{8} r - \frac{1}{8} r^3\right) + C_1(r^{-1} - r), \end{aligned}$$

where we may again set $C_0 = 1$ and $C_1 = 0$ without loss of generality. Hence for the remote region we obtain for small r

$$\theta \sim \pi + s(t)(r^{-1} - r) + s'(t)\left(\frac{1}{2} r \ln r + \frac{1}{8} r\right).$$

Matching in a way similar to that of section 2.4 to the inner solution (2.3), we find that

$$\begin{aligned} O(r^{-1}): &\qquad s \sim -2R, \\ O(r \ln r): &\qquad \frac{1}{2} s' \sim -R', \\ O(r): &\quad -s + \frac{1}{8} s' \sim R'(\ln R + 1). \end{aligned}$$

Hence for large t

$$R(t) \sim e^{-2\sqrt{t} - 5/4} \qquad \text{as } t \to \infty.$$

We note that this is the one instance where the term $R'R\xi = R'r$ in the inner approximation has an influence on the leading order result. (It is needed to calculate the multiplicative constant $e^{-5/4}$.) The most striking difference between this result and the situation in section 2.4 is that blowup now occurs in infinite time. We remark that, for suitable initial data, blowup may happen in finite time via the scenario in the previous sections, after which $\theta \to \pi$ uniformly as $t \to \infty$ (generically at rate $e^{-\lambda_1^2 t}$, where λ_1 is the first zero of the Bessel function J_1; see also below). However, this cannot happen for initial profiles $|\theta(0, r)| \le \pi$ for all $r \in [0, 1]$, since then $|\theta(t, r)| \le \pi$ for all $t > 0$ (by the comparison principle) and blowup is postponed until $t = \infty$.

As in the previous section, there is a hierarchy of degenerate cases. Looking for almost separable solutions, one tries, with $\lambda > 0$,

$$w \sim \pi + e^{-\lambda^2 t} \sum_{i=0}^{\infty} \frac{d^i s(t)}{dt^i} W_i(r).$$

Now

$$W_0 = J_1(\lambda r) - \frac{J_1(\lambda)}{Y_1(\lambda)} Y_1(\lambda r),$$

where J_i and Y_i are the Bessel functions of order i (take $W_0 = Y_1(\lambda r)$ if $Y_1(\lambda) = 0$). Analogous to section 2.4, this does not lead to self-consistent matching unless W_0 is regular at $r = 0$. Therefore we require that $J_1(\lambda) = 0$, and we obtain a nice eigenvalue problem with Dirichlet boundary conditions.

Let $\lambda_k > 0$ be the kth zero of J_1 and $W_0(r) = J_1(\lambda_k r)$. Then one finds that

$$\begin{aligned} W_1 = \frac{\pi}{8} \big[& J_1(\lambda_k r) J_2(\lambda_k r) Y_0(\lambda_k r) r^2 + J_1(\lambda_k r) J_0(\lambda_k r) Y_2(\lambda_k r) r^2 \\ & - 2 J_0(\lambda_k r) J_2(\lambda_k r) Y_1(\lambda_k r) r^2 + D_k Y_1(\lambda_k r) \big], \end{aligned}$$

where

$$D_k \stackrel{\text{def}}{=} 2 J_0(\lambda_k) J_2(\lambda_k) < 0.$$

The matching conditions become

$$\begin{aligned} O(r^{-1}): &\quad -e^{-\lambda_k^2 t}\frac{D_k}{4\lambda_k}s' \sim -2R, \\ O(r\ln r): &\quad e^{-\lambda_k^2 t}\frac{D_k\lambda_k}{8}s' \sim -R', \\ O(r): &\quad e^{-\lambda_k^2 t}\frac{1}{2}\lambda_k s \sim R'(\ln R + 1). \end{aligned}$$

We infer that, with codimension $k > 0$,

$$R(t) \sim \kappa\, t^{-1+4/D_k\lambda_k^2}\, e^{-\lambda_k^2 t} \qquad \text{as } t \to \infty$$

for some $\kappa > 0$.

We stress that the generic case (discussed at the beginning of this subsection) does not obey the boundary condition $W_0(0) = 0$ but nevertheless leads to consistent matching. This is, however, the only consistent case that has a spatial singularity near the origin in the remote region. (The solution is, of course, not singular in the inner variable.) This may be somewhat surprising. Let us clarify the role of the generic and nongeneric scenarios.

Notice that for any $\lambda > 0$ the associated eigenfunction $W_0(r)$ has sign changes on $(0, 1)$, leading to intersections of the solution with π close to blowup. Consider initial profiles that lie entirely below π, i.e., $-\pi < \theta(0, r) < \pi$ for $r \in (0, 1)$. A comparison argument shows that for such initial data these blowup profiles are excluded. This indicates that a generic scenario is associated with $\lambda = 0$ (even though the "eigenfunction" (which obeys the boundary condition at $r = 1$) for $\lambda = 0$ is singular).

The degenerate cases act as borderline cases between finite time blowup and infinite time blowup. For example, when the initial data have a sufficiently large bump above π, the solution will blowup in finite time, whereas solutions starting from initial data below π blow up in infinite time. The codimension 1 scenario found in this subsection acts as the borderline between generic infinite and generic finite time blowup. This is most easily understood for initial conditions which have only one crossing with π, since for such initial data the finite time codimension 1 scenario plays no role (because it has two crossings with π close to blowup). As we have seen in this section, the infinite time blowup phenomenon for $\theta_1 = \pi$ is essentially driven by the linear equation (2.14), and for such an equation the presence of high codimension cases with many sign changes (of $\theta - \pi$) is not unexpected.

2.7. Reverse jumps. As explained in section 1, weak solutions have the possibility to make a jump in which they *increase* their energy $\mathcal{E}(t)$; see [5, 2]. We consider the situation where $\theta(t, 0)$ jumps from π to 0 ("jumping back") at $t = t_0$. In these jumps the energy $\mathcal{E}$ necessarily increases by 4π at $t = t_0$.

The inner approximation is the same as in section 2.2, although now $R(t) \to 0$ as $t \downarrow t_0$. Concerning the outer approximation, we argue as follows. As in section 2.3, we turn to self-similar variables ($t > t_0$),

$$\tau = \ln(t - t_0), \qquad y = e^{-\tau/2}r,$$

and obtain for $\zeta(\tau, y) = \theta(t, r)$

$$\zeta_\tau = \zeta_{yy} + \left(\frac{1}{y} + \frac{y}{2}\right)\zeta_y - \frac{\sin 2\zeta}{2y^2}.$$

Since a reverse jump can occur at any moment, one generically has $\theta_r(t_0, 0) = C \neq 0$, i.e., the dominant term in the local expansion is $Cr = Ce^{\tau/2}y$. This being a separable solution of the linearized equation, one proposes a solution of the form $e^{\tau/2}\eta(y)$ (cf. section 2.3), where η satisfies

$$\eta_{yy} + \left(\frac{1}{y} + \frac{y}{2}\right)\eta_y - \left(\frac{1}{y^2} + \frac{1}{2}\right)\eta = 0.$$

The general solution is

$$\eta = C_0 y + C_1 y \int_y^\infty \frac{e^{-s^2/4}}{s^3} ds,$$

with arbitrary constants C_0, $C_1 \in \mathbb{R}$ and $y_0 > 0$. We see that, as opposed to the situation in section 2.3, the second independent solution has reasonable growth. In fact, it tends to 0 as $y \to \infty$ and behaves as $y^{-1} + \frac{1}{2}y \ln y$ for small y. Therefore the outer approximation becomes

$$\zeta \sim \pi + e^{\tau/2}\left[\alpha(\tau)y + \beta(\tau)(4y^{-1} + 2y \ln y)\right] \qquad \text{for small } y \text{ and } -\tau \gg 1.$$

As matching conditions for $\tau \to -\infty$ we find (writing $\tilde{R}(\tau) = R(t)$)

$$\begin{aligned} O(y^{-1}): &\quad 4\beta e^{\tau/2} \sim -2\tilde{R}e^{-\tau/2}, \\ O(y \ln y): &\quad 2\beta e^{\tau/2} \sim -\tilde{R}' e^{-\tau/2}, \\ O(y): &\quad \alpha e^{\tau/2} \sim \tilde{R}' e^{-\tau/2}\left(\ln \tilde{R} - \frac{\tau}{2} + 1\right). \end{aligned}$$

We infer that $\tilde{R} \sim e^\tau \rho(\tau)$ with $\rho(\tau) \sim -\frac{2\ell}{\tau}$ for some $\ell = -\alpha(-\infty) > 0$, and $\beta \sim \frac{\ell}{\tau}$, so that $|\beta(\tau)| \ll |\alpha(\tau)|$ as $\tau \to -\infty$. Hence

$$R(t) \sim 2\ell \frac{t - t_0}{|\ln(t - t_0)|} \qquad \text{as } t \downarrow t_0$$

for some $\ell > 0$. The limit profile is

$$\theta(t_0, r) \sim \pi - \ell r \quad \text{for small } r,$$

which is consistent with the assumption at the start, so that $\ell = -C$. Notice that, when the jump is downwards from π to 0, then the profile at $t = t_0$ has to be decreasing for small r.

A whole hierarchy of degenerate cases in which $\theta_r(0, 0) = 0$ can be calculated as well. At some $t_0 > 0$ (i.e., after the solution has started to evolve) it happens with codimension k that $\theta(t_0, r) \sim \pi - \tilde{C}r^{2k+1}$ as $r \downarrow 0$ for some $\tilde{C} \neq 0$. Reverse jumps at such nongeneric instances can be analyzed via the method presented above. The jump (of $\theta(t, 0)$) is downwards when $\tilde{C} > 0$ and upwards when $\tilde{C} < 0$. For example, for the codimension 1 scenario one finds

$$R \sim \frac{8}{3}\ell \frac{(t - t_0)^2}{|\ln(t - t_0)|} \qquad \text{as } t \downarrow t_0 \tag{2.16}$$

for some $\ell > 0$, with $\theta(t_0, r) \sim \pi - \ell r^3$ as $r \to 0$. Reverse jumps can also happen at the moment of a forward jump (i.e., $t_0 = T$); see also [18, section 5]. Then at

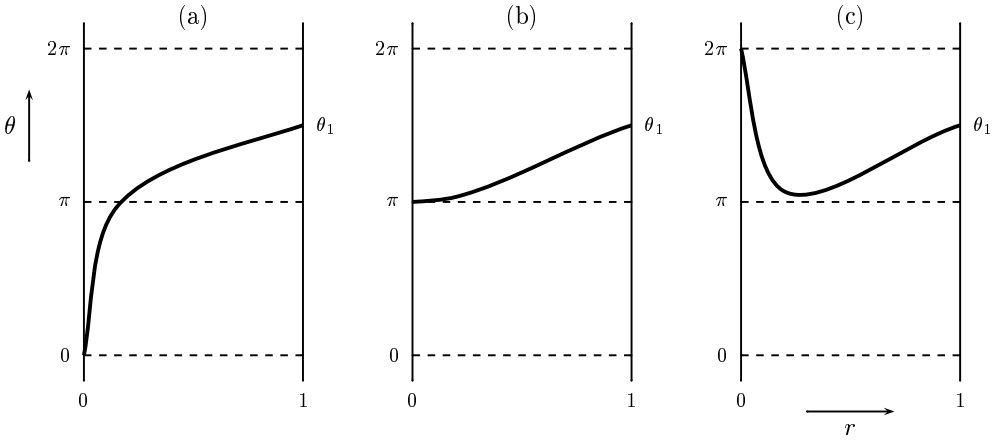


FIG. 2.2. *Picture of the situation when $t_0 = T$; i.e., a reverse jump happens at the moment of a forward jump:* (a) *a time just before $t = T$,* (b) $t = T = t_0$, (c) *a time just after $t = t_0$.*

the self-similar scale a logarithmic correction needs to be applied (cf. the recovery of analyticity in section 2.4). We note that if $t_0 = T$ and if the forward jump from 0 to π behaves according to the generic scenario, then the reverse jump must be from π to 2π. (This follows from a comparison of the limiting profiles as $t \uparrow T$ and $t \downarrow t_0$.) A schematic picture of the situation is given in Figure 2.2. We leave the details to the reader and just state the result:

$$R(t) \sim \kappa \frac{t - t_0}{|\ln(t - t_0)|^2} \qquad \text{as } t \downarrow t_0, \tag{2.17}$$

where $\kappa > 0$ is the same constant as in the forward jump (see (2.10)). For $t_0 = 0$ one may choose, for example, initial data with $\theta(0, r) \sim \pi - \hat{C} r^a$ as $r \downarrow 0$ for any $a > 0$ and some $\hat{C} \neq 0$, and a reverse jump can then happen instantaneously. The analysis is again along the same lines.

Finally, our analysis suggests that, given a jump time t_0, the asymptotic profile of $\theta(t_0, r)$ as $r \downarrow 0$ completely determines the blowup rate $R(t)$. In [2] it is conjectured that from a physical perspective it is most likely that if the solution has first made a forward jump from 0 to π, the reverse jump happens at the first instance at which $\theta(t_0, r) < 0$ close to the origin. This corresponds to the system selecting an otherwise degenerate scenario, whereby $\theta(t_0, r) \sim \pi - \ell r^3$ as $r \to 0$ for some $\ell > 0$, and the blowup rate is then given by (2.16). More degenerate scenarios can of course occur with higher codimension. We note that our scenario is subtly different from the one studied in [2], which requires linear behavior of θ at the origin.

3. The general case.

3.1. Preamble.

We now analyze the generalization

$$\theta_t = \theta_{rr} + \frac{1}{r}\theta_r - n^2 \frac{\sin 2\theta}{2r^2}, \qquad 0 \leq r \leq 1. \tag{3.1}$$

Apart from the fact that the equation yields physically relevant solutions for $n = 1, 2, 3, \ldots$ as explained in section 1, analyzing the dependence of the blowup behavior on n also enhances the understanding of the special case $n = 1$.

We deal with the inner approximation in section 3.2 and then have a first attempt at matching without using self-similar coordinates to get the general idea in section 3.3. In section 3.4 we deal with $n < 2$, while section 3.5 deals with $n \geq 2$. In section 3.6 we discuss simultaneous blowup at several scales. The special boundary condition $\theta(t,1) = \pi$ is dealt with in section 3.7. We remark on the case of an infinite domain in section 3.8, and section 3.9 is devoted to reverse jumps.

3.2. The inner approximation. The stationary solutions of (3.1) are

$$\theta(r) = m\pi + 2\arctan(qr^n), \qquad m \in \mathbb{Z},\ q \in \mathbb{R}.$$

Choosing the scaling $\xi = r/R(t)$ so that

$$R(t)^n\theta(t,r) \sim 2r^n \quad \text{as } r \downarrow 0 \qquad \text{for all } t \text{ up to the blowup time,}$$

the outer limit of the inner approximation $v(t,\xi) = \theta(t,r)$ (cf. section 2.2) becomes for $n > \frac{1}{2}$, $n \neq 1$,

$$v \sim \pi - 2\,\xi^{-n} + R'R\left(\frac{n}{2n-2}\xi^{-n+2} - E_n\xi^n\right). \tag{3.2}$$

The coefficient of the term ξ^{-n+2} can be obtained from the asymptotic equation for Φ_1 at large ξ, as explained at the end of section 2.2. The coefficient E_n of the term ξ^n can be obtained only by solving the full problem for Φ_1 (borrowing the notation from section 2.2):

$$\Phi_{1\xi\xi} + \frac{1}{\xi}\Phi_{1\xi} - n^2\frac{\cos(4\arctan\xi^n)}{\xi^2}\Phi_1 = -\frac{2n\xi^n}{1+\xi^{2n}},$$

with boundary condition $\lim_{\xi\downarrow 0}\frac{\Phi_1(\xi)}{\xi^n} = 0$. We rewrite

$$\cos(4\arctan\xi^n) = \frac{-6\xi^{2n}+1+\xi^{4n}}{(1+\xi^{2n})^2}$$

and note that

$$\frac{d(2\arctan q\xi^n)}{dq}\Big|_{q=1} = \frac{2n\xi^n}{1+\xi^{2n}}$$

is a solution of the homogeneous equation. Using variation of constants, one finds that the solution we are looking for is

$$\Phi_1 = \frac{(1-\xi^{4n}-4n\xi^{2n}\ln\xi)\int_0^\xi \frac{s^{2n+1}}{(1+s^{2n})^2}ds + \xi^{2n}\int_0^\xi \frac{s(s^{4n}-1+4ns^{2n}\ln s)}{(1+s^{2n})^2}ds}{\xi^n(1+\xi^{2n})}.$$

This gives for $n > 1$

$$E_n = \int_0^\infty \frac{s^{2n+1}}{(1+s^{2n})^2}ds = \frac{\pi}{2n^2\sin(\frac{\pi}{n})}.$$

For $n < 1$ the value of E_n is somewhat irrelevant, since the term ξ^n in (3.2) is non-dominant in that case. Nevertheless, for $\frac{1}{2} < n < 1$,

$$E_n = \int_0^1 \frac{s^{2n+1}-2s^{4n-3}-s^{6n-3}}{(1+s^{2n})^2}ds - 1.$$

For $n = \frac{1}{2}$ the outer limit of the inner approximation is

$$v \sim \pi - 2\xi^{-1/2} + R'R\left(-\frac{1}{2}\xi^{3/2} + \left(-\frac{9}{2} + 2\ln\xi\right)\xi^{1/2}\right),$$

and for $n < \frac{1}{2}$ it becomes

$$v \sim \pi - 2\xi^{-n} + R'R\left(\frac{n}{2n-2}\xi^{2-n} + \frac{n(2n^2-n+1)}{4(1-n)^2(\frac{1}{2}-n)}\xi^{2-3n}\right).$$

3.3. A first try. To get a preliminary idea, let us first attempt to match without going to self-similar coordinates. (This turns out to produce the correct generic rate for $n \neq 1$; nongeneric cases have to be analyzed in self-similar coordinates.) Near $\theta = \pi$ the equation can be linearized to

$$w_t = \mathcal{L}_2 w \stackrel{\text{def}}{=} w_{rr} + \frac{1}{r}w_r - \frac{n^2}{r^2}w. \tag{3.3}$$

The stationary solution is $w = C_0 r^n + C_1 r^{-n}$, with C_0, $C_1 \in \mathbb{R}$, and this forms the inspiration for the formal solution

$$w = \sum_{i=0}^{\infty} \frac{d^i\alpha}{dt^i}\psi_i(r) + \sum_{i=0}^{\infty} \frac{d^i\beta}{dt^i}\chi_i(r),$$

for some functions $\alpha(t)$ and $\beta(t)$, and

$$\begin{aligned} &\psi_0 = r^n, && \mathcal{L}_2\psi_i = \psi_{i-1}, \quad i = 1, 2, \ldots, \\ &\chi_0 = r^{-n}, && \mathcal{L}_2\chi_i = \chi_{i-1}, \quad i = 1, 2, \ldots. \end{aligned}$$

The outer approximation becomes

$$\theta \sim \pi + \alpha(t)r^n + \beta(t)r^{-n} + \beta'(t)\frac{1}{4(1-n)}r^{-n+2}.$$

Matching yields (for $n > \frac{1}{2}$)

$$\begin{aligned} O(r^{-n}): &\qquad \beta \sim -2R^n, \\ O(r^{-n+2}): &\quad \beta'\frac{1}{4(1-n)} \sim \frac{n}{2n-2}R'R^{n-1}, \\ O(r^n): &\qquad \alpha \sim -E_n R'R^{1-n}, \end{aligned}$$

from which we conclude that

$$\begin{aligned} &R \sim \left[\frac{\beta_0}{2}(T-t)\right]^{1/n}, && n < 1, && \text{as } t \uparrow T, \\ &R \sim \left[\frac{(2-n)\alpha_0}{E_n}(T-t)\right]^{1/(2-n)}, && 1 < n < 2, && \text{as } t \uparrow T, \\ &R \sim \kappa e^{-\frac{\alpha_0}{E_2}t}, && n = 2, && \text{as } t \to \infty, \\ &R \sim \left(\frac{(n-2)\alpha_0}{E_n}t\right)^{-1/(n-2)}, && n > 2, && \text{as } t \to \infty, \end{aligned}$$

for some $\kappa > 0$ and $\alpha_0 = \alpha(T) > 0$ and $\beta_0 = \beta'(T) > 0$ (i.e., $\beta(t) \sim \beta_0(t-T)$). For $n < 1$ we have to separately consider the three cases ($n < \frac{1}{2}$, $n = \frac{1}{2}$, $n > \frac{1}{2}$) with different asymptotic behavior for the inner solutions; they all lead to results of the same form.

Notice that when we perform the same analysis for $n = 1$ we obtain

$$\begin{aligned} O(r^{-1}): &\quad \beta \sim -2R, \\ O(r\ln r): &\quad \frac{1}{2}\beta' \sim -R', \\ O(r): &\quad \alpha \sim R'(\ln R + 1). \end{aligned}$$

One deduces that $R \sim \alpha_0 \frac{(T-t)}{|\ln(T-t)|}$, which is the wrong asymptotic behavior (although it is almost right); see section 2. The artefact is caused by the fact that one a priori assumes the limit profile to be $\theta(T,r) \sim \pi + Cr$ for some $C \neq 0$, whereas this cannot be fixed a priori but should be determined by matching with the region in which θ is near π; for this one needs to analyze what happens at an intermediate scale, the self-similar one. When one performs the matching at the self-similar scale (with variable $y = r/(T-t)^{1/2}$), as will be done in section 3.4, it turns out that in the generic (codimension 0) case the solution is y^n ($1 < n < 2$) or y^{-n} ($n < 1$); hence the self-similar scale seems to have no influence. However, this region is crucial since it introduces a selection mechanism (the requirement that the solution does not grow exponentially for large y). Therefore, the analysis of the self-similar region does not appear to influence the result for $n \neq 1$ (in the generic scenario), but for $n = 1$ it corrects the result from the above naive approach.

Finally, it is important to note that our analysis thus far suggests that blowup occurs in infinite time for $n \geq 2$ and in finite time for $n < 2$. This difference causes us to investigate these cases separately.

3.4. The case $n < 2$: Finite time blowup. The analysis goes along the same lines as for $n = 1$. The inner approximation has been obtained in section 3.2. Let us here pay some extra attention to the outer solution. One could formulate this analysis in the same terms as used in section 2.3 for $n = 1$. The matching for $n \neq 1$ is easier because it turns out that only the dominant term needs to be taken into account. We can therefore use a slightly more straightforward approach.

We look for a self-similar solution to (3.3) of the form

$$w = (T-t)^{\gamma}\phi\left(\frac{r}{\sqrt{T-t}}\right), \tag{3.4}$$

where γ is not known a priori but has to be determined as part of the process. As the first boundary conditions we require that $\phi(y)$ does not grow exponentially for $y \to \infty$. The second boundary condition is different for $n > 1$ and $n < 1$. For $n > 1$ we require that $\phi(0) = 0$, while for $n < 1$ we require that $\phi = Cy^{-n} + o(y^n)$ for some $C \neq 0$. These boundary conditions are suggested by our preliminary results in section 3.3: the terms of order y^n and y^{-n} are dominant for $n > 1$ and $n < 1$, respectively. Another, equivalent, point of view is that this boundary condition is in fact a matching condition, as explained in section 2.4. For both $n > 1$ and $n < 1$ there is a sequence of self-similar solutions of the boundary value problem.

For $n > 1$ the first one is $\gamma = n/2$ and $\phi_0 = C_0 y^n$ with $C_0 \neq 0$; the second one is $\gamma = n/2 + 1$ and $\phi_1 = C_1 y^n(1 - \frac{1}{4n+4}y^2)$. In general, there is a family of solutions $\gamma = n/2 + k$ and $\phi_k = C_k y^n f_{k,n}(y^2)$ for $k = 0, 1, 2, \ldots$, where $f_{k,n}$ is a polynomial

of degree k and $f_{k,n}(0) = 1$. (We note that $f_{k,n}$ can be expressed in terms of a generalized Laguerre polynomial.) Since the inner approximation (3.2) thus needs to match into $C_0(T-t)^k r^n$ for some $C_0 \neq 0$, one obtains the matching condition

$$O(r^n): \quad C_0(T-t)^k \sim -E_n R'(t) R(t)^{1-n}.$$

We only use this one term since we have already seen in section 3.3 that it is the dominant one. Hence with codimension $k = 0, 1, 2, \ldots$ (and $1 < n < 2$)

$$R(t) = \kappa(T-t)^{(k+1)/(2-n)} \qquad \text{as } t \uparrow T$$

for some $\kappa > 0$, with limit profile

$$\theta(T, r) \sim \pi + A_{k,n}\kappa^{2-n} r^{n+2k} \qquad \text{as } r \downarrow 0$$

for some constant $A_{k,n}$, which can be calculated from the coefficient of the highest order term in the polynomial $f_{k,n}$ (e.g., $A_{0,n} = \frac{E_n}{4(n+1)(2-n)}$).

For $n < 1$ the first possibly relevant self-similar solution of the form (3.4) is $(T-t)^{-n/2} y^{-n}$, but this is simply r^{-n} and does not tend to 0 as $t \to T$, and hence it is not suitable (it does not correspond to blowup at $t = T$). The next is $\gamma = -n/2 + 1$ and $\phi_0 = \tilde{C}_0 y^{-n}(1 - \frac{1}{4(1-n)} y^2)$ with $\tilde{C}_0 \neq 0$. There is again a family of solutions $\gamma = -n/2 + k + 1$ and $\phi_k = \tilde{C}_k y^{-n} \tilde{f}_{k,n}(y^2)$, where $\tilde{f}_{k,n}$ is a polynomial of degree $k+1$ and $\tilde{f}_{k,n}(0) = 1$. (Again $f_{k,n}$ can be expressed in terms of a generalized Laguerre polynomial.) Because (3.2) thus needs to match into $\tilde{C}_0(T-t)^{k+1} r^{-n}$ for some $\tilde{C}_0 \neq 0$, one obtains the matching condition

$$O(r^{-n}): \quad \tilde{C}_0(T-t)^{k+1} \sim -2R^n,$$

and hence with codimension $k = 0, 1, 2, \ldots$ (and $0 < n < 1$)

$$R(t) \sim \kappa(T-t)^{(k+1)/n} \qquad \text{as } t \uparrow T$$

for some $\kappa > 0$, and the limit profile is

$$\theta(T, r) \sim \pi + A_{k,n}\kappa^n r^{2-n+2k} \qquad \text{as } r \downarrow 0$$

for some constant $A_{k,n}$, which can be calculated from the coefficient of the highest order term in the polynomial $\tilde{f}_{k,n}$ (e.g., $A_{0,n} = \frac{1}{2(1-n)}$).

Immediately after blowup, an inner layer near $r = 0$ appears. For $n < 1$ the leading order behavior in this layer is simpler than for $n = 1$, being exactly self-similar. In the generic case ($k = 0$) one finds

$$\theta(t, r) \sim \pi + d_n \kappa^n (t-T)^{(2-n)/2} g_n\left(\frac{r}{\sqrt{t-T}}\right) \qquad \text{as } t \downarrow T \text{ for small } r$$

for some $d_n > 0$ to be determined. Here

$$g_n(z) = \left(8n(1-n)z^{2-n} + 32n(1-n)^2 z^{-n}\right) \int_0^z \frac{s^{2n-1} e^{-s^2/4}}{(4(1-n) + s^2)^2}\, ds$$

is the solution of the linearized equation in self-similar coordinates with the property that $g_n(z) \sim z^n$ as $z \to 0$. It follows that $g_n(z) \sim 4^{n-1}\Gamma(n+1) z^{2-n}$ as $z \to \infty$. Matching with the limit profile at $t = T$ yields $d_n = \left[2^{2n-1}(1-n)\Gamma(n+1)\right]^{-1}$. Notice that for $1 < n < 2$ special treatment of the short time behavior after blowup is not necessary. (One would just find that $\theta(t, r) \sim \pi + (t-T)^{n/2} g_n(r/\sqrt{t-T})$, with $g_n(z) = \frac{E_n}{4(n+1)(2-n)} z^n$.)

3.5. The case $n \geq 2$: Infinite time blowup. For $n \geq 2$ blowup occurs in infinite time, and there are only two scales: an inner and a remote region. In the remote region the limit profile is now known. Namely, for boundary condition $\theta(t,1) = \theta_1 \in (\pi, 2\pi)$ the limit profile is

$$\lim_{t\to\infty} \theta(t,r) \sim \pi + 2\arctan\left(\tan\left(\frac{\theta_1 - \pi}{2}\right) r^n\right). \tag{3.5}$$

The special case $\theta_1 = \pi$ is dealt with in section 3.7, and $\theta_1 > 2\pi$ is discussed in section 3.6. For small r the limit profile (3.5) behaves as $\pi + 2\tan(\frac{\theta_1-\pi}{2})r^n$. For $\theta_1 \in (\pi, 2\pi)$ we define $\alpha_0 \stackrel{\text{def}}{=} 2\tan(\frac{\theta_1-\pi}{2})$.

Now the inner approximation (3.2) has to match into $\alpha_0 r$, and hence

$$O(r^n): \quad \alpha_0 \sim -E_n R'(t) R(t)^{1-n}.$$

One obtains

$$R \sim \begin{cases} \kappa e^{-\frac{\alpha_0}{E_2}t}, & n = 2, \\ \left(\frac{(n-2)\alpha_0}{E_n} t\right)^{-1/(n-2)}, & n > 2, \end{cases} \qquad \text{as } t \to \infty \tag{3.6}$$

for some $\kappa > 0$. Because the limit profile is known a priori (as opposed to $n < 2$), the unknown constant appears to leading order only when $n = 2$, where it is required due to translation invariance in time. Furthermore, we note that there are no degenerate cases; since blowup occurs as $t \to \infty$, it can be determined a priori which of the stationary states will be the final profile. This depends only on the value of θ_1 and $\theta(0,0)$, and since $\theta(0,0) \in \pi\mathbb{Z}$ is not a continuous parameter, degenerate (borderline) cases are not needed.

3.6. Multiple blowup. In certain situations it may happen that blowup occurs at several scales simultaneously (a so-called bubble tree; see also [17]), for example, when $\theta_1 \geq 2\pi$. For $n < 2$, i.e., finite time blowup, double (or multiple) blowup does not need to happen since the necessary jumps can occur at different instances, and simultaneous blowup is indeed not possible, at least for $n = 1$; see [12] (and our analysis reveals no possible finite time bubble tree). On the other hand, for $n \geq 2$ multiscale blowup necessarily happens if $\theta_1 \geq 2\pi$, since all blowup occurs as $t \to \infty$.

Let us take $\theta_1 \in (2\pi, 3\pi)$ as an example. The limit profile is now

$$\lim_{t\to\infty} \theta(t,r) = 2\pi + 2\arctan\left(\tan\left(\frac{\theta_1}{2}\right) r^n\right).$$

Define $\alpha_1 = 2\tan(\frac{\theta_1}{2})$ for $\theta_1 \in (2\pi, 3\pi)$. The two blowup rates are $R_1(t)$ and $R_2(t)$ for the jumps from 0 to π and from π to 2π, respectively, with $R_1 \ll R_2 \ll 1$. The first blowup rate $R_1(t)$ is defined as before, namely, so that $R_1(t)^n \theta(t,r) \sim 2r^n$ as $r \downarrow 0$ for all $t > 0$. The second blowup rate $R_2(t)$ cannot be defined in the same way, and instead we set

$$\theta(t, R_2(t)) = \frac{3\pi}{2} \qquad \text{for all } t \text{ close to blowup,}$$

so that in the limit we have for all $\rho > 0$

$$\lim_{t\to\infty} \theta(t, \rho R_2(t)) = \pi + 2\arctan\rho^n \qquad \text{and} \qquad \lim_{t\to\infty} \theta(t, \rho R_1(t)) = 2\arctan\rho^n.$$

The inner-inner region, $r = O(R_1(t))$, is as analyzed in section 3.2. The analysis of the inner region, $r = O(R_2(t))$, is the same except for the boundary condition on the left. Let $x = r/R_2(t)$; then $v(t,x) = u(t,r)$ behaves for large t as

$$v \sim \pi + 2\arctan x^n + R_2' R_2 \Psi,$$

where Ψ obeys

$$\Psi_{xx} + \frac{1}{x}\Psi_x - n^2 \frac{\cos(4\arctan x^n)}{x^2}\Psi = -\frac{2nx^n}{1+x^{2n}},$$

with "boundary" condition $\Psi(1) = 0$. The solution is

$$\Psi = \frac{(1 - x^{4n} - 4nx^{2n}\ln x)(A + \int_1^x \frac{s^{2n+1}}{(1+s^{2n})^2} ds) + x^{2n}\int_1^x \frac{s(s^{4n}-1+4ns^{2n}\ln s)}{(1+s^{2n})^2} ds}{x^n(1+x^{2n})}$$

with arbitrary $A \in \mathbb{R}$. For convenience we define

$$B_n = \int_0^1 \frac{s^{2n+1}}{(1+s^{2n})^2} ds.$$

One infers that

$$\begin{aligned} &\Psi \sim (A - B_n)x^{-n} && \text{for small } x, \\ &\Psi \sim \left(-\int_1^\infty \frac{s^{2n+1}}{(1+s^{2n})^2} ds - A\right) x^n + \frac{n}{2n-2} x^{-n+2} && \text{for large } x. \end{aligned}$$

Matching with the remote solution gives (recalling that $E_n = \int_0^\infty \frac{s^{2n+1}}{(1+s^{2n})^2} ds$)

$$O(r^n): \quad \alpha_1 \sim (B_n - E_n - A) R_2' R_2^{1-n};$$

hence

$$R_2 \sim \begin{cases} c_0 e^{-\frac{\alpha_1}{E_2 - B_2 + A} t}, & n = 2, \\ \left(\frac{(n-2)\alpha_1}{E_n - B_n + A} t\right)^{-1/(n-2)}, & n > 2, \end{cases} \tag{3.7}$$

for an arbitrary constant $c_0 > 0$.

Matching the inner-inner with the inner gives

$$O(r^{-n}): \qquad -2R_1^n \sim (A - B_n) R_2' R_2^{1+n}, \tag{3.8}$$

$$O(r^n): \qquad -E_n R_1' R_1^{1-n} \sim 2R_2^{-n}. \tag{3.9}$$

From (3.7) and (3.9) we deduce that

$$R_1 \sim \begin{cases} c_2 e^{-c_1 e^{\frac{2\alpha_1}{E_2 - B_2 + A} t}}, & n = 2, \\ \left(\frac{(n-2)(E_n - B_n + A)}{E_n \alpha_1 (n-1)}\right)^{-1/(n-2)} \left(\frac{(n-2)\alpha_1}{E_n - B_n + A} t\right)^{-(2n-2)/(n-2)^2}, & n > 2, \end{cases}$$

for arbitrary constants c_1, $c_2 > 0$ ($c_1 = \frac{2}{E_2} c_0$). From (3.8) we then conclude that $A = B_n$, and hence

$$R_1 \sim \begin{cases} c_2 e^{-c_1 e^{\frac{2\alpha_1}{E_2} t}}, & n = 2, \\ \left(\frac{n-2}{\alpha_1 (n-1)}\right)^{-1/(n-2)} \left(\frac{(n-2)\alpha_1}{E_n} t\right)^{-(2n-2)/(n-2)^2}, & n > 2, \end{cases}$$

for arbitrary constants c_1, $c_2 > 0$ (which only appear for $n = 2$), and $\alpha_1 = 2\tan(\frac{\theta_1}{2})$, $E_n = \frac{\pi}{2n^2 \sin(\frac{\pi}{n})}$. Notice the doubly exponential decay for $n = 2$; there are two unknown constants c_1 and c_2 in the leading order asymptotic expression because at both blowup scales there is a scaling invariance. While the first one could be attributed to translation invariance in time, the criticality of the case $n = 2$ is apparent in the appearance of a second unknown constant; for $n > 2$ there are no free constants in the leading order expression. One could generalize this to triple and higher multijumps, but we leave this to the puzzle-minded reader.

When one tries to perform the above analysis for $n < 2$, one readily encounters matching conditions which cannot be fulfilled. We therefore conjecture that there exist no bubble trees for $n < 2$; in particular, there are no finite time bubble trees.

3.7. Boundary condition $\theta_1 = \pi$. In the case of the special boundary condition $\theta(t, 1) = \pi$ we follow the argument of section 2.6. The remote solution w behaves as

$$w \sim \pi + s(t)(r^n - r^{-n}) + s'(t)\frac{1}{4}\left(\frac{1}{n-1}r^{2-n} + \frac{1}{n+1}r^{2+n} - \frac{2n}{n^2-1}r^n\right).$$

Matching this with the inner solution, we find

$$\begin{aligned} O(r^{-n}): &\qquad -s \sim -2R^n, \\ O(r^{-n+2}): &\quad \frac{1}{4(n-1)}s' \sim \frac{n}{2(n-1)}R'R^{n-1}, \\ O(r^n): &\qquad s \sim -E_n R' R^{1-n}. \end{aligned}$$

We obtain

$$R(t) \sim \left(\frac{4(n-1)}{E_n}t\right)^{-1/(2n-2)} \qquad \text{as } t \to \infty \qquad \text{for } n > 1.$$

For $n < 1$ matching suggests $R(t) \sim \left(\frac{4(1-n)}{E_n}(T-t)\right)^{1/(2-2n)}$. However, this does not provide consistent matching since it implies that $s' \gg s$. This suggests that we need a left boundary condition on the remote region of the form $w \sim Cr^{-n} + o(r^n)$ as $r \to 0$ for some $C \neq 0$. Hence we need to consider solutions of the form (compare to the degenerate case in section 2.6)

$$w \sim \pi + C_0 e^{-\nu_n^2 t} r^{-n}$$

for some $C_0 \neq 0$. Here ν_n is the first zero of the nth order singular Bessel function $\tilde{Y}_n$, which (for the occasion) is defined with the choice that $\tilde{Y}_n(r) \sim \tilde{C}r^{-n} + o(r^n)$, with $\tilde{C} \neq 0$ arbitrary. We remark that $\nu_n \to 0$ as $n \uparrow 1$. Matching now yields

$$O(r^{-n}): \quad C_0 e^{-\nu_n^2 t} \sim -2R^n,$$

and hence

$$R(t) \sim \kappa e^{-\frac{\nu_n^2}{n}t} \qquad \text{for } n < 1$$

for some $\kappa > 0$. To summarize, one finds that for $\theta_1 = \pi$ generically

$$R(t) \sim \begin{cases} \kappa e^{-\nu_n^2 t/n}, & n < 1, \\ e^{-2\sqrt{t}-5/4}, & n = 1, \\ \left(\frac{4(n-1)}{E_n}t\right)^{-1/(2n-2)}, & n > 1, \end{cases}$$

provided that no finite time blowup occurs (for $n < 2$), for example, when one takes initial data which lie entirely between 0 and π.

For $n < 2$ there is again a family of degenerate cases (cf. section 2.6); the blowup is at an exponential rate determined by the zeros of the Bessel functions $\tilde{Y}_n$ for $n < 1$ and J_n for $1 < n < 2$. For $n \geq 2$ degenerate scenarios do not exist. We leave the details to the diligent reader. For $n \geq 2$ the cases $\theta_1 = m\pi$, $m = 2, 3, \ldots$, involve multiple blowup, and the technique from the previous section may be used.

3.8. The infinite domain. We will now discuss how the results obtained so far have to be adapted when we consider an infinite domain, i.e., $r \in (0, \infty)$, instead of a finite one. In order to have a solution with finite energy $\mathcal{E}(t) = \pi \int_0^\infty \left(r\theta_r^2 + n^2 \frac{\sin^2\theta}{r}\right) dr$ the profile has to approach a multiple of π at a reasonably fast rate as $r \to \infty$; we denote this "boundary" condition by $\lim_{r\to\infty} \theta(t, r) = \tilde{\theta}_1 \in \pi\mathbb{Z}$. We focus on initial data which have compact support in the sense that $\theta(0, r) = m\pi$ for all sufficiently large r or, more generally, data which decay to $m\pi$ exponentially (thus, in particular, excluding algebraic decay).

Let us first discuss the case $\tilde{\theta}_1 = \pi$. There are several possibilities, depending on the initial data. For $n < 2$ a generic possibility is finite time blowup (see section 3.4). A priori, another possibility is that there is no blowup and that for large time the solution converges to one of the equilibria $\theta(r) = 2\arctan qr^n$ for some $q > 0$. For $n \geq 2$ this is in fact the only feasible scenario, and no blowup turns out to be a generic scenario for $1 < n < 2$; on the other hand, for $n \leq 1$ blowup always occurs (as is explained below). Regarding nongeneric possibilities, consider, for example, the parameter range $1 < n < 2$ and initial data which have one crossing with π (so that the finite time codimension 1 blowup scenario is not possible). We deduce that there should be at least one nongeneric infinite time blowup scenario which acts as the borderline between the two generic possibilities.

The large time behavior away from the origin is described in terms of self-similar variables $\tau = \ln t$ and $y = r/\sqrt{t} = e^{-\tau/2} r$, which leads to the linearized equation

$$\eta_\tau = \eta_{yy} + \left(\frac{1}{y} + \frac{y}{2}\right)\eta_y - \frac{n^2}{y^2}\,\eta.$$

We now analyze the codimension 0 and 1 scenarios for various ranges of n. (Higher codimension cases can be analyzed in a similar manner.)

For $n > 1$ the generic behavior is described by the solution

$$\eta = C_0 e^{-n\tau/2}\, y^{-n} \int_y^\infty s^{2n-1} e^{-s^2/4}\, ds \tag{3.10}$$

for some $C_0 \neq 0$; it decays faster than exponentially as $y \to \infty$, and as $y \to 0$ it asymptotically satisfies $\eta \sim C_0 2^{2n-1}\Gamma(n) e^{-n\tau/2} y^{-n} = C_0 2^{2n-1}\Gamma(n) r^{-n}$. Since matching into the inner solution (3.2) leads to $R(t) \to \kappa$ as $t \to \infty$ for some $\kappa > 0$ with limit profile $\theta_\infty(r) = 2\arctan(r/\kappa)^n$, the generic scenario corresponds to no blowup. To be more precise, for $n > 1$ the matching conditions are

$$\begin{aligned} O(r^{-n}): &\qquad C_0 2^{2n-1}\Gamma(n) \sim -2R^n, \\ O(r^n): &\qquad -\tfrac{1}{2n} C_0 t^{-n} \sim -E_n R' R^{1-n}. \end{aligned}$$

Hence $R(t) \sim \kappa + Q_{q,n} t^{1-n}$ as $t \to \infty$, where

$$Q_{q,n} = \left[2^{2n-1}\kappa^{1-2n}(n-1)\Gamma(n+1)E_n\right]^{-1} > 0.$$

In between the generic possibilities of no blowup and finite time blowup there is a degenerate infinite time blowup scenario. For this codimension 1 case we find the outer approximation (writing $\zeta(\tau, y) = \theta(t, r)$)

$$\zeta(\tau, y) \sim \pi + C_1 e^{-(2+n)\tau/2}\, y^n\, e^{-y^2/4} \qquad \text{as } \tau \to \infty$$

for some $C_1 \neq 0$. Matching this with the inner solution, we infer that for $1 < n < 2$ the codimension 1 blowup rate is

$$R(t) \sim \kappa t^{-n/(2-n)} \qquad \text{as } t \to \infty \qquad \text{for } 1 < n < 2 \tag{3.11}$$

for some $\kappa > 0$. Away from the origin the decay towards π is at algebraic rate $O(t^{-1-n})$. For $n \geq 2$ matching in the nongeneric case is impossible (looking at (3.11), one could anticipate this), implying that for $\tilde{\theta}_1 = \pi$ there is never blowup when $n \geq 2$.

The case $n = 1$ is again a borderline one; we find that (3.10) again describes the generic behavior in the outer region, but we need C_0 to depend on τ in a nonexponential manner. Matching this with the inner solution (2.3) gives matching conditions (with $\tilde{R}(\tau) = R(t)$)

$$\begin{aligned} O(y^{-1}): &\qquad 2C_0(\tau)e^{-\tau/2} \sim -2e^{-\tau/2}\tilde{R}, \\ O(y \ln y): &\qquad C_0'(\tau)e^{-\tau/2} \sim -e^{-\tau/2}\tilde{R}', \\ O(y): &\qquad -\frac{1}{2}C_0(\tau)e^{-\tau/2} \sim e^{-\tau/2}\tilde{R}'\left(\ln \tilde{R} - \frac{\tau}{2} + 1\right). \end{aligned}$$

This generic scenario thus describes blowup at rate

$$R(t) \sim \frac{\kappa}{\ln t} \qquad \text{as } t \to \infty \qquad \text{for } n = 1$$

for some $\kappa > 0$. (Since the problem on the infinite domain is scaling invariant, a multiplicative constant, whose value depends on the initial data, must again be present.) For $r = O(1)$ the solution approaches π at rate $O(1/\ln t)$ as $t \to \infty$. The outer approximation in the codimension 1 case is

$$\zeta(\tau, y) \sim \pi + e^{-3\tau/2} e^{-y^2/4}\big[\sigma(\tau)y + \sigma'(\tau)(4y^{-1} - 2y \ln y)\big] \qquad \text{as } \tau \to \infty$$

for some $\sigma(\tau)$, and matching gives the blowup rate $R \sim \kappa t^{-1}(\ln t)^{-4/3}$ as $t \to \infty$ (i.e., a logarithmically corrected version of (3.11) with $n = 1$), where $\kappa > 0$ is arbitrary. Away from the origin the rate of decay towards π is of order $O(t^{-1}(\ln t)^{-4/3})$.

For $n < 1$ the outer approximation described by (3.10) does not lead to consistent matching; we have seen previously that for $n < 1$ the outer solution should behave as $\tilde{C}y^{-n} + o(y^n)$ for small y with $\tilde{C} \neq 0$, which is not satisfied by (3.10). Therefore, for $n < 1$ the outer approximation in the generic case is

$$\zeta(\tau, y) \sim \pi + C_2 e^{-(2-n)\tau/2}\, y^{-n}\, e^{-y^2/4} \qquad \text{as } \tau \to \infty$$

for some $C_2 \neq 0$. Matching with the inner solution, we find that generically

$$R(t) \sim \kappa t^{-(1-n)/n} \qquad \text{as } t \to \infty \qquad \text{for } n < 1$$

for some $\kappa > 0$. Away from the origin the decay towards π as $t \to \infty$ is at algebraic rate $O(t^{-1})$. For the codimension 1 scenario we find $R \sim \kappa t^{-(2-n)/n}$ as $t \to \infty$ for some $\kappa > 0$.

We note that, although the infinite domain allows scaling invariance, spreading of the form $\theta(t,r) = \Theta(r/\sqrt{t})$ is seen to be impossible by an argument analogous to that given at the end of section 1. On the other hand, taking initial data with $\theta(t_0,r) \sim \hat{C}r^{-a}$ as $r \to \infty$ for some $\hat{C} \neq 0$, where $0 < a < n$, spreading at a rate slower than the self-similar one will occur as $t \to \infty$. The matching conditions in that case imply that the term of order r^{-n} is dominant for all $n > 0$, which yields $R(t) \sim \kappa t^{(n-a)/2n}$ as $t \to \infty$ for some $\kappa > 0$ and any $0 < a < n$; notice that $\frac{n-a}{2n} < \frac{1}{2}$ so that the self-consistency condition $R'R \to 0$ as $t \to \infty$ holds. For $a = n$ (i.e., $\theta(t_0,r) \sim \hat{C}r^{-n}$ as $r \to \infty$, which is the same rate as a stationary solution) no blowup occurs, while for $a > n$ there is either blowup (for $n \leq 1$) or no blowup (for $n > 1$), but we will not pursue the issue of algebraically decaying initial data any further.

The analysis is similar for $\lim_{r\to\infty} \theta(t,r) = \tilde{\theta}_1 = m\pi$, $m = 2,3,\ldots$. For $n < 2$ a (finite) number of finite time jumps essentially reduces the situation to the case $\tilde{\theta}_1 = \pi$. For $n \geq 2$ and boundary value $\tilde{\theta}_1 = 2\pi$, blowup will happen as $t \to \infty$ and one of the stationary states $\theta_\infty(r) = \pi + 2\arctan qr^n$ for some $q > 0$ is selected. The analysis is completely analogous to the finite domain case $\theta_1 \in (\pi, 2\pi)$ discussed in section 3.5 (notice that it thus differs from the finite domain with $\theta_1 = 2\pi$). The result is the same as in section 3.5 except that the constant $q = \alpha_0$ cannot be determined a priori in the case of an infinite domain (and it should not be since the equation has a scaling invariance). Hence, the asymptotic blowup rate is given by (3.6), the only alteration being that, in the case of an infinite domain, $\alpha_0 > 0$ is an arbitrary constant whose value depends on the initial data. For $\tilde{\theta}_1 = m\pi$ with $m = 3,4,\ldots$ (and $n \geq 2$) multiscale blowup occurs (see section 3.6), and the analysis of the infinite domain is analogous to that of the finite domain with $\theta_1 \in ((m-1)\pi, m\pi)$.

3.9. Jumping back. Finally, we analyze the possibility of reverse jumps for (3.1). Analogous to section 2.7, the profile at any time generically behaves as Cr^n for small r with $C \neq 0$. Hence in the outer region, in self-similar coordinates $\tau = \ln(t-t_0)$ and $y = e^{-\tau/2}r$, we look for a solution of the form $e^{n\tau/2}\psi(y)$. The linear equation for ψ has as solution

$$\psi = C_0 y^n + C_1 y^n \int_y^\infty e^{-s^2/4} s^{-2n-1} ds,$$

with arbitrary constants C_0, $C_1 \in \mathbb{R}$, and $y_0 > 0$. The inner limit of the outer approximation thus becomes

$$\zeta \sim \pi + e^{n\tau/2}\left[\alpha(\tau)y^n + \beta(\tau)\left(\frac{1}{2n}y^{-n} - \frac{1}{8(n-1)}y^{2-n}\right)\right] \qquad \text{for small } y \text{ and } -\tau \gg 1.$$

Matching with the inner solution gives (writing $\tilde{R}(\tau) = R(t)$)

$$\begin{aligned} O(y^{-n}): &\qquad \frac{1}{2n}\beta e^{n\tau/2} \sim -2\tilde{R}^n e^{-n\tau/2}, \\ O(y^{2-n}): &\quad -\frac{1}{8(n-1)}\beta e^{n\tau/2} \sim \frac{n}{2(n-1)}\tilde{R}'\tilde{R}^{n-1}e^{-n\tau/2}, \\ O(y^n): &\qquad \alpha e^{n\tau/2} \sim -E_n \tilde{R}'\tilde{R}^{1-n}e^{(-1+n/2)\tau}. \end{aligned}$$

One concludes that for $n > 1$ the term of order y^n is dominant, and hence as $t \downarrow t_0$

$$R \sim \left[\frac{\ell(2-n)}{E_n}(t-t_0)\right]^{1/(2-n)} \qquad \text{for } 1 < n < 2$$

for some $\ell > 0$, the limit profile being $\theta(t_0, r) \sim \pi - \ell r^n$. Considerations about nongeneric cases are analogous to those in section 2.7; for the codimension 1 case one finds that

$$R \sim \left[\frac{2\ell(n+1)(2-n)}{E_n}\right]^{1/(2-n)} (t-t_0)^{2/(2-n)} \qquad \text{as } t \downarrow t_0$$

for some $\ell > 0$ with $\theta(t_0, r) \sim \pi - \ell r^{2+n}$ as $r \to 0$. For $n \geq 2$ no consistent matching is found, implying the rather strong result that reverse jumps do not seem possible. Notice that these conclusions are in line with what one could expect from section 3.3.

For $n < 1$ the term of order y^{-n} is dominant; hence the matching conditions give $R \sim \omega_n(t - t_0)$ as $t \downarrow t_0$ for some $\omega_n > 0$, which can be determined using the limit profile $\theta(t_0, r) \sim \pi - \ell r^n$ as $r \to 0$. Since the inner limit of the outer approximation behaves as $(\zeta - \pi)e^{-n\tau/2} \sim -4n\omega_n^n y^{-n} + o(y^n)$ as $y \to 0$, the outer limit of the outer approximation becomes

$$\zeta \sim \pi - 4n\omega_n^n e^{n\tau/2} y^n \left[-2^{-2n-1}\Gamma(-n) + \int_y^\infty e^{-s^2/4} s^{-2n-1} ds\right].$$

Hence as $t \downarrow t_0$

$$R \sim \left(\frac{\ell\, 2^{2n-1}}{\Gamma(1-n)}\right)^{1/n} (t - t_0) \qquad \text{for } n < 1$$

for some ℓ with $\theta(t_0, r) \sim \pi - \ell r^n$ as $r \downarrow 0$. Nongeneric cases can also be considered; for example, for the codimension 1 scenario one finds

$$R \sim \left(\frac{\ell\, 2^{2n+1}}{\Gamma(1-n)}\right)^{1/n} (t - t_0)^{(n+1)/n} \qquad \text{as } t \downarrow t_0$$

for some ℓ with $\theta(t_0, r) \sim \pi - \ell r^{n+2}$ as $r \downarrow 0$.

Finally, we do not find any self-consistent scenarios for reverse jumps which increase the energy by more than 4π, i.e., no reverse bubble trees for any n. (This is analogous to the conclusion in section 3.6 that there are no normal (energy-decreasing) bubble trees for $n < 2$; for $n \geq 2$ there are normal bubble trees but they are of the infinite time blowup type.)

4. Conclusion. We have analyzed the blowup rate in the harmonic map heat flow in a family of symmetric settings, leading to a parabolic problem in one space dimension:

$$\theta_t = \theta_{rr} + \frac{1}{r}\theta_r - n^2 \frac{\sin 2\theta}{2r^2}, \qquad 0 < r < 1, \tag{4.1}$$

with boundary conditions $\theta(t, 0) \in \pi\mathbb{Z}$ and $\theta(1, t) = \theta_1$. Here $n > 0$ is a parameter, and it corresponds to a well-defined physical situation (e.g., in the context of aligned nematic liquid crystals) when $n = 1, 2, 3, \ldots$. The initial value problem has a unique energy-decreasing solution, the energy being $\mathcal{E}(t) = \pi \int_0^1 (\theta_r^2 + r^{-2} n^2 \sin^2\theta) r dr$. Equation (4.1) is the gradient flow associated with this energy.

Without loss of generality we may assume that $\theta(0, 0) = 0$. During the evolution of a solution the value of θ at the origin may jump at some time(s) $t = T \in (0, \infty]$.

As the time approaches T the quantity $\lim_{r\downarrow 0} r^{-n}\theta(t,r)$ blows up. In this paper we have determined the asymptotic behavior of the blowup rate $R(t)$, defined by

$$R(t)^n\theta(t,r) \sim 2r^n \quad \text{as } r \downarrow 0 \qquad \text{for all } t \text{ up to the blowup time,}$$

using formal matched asymptotic expansions. After rescaling with this blowup rate, the profile approaches a harmonic map (a stationary state):

$$\lim_{t\uparrow T} \theta(t, \xi R(t)) = 2\arctan \xi^n \qquad \text{for all fixed } \xi > 0.$$

Since our results suggest important differences between $n < 2$ and $n \geq 2$, let us first summarize the behavior for $n < 2$. The generic behavior for $\theta_1 > \pi$ is

$$\begin{aligned} n < 1: &\qquad R \sim \kappa\,(T-t)^{1/n} && \text{as } t \uparrow T,\\ n = 1: &\qquad R \sim \kappa\,\frac{T-t}{|\ln(T-t)|^2} && \text{as } t \uparrow T,\\ 1 < n < 2: &\qquad R \sim \kappa\,(T-t)^{1/(2-n)} && \text{as } t \uparrow T, \end{aligned}$$

where $\kappa > 0$ is an arbitrary constant and $T < \infty$ is the time of blowup. There is also a countable family of degenerate cases. We find that with codimension k $(k = 0, 1, 2, \dots)$

$$\begin{aligned} n < 1: &\qquad R \sim \kappa\,(T-t)^{(k+1)/n} && \text{as } t \uparrow T,\\ n = 1: &\qquad R \sim \kappa\,\frac{(T-t)^{k+1}}{|\ln(T-t)|^{(2k+2)/(2k+1)}} && \text{as } t \uparrow T,\\ 1 < n < 2: &\qquad R \sim \kappa\,(T-t)^{(k+1)/(2-n)} && \text{as } t \uparrow T, \end{aligned}$$

where $\kappa > 0$ is again an arbitrary constant. In the codimension k scenario the profile $\theta(t,r)$ just before blowup has $k+1$ intersections with π, which approach the origin as $t \uparrow T$. Countable families of nongeneric blowup rates are encountered in a wide variety of problems; see [10] for an illustrative example.

For boundary data $\theta_1 = \pi$, blowup can occur (in finite time) via the scenario described above, but there is another generic blowup behavior (in infinite time) of the form

$$\begin{aligned} n < 1: &\qquad R \sim \kappa\, e^{-\frac{\nu_n^2}{n}t} && \text{as } t \to \infty,\\ n = 1: &\qquad R \sim e^{-2\sqrt{t}-5/4} && \text{as } t \to \infty,\\ 1 < n < 2: &\qquad R \sim \left(\frac{4(n-1)}{E_n}t\right)^{-1/(2n-2)} && \text{as } t \to \infty, \end{aligned}$$

with arbitrary constant $\kappa > 0$. Here $E_n = \frac{\pi}{2n^2\sin(\frac{\pi}{n})}$ and ν_n is the first zero of the Bessel function $\tilde{Y}_n$ (see section 3.7 for details). There is also a family of nongeneric infinite time blowup possibilities.

For $n \geq 2$ all blowup occurs as $t \to \infty$, and there is always a unique blowup scenario. For $\theta_1 \in (0, \pi)$ blowup never occurs (whereas for $n < 2$ this depends on the initial profile). For $\theta_1 \in (\pi, 2\pi)$ one finds

$$\begin{aligned} n = 2: &\qquad R \sim \kappa\, e^{-\frac{\alpha_0}{E_2}t} && \text{as } t \to \infty,\\ n > 2: &\qquad R \sim \left(\frac{(n-2)\alpha_0}{E_n}t\right)^{-1/(n-2)} && \text{as } t \to \infty, \end{aligned}$$

where $\kappa > 0$ is arbitrary, $E_n = \frac{\pi}{2n^2 \sin(\frac{\pi}{n})}$, and $\alpha_0 = \tan(\frac{\theta_1 - \pi}{2})$. No other (nongeneric) scenarios are found. In the case of boundary data $\theta_1 = \pi$ the result is

$$n \geq 2: \qquad R \sim \left(\frac{4(n-1)}{E_n} t \right)^{-1/(2n-2)} \qquad \text{as } t \to \infty,$$

which is the same expression as for $1 < n < 2$.

For $\theta_1 \geq 2\pi$ there is blowup over two or more scales (the number being known a priori from the value of θ_1). For $2\pi \leq \theta < 3\pi$ there is blowup over two scales; these are the simplest examples of so-called bubble trees. For $\theta_1 = 2\pi$ the blowup rate in the innermost region is

$$\begin{aligned} n = 2: &\qquad R \sim \kappa\, e^{-4E_2^{-2} t^2} && \text{as } t \to \infty, \\ n > 2: &\qquad R \sim \left(\frac{n-2}{3n-2} \right)^{-1/(n-2)} \left(\frac{4(n-1)}{E_n} t \right)^{-(3n-2)/(2(n-1)(n-2))} && \text{as } t \to \infty, \end{aligned}$$

where $\kappa > 0$ is arbitrary. For $\theta_1 \in (2\pi, 3\pi)$ we obtain (for the innermost blowup)

$$\begin{aligned} n = 2: &\qquad R \sim \kappa_2\, e^{-\kappa_1 e^{\frac{2\alpha_1}{E_2} t}} && \text{as } t \to \infty, \\ n > 2: &\qquad R \sim \left(\frac{n-2}{\alpha_1 (n-1)} \right)^{-1/(n-2)} \left(\frac{(n-2)\alpha_1}{E_n} t \right)^{-(2n-2)/(n-2)^2} && \text{as } t \to \infty, \end{aligned}$$

where $\kappa_1, \kappa_2 > 0$ are arbitrary and $\alpha_1 = 2\tan(\frac{\theta_1}{2})$. Analogous results hold for other values of $\theta_1 \geq 3\pi$.

We now describe the global picture suggested by these formal results. Since the equation is invariant under the discrete symmetries $\theta \mapsto -\theta$ and $\theta \mapsto \theta + \pi$, we may without loss of generality assume that $\theta(0,0) = 0$ and $\theta_1 \geq 0$. For convenience of notation, let the stationary solution $2\arctan(r^n \tan\frac{\theta_1}{2})$ be denoted by $\vartheta_{\theta_1}(r)$. We make a subdivision depending on the value of the boundary data θ_1.

Case 1: $0 \leq \theta_1 < \pi$. No blowup occurs for $n \geq 2$ and the limit profile $\lim_{t\to\infty} \theta(t,r) = \theta_\infty(r) = \vartheta_{\theta_1}(r)$ for all fixed $r > 0$. For $n < 2$ there is, depending on the initial data, either no blowup or blowup at a finite set of finite time moments $\{T_i\}_{i=1}^K$ for some integer K; if $\theta_1 = 0$, then K must be even. At each blowup time T_i the value of θ at the origin jumps by $\pm\pi$, and an amount $\mathcal{E}(T_i) - \lim_{t\uparrow T_i} = 4\pi n$ of energy is lost (a sphere bubbles off). This may happen via either a generic or a nongeneric scenario. The blowup instances T_i are all different, and there is no blowup as $T \to \infty$. If K is even, then the limit profile is $\theta_\infty(r) = \vartheta_{\theta_1}(r)$, while if K is odd, it is $\theta_\infty(r) = \pi - \vartheta_{\pi-\theta_1}(r)$. The number of jumps K is a priori bounded from above by $\frac{1}{4\pi n} \max\{\mathcal{E}(0) - \mathcal{E}_{\theta_1}, \mathcal{E}(0) - \mathcal{E}_{\pi-\theta_1}\}$, where $\mathcal{E}(0)$ is the energy of the initial data and $\mathcal{E}_{\theta_1}$ is the energy of the stationary state ϑ_{θ_1}; if $\mathcal{E}(0) < \mathcal{E}_{\pi-\theta_1} + 4\pi n$, then no jump can occur.

Case 2: $\theta_1 = \pi$. For any $n > 0$ the limit profile is $\theta_\infty(r) \equiv \pi$, and blowup has to occur at at least one time $T \in (0, \infty]$. For $n \geq 2$ blowup occurs only as $t \to \infty$, and an energy loss of $4\pi n$ occurs in this limit (i.e., $\lim_{t\to\infty} \mathcal{E}(t) = 4\pi n$, while $\mathcal{E}(\theta_\infty) = 0$). There is only one possible blowup rate. For $n < 2$ there is a finite set of time moments $\{T_i\}_{i=1}^K$ for some integer K with the same properties as for $\theta_1 < \pi$, except that K is odd (and thus $K \geq 1$) and one of the T_i may be equal to ∞, in which case the energy

jump at infinity is $4\pi n$ (i.e., the same as for a finite time jump). Both the finite time and the infinite time blowup can happen via a generic or nongeneric scenario.

Case 3: $\pi < \theta_1 < 2\pi$. Blowup has to occur at at least one time $T \in (0, \infty]$. For $n \geq 2$ there is blowup only as $t \to \infty$, and the limit profile is $\theta_\infty(r) = \pi + \vartheta_{\theta_1 - \pi}(r)$. The energy loss at infinity is $4\pi n$, and there is only one possible blowup rate. For $n < 2$ the scenario is the same as for $\theta < \pi$ (in particular, there is no infinite time blowup), with the adaptation that $K \geq 1$. If K is odd, then the limit profile is $\theta_\infty(r) = \pi + \vartheta_{\theta_1 - \pi}(r)$, while if K is even, it is $\theta_\infty(r) = 2\pi - \vartheta_{2\pi - \theta_1}(r)$.

Case 4: $\theta_1 = m\pi$, $m = 2, 3, \ldots$. For any $n > 0$ the limit profile is $\theta_\infty(r) \equiv k\pi$, and blowup has to occur at at least one time moment $T \in (0, \infty]$. For $n \geq 2$ blowup occurs only as $t \to \infty$, and an energy loss of $4\pi nm$ occurs in this limit. There is only one possible blowup rate, and blowup occurs over m different scales (a so-called bubble tree, the current analysis furnishing simple concrete examples of how such behavior can occur); there is a unique blowup scenario. For $n < 2$ there is a finite set of time moments $\{T_i\}_{i=1}^K$ for some integer K with the same properties as for $\theta_1 = \pi$ (one of the T_i can be equal to ∞), except that the number of blowup times $K \geq m$ and $K - m$ is always even. There is no bubble tree, and the energy loss at each T_i is $4\pi n$.

Case 5: $\theta_1 > 2\pi$, $\theta_1 \notin \pi\mathbb{Z}$. Blowup has to occur at at least one time $T \in (0, \infty]$. For $n \geq 2$ there is blowup only as $t \to \infty$, and the limit profile is $\theta_\infty(r) = M\pi + \vartheta_{\theta_1 - M\pi}(r)$, where M is the largest integer smaller than θ_1. There is blowup over M different scales, and the energy loss at infinity is $4\pi nM$ (i.e., a bubble tree). There is just one possible scenario. For $n < 2$ the scenario is the same as for $\pi < \theta < 2\pi$, with the adaptation that $K \geq M$. If $K - M$ is even, then the limit profile is $\theta_\infty(r) = M\pi + \vartheta_{\theta_1 - M\pi}(r)$, while if K is even, it is $\theta_\infty(r) = (M+1)\pi - \vartheta_{(M+1)\pi - \theta_1}(r)$. Again, at each blowup time one quantum of energy (i.e., $4\pi n$) is lost.

In the case of an infinite domain $r \in (0, \infty)$ and boundary conditions at infinity $\lim_{r \to \infty} \theta(t, r) = \tilde{\theta}_1 = m\pi$, $m = 0, 1, 2, \ldots$ (in order for profiles to have finite energy), we restrict our attention to initial data that approach $\tilde{\theta}_1$ sufficiently fast as $r \to \infty$. We again describe the results for $n \geq 2$ and $n < 2$ separately.

For $n \geq 2$ and $\tilde{\theta}_1 = m\pi$, $m = 1, 2, \ldots$, the situation is very similar to that of a finite domain with $\theta_1 \in ((m-1)\pi, m\pi)$. (For $m = 0$ there is no blowup, and the solution converges to 0 uniformly as $t \to \infty$.) Blowup occurs over $m - 1$ scales, and the limit profile is a stationary state $\theta_\infty(r) = (m-1)\pi + 2\arctan qr^n$ for some $q > 0$ (which depends on the initial data). The blowup rate is the same as for the finite domain described previously (with q replacing the constants α_0 and α_1). There are no nongeneric scenarios.

For $n < 2$ the solution has $K \geq m$ blowup times (cf. the finite domain case), one of which may be infinity. If $K - m$ is odd, then the limit profile is $\theta_\infty(r) = (m \pm 1)\pi \mp 2\arctan qr^n$ for some $q > 0$, while if $K - m$ is even, then $\theta_\infty(r) = m\pi$. The finite time blowup rates are the same as for the finite domain (including the possibility of nongeneric finite time blowup). Whereas for $1 < n < 2$ the blowup times are all generically finite (and there is thus no generic blowup as $t \to \infty$), for $n \leq 1$ infinite time blowup is generic and the rate is

$$\begin{aligned} n < 1: &\qquad R \sim \kappa\, t^{-(1-n)/n} &&\text{as } t \to \infty, \\ n = 1: &\qquad R \sim \kappa\, (\ln t)^{-1} &&\text{as } t \to \infty, \end{aligned}$$

with $\kappa > 0$ arbitrary. For $1 < n < 2$ infinite time blowup can occur with codimen-

sion 1; the corresponding blowup rate is

$$\begin{aligned} n<1: &\qquad R\sim \kappa\, t^{-(2-n)/n} &&\text{as } t\to\infty,\\ n=1: &\qquad R\sim \kappa\, t^{-1}(\ln t)^{-4/3} &&\text{as } t\to\infty,\\ 1<n<2: &\qquad R\sim \kappa\, t^{-n/(2-n)} &&\text{as } t\to\infty \end{aligned}$$

for arbitrary $\kappa>0$. There is again a countable family of infinite time blowup scenarios with higher codimension.

Finally, when one allows for solutions that do not necessarily have decreasing energy (thereby introducing nonuniqueness), then, depending on n, jumps can occur in which the energy increases. The physical interpretation is that the energy stored in the origin at a forward (energy-decreasing) jump is released.

For $n<2$ reverse jumps can happen at any time. In the nominally generic (codimension 0) case we have

$$\begin{aligned} n<1: &\qquad R\sim \left(\frac{\ell\, 2^{2n-1}}{\Gamma(1-n)}\right)^{1/n}(t-t_0) &&\text{as } t\downarrow t_0,\\ n=1: &\qquad R\sim 2\ell\,\frac{t-t_0}{|\ln(t-t_0)|} &&\text{as } t\downarrow t_0,\\ 1<n<2: &\qquad R\sim \left[\frac{\ell(2-n)}{E_n}(t-t_0)\right]^{1/(2-n)} &&\text{as } t\downarrow t_0, \end{aligned}$$

where $\ell>0$ with $\theta(t_0,r)\sim\pi-\ell r^n$. For the codimension 1 scenario one finds

$$\begin{aligned} n<1: &\qquad R\sim \left(\frac{\ell\, 2^{2n+1}}{\Gamma(1-n)}\right)^{1/n}(t-t_0)^{(n+1)/n} &&\text{as } t\downarrow t_0,\\ n=1: &\qquad R\sim \frac{8}{3}\ell\,\frac{(t-t_0)^2}{|\ln(t-t_0)|} &&\text{as } t\downarrow t_0,\\ 1<n<2: &\qquad R\sim \left(\frac{2\ell(n+1)(2-n)}{E_n}\right)^{1/(2-n)}(t-t_0)^{2/(2-n)} &&\text{as } t\downarrow t_0, \end{aligned}$$

where $\ell>0$ with $\theta(t_0,r)\sim\pi-\ell r^{n+2}$. It is conjectured in [2] that a physical system selects the codimension 1 scenario to release the energy stored in the origin. When a forward and a reverse jump occur at the same instant (cf. [18]), the rate is given by (2.17).

For $n\geq 2$ no reverse jumps are possible; this is not surprising since energy can be stored in the origin only as $t\to\infty$, and thus none is available for release.

Acknowledgments. We would like to thank Sigurd Angenent, Marek Fila, Rein van der Hout, and Giles Richardson for a series of pleasant and enlightening discussions.

REFERENCES

[1] S.B. Angenent, J. Hulshof, and H. Matano, *Asymptotics for Gradient Blow-Up in Equivariant Harmonic Map Flows from D^2 to S^2*, in preparation.

[2] M. Bertsch, R. Dal Passo, and R. van der Hout, *Nonuniqueness for the heat flow of harmonic maps on the disk*, Arch. Ration. Mech. Anal., 161 (2002), pp. 93–112.

[3] M. Bertsch, P. Podio-Guidugli, and V. Valente, *On the dynamics of deformable ferromagnets:* I. *Global weak solutions for soft ferromagnets at rest*, Ann. Mat. Pura Appl., 179 (2001), pp. 331–360.

[4] F. Bethuel, H. Brezis, and F. Hélein, *Ginzburg–Landau Vortices*, Birkhäuser-Verlag, Berlin, 1994.
[5] K.-C. Chang, W.-Y. Ding, and R. Ye, *Finite-time blow-up of the heat flow from surfaces*, J. Differential Geom., 36 (1992), pp. 507–515.
[6] A. DeSimone and P. Podio-Guidugli, *On the continuum model of deformable ferromagnetic solids*, Arch. Ration. Mech. Anal., 136 (1996), pp. 201–233.
[7] A. Freire, *Uniqueness of the harmonic map flow from surfaces to general targets*, Comment Math. Helv., 70 (1995), pp. 310–338.
[8] V.A. Galaktionov and J.L. Vázquez, *The problem of blow-up in nonlinear parabolic equations*, Discrete Contin. Dyn. Syst. Ser. A, 8 (2002), pp. 399–433.
[9] R.M. Hardt, *Singularities of harmonic maps*, Bull. Amer. Math. Soc., 34 (1997), pp. 15–34.
[10] M.A. Herrero and J.J.L. Velázquez, *On the melting of ice balls*, SIAM J. Math. Anal., 28 (1997), pp. 1–32.
[11] R. van der Hout, *Flow alignment in nematic liquid crystals in flows with cylindrical symmetry*, Differential Integral Equations, 14 (2001), pp. 189–211.
[12] R. van der Hout, *On the nonexistence of finite time bubble trees in symmetric harmonic map heat flows from the disk to the* 2*-sphere*, J. Differential Equations, 192 (2003), pp. 188–201.
[13] J. Hulshof, J.R. King, and M. Bowen, *Intermediate asymptotics of the porous medium equation with sign changes*, Adv. Differential Equations, 6 (2001), pp. 1115–1152.
[14] G. Guidone Peroli and E.G. Virga, *Nucleation of topological dipoles in nematic liquid crystals*, Comm. Math. Phys., 200 (1999), pp. 195–210.
[15] M. Struwe, *Variational Methods*, Springer-Verlag, New York, Berlin, 1990.
[16] M. Struwe, *Geometric evolution problems*, in Nonlinear Partial Differential Equations in Differential Geometry, IAS Park City Math. Ser. 2, AMS, Providence, RI, 1996, pp. 257–339.
[17] P. Topping, *An example of a nontrivial bubble tree in the harmonic map heat flow*, in Harmonic Morphisms, Harmonic Maps and Related Topics, C.K. Anand, P. Baird, E. Loubeau, and J.C. Wood, eds., CRC Press, Boca Raton, FL, 1999, pp. 185–191.
[18] P. Topping, *Reverse bubbling and nonuniqueness in the harmonic map flow*, Internat. Math. Res. Notices, 101 (2002), pp. 505–520.

SIAM J. APPL. MATH.
Vol. 63, No. 5, pp. 1718–1736

BACKSCATTERING AND NONPARAXIALITY ARREST COLLAPSE OF DAMPED NONLINEAR WAVES*

G. FIBICH†, B. ILAN†, AND S. TSYNKOV‡

Abstract. The critical nonlinear Schrödinger equation (NLS) models the propagation of intense laser light in Kerr media. This equation is derived from the more comprehensive nonlinear Helmholtz equation (NLH) by employing the paraxial approximation and neglecting the backscattered waves. It is known that if the input power of the laser beam (i.e., L_2 norm of the initial solution) is sufficiently high, then the NLS model predicts that the beam will self-focus to a point (i.e., collapse) at a finite propagation distance. Mathematically, this behavior corresponds to the formation of a singularity in the solution of the NLS. A key question which has been open for many years is whether the solution to the NLH, i.e., the "parent" equation, may nonetheless exist and remain regular everywhere, particularly for those initial conditions (input powers) that lead to blowup in the NLS. In the current study we address this question by introducing linear damping into both models and subsequently comparing the numerical solutions of the damped NLH (boundary-value problem) with the corresponding solutions of the damped NLS (initial-value problem) for the case of one transverse dimension. Linear damping is introduced in much the same way as is done when analyzing the classical constant-coefficient Helmholtz equation using the limiting absorption principle. Numerically, we have found that it provides a very efficient tool for controlling the solutions of both the NLH and NLS. In particular, we have been able to identify initial conditions for which the NLS solution does become singular, while the NLH solution still remains regular everywhere. We believe that our finding of a larger domain of existence for the NLH than for the NLS is accounted for by precisely those mechanisms that have been neglected when deriving the NLS from the NLH, i.e., nonparaxiality and backscattering.

Key words. Kerr medium, nonlinear wave propagation, self-focusing, singularity formation, linear damping, limiting absorption, two-way ABCs

AMS subject classifications. 65N06, 65Z05, 78A10, 78A40, 78A45, 78A60, 78M20

DOI. 10.1137/S0036139902411855

1. Introduction.

The focusing critical nonlinear Schrödinger equation (NLS)

$$i\psi_z(z,\boldsymbol{x}) + \Delta_\perp \psi + |\psi|^{4/d}\psi = 0, \qquad \psi(0,\boldsymbol{x}) = \psi_0(\boldsymbol{x}), \tag{1.1}$$

where $\boldsymbol{x} \in \mathbb{R}^d$ and $\Delta_\perp = \partial_{x_1 x_1} + \cdots + \partial_{x_d x_d}$, arises in a variety of physical contexts. Of foremost interest is the case $d = 2$, which corresponds to the propagation of intense laser beams in Kerr media. In this case, z is the axial coordinate in the direction of propagation, $\boldsymbol{x} = (x, y)$ are the spatial coordinates in the transverse plane, $\Delta_\perp = \partial_{xx} + \partial_{yy}$ is the diffraction term (transverse Laplacian), and $|\psi|^2\psi$ describes the nonlinear polarization of the Kerr medium. It is well known that solutions to the critical NLS (1.1) can self-focus and eventually collapse, i.e., become singular, at a finite propagation distance, provided that their *initial power* $N(0) = \int |\psi_0|^2 \, d\boldsymbol{x}$ exceeds a threshold power $N_{\rm c}$, whose value depends only on the dimension d (see [7,28]). Since,

*Received by the editors July 22, 2002; accepted for publication (in revised form) January 23, 2003; published electronically July 26, 2003. This research was supported by grant 2000311 from the United States–Israel Binational Science Foundation (BSF), Jerusalem, Israel, and also by the National Aeronautics and Space Administration under NASA contract NAS1–97046 while the third author was in residence at ICASE, NASA Langley Research Center, Hampton, VA.

http://www.siam.org/journals/siap/63-5/41185.html

†School of Mathematical Sciences, Tel Aviv University, Ramat Aviv, Tel Aviv 69978, Israel (fibich@math.tau.ac.il, www.math.tau.ac.il/∼fibich; bazooka@math.tau.ac.il, www.math.tau.ac.il/∼bazooka).

‡Department of Mathematics, North Carolina State University, Box 8205, Raleigh, NC 27695 (tsynkov@math.ncsu.edu, www.math.ncsu.edu/∼stsynkov).

however, physical quantities do not become infinite, and since in experiments laser beams continue to propagate beyond the NLS blowup point, the question arises as to what specific physical mechanism(s), among those that have been neglected when deriving the NLS from the Maxwell's equations, actually arrest(s) the collapse. We recall that the final stage in the derivation of the NLS is to disregard the backscattering and apply the paraxial approximation (see section 2.2) to the critical nonlinear Helmholtz equation (NLH)

$$\Delta E(z,\boldsymbol{x}) + k_0^2(1+\epsilon|E|^{4/d})E = 0, \qquad \Delta \equiv \partial_{zz} + \Delta_\perp, \tag{1.2}$$

where k_0 is the linear wavenumber and the extent of nonlinearity is measured by the quantity $\epsilon = 4\epsilon_0 c n_2$, where n_2 is the Kerr coefficient; see, e.g., [3, 19]. Therefore, it is natural to ask whether going back from the NLS to the NLH, i.e., adding nonparaxiality and backscattering, is sufficient to guarantee existence of the solution with no singularities. In other words, for a given initial condition that leads to blowup in the critical NLS, does the NLH (always) have a solution that remains regular everywhere?

The foregoing question has been open for many years. In his celebrated 1965 paper [15], which was the first paper in the literature to predict that the solutions to the critical NLS could become singular, Kelley was careful to note that the paraxial approximation, and hence the entire NLS model, breaks down near the singularity. Feit and Fleck [4] were the first to demonstrate that nonparaxiality of the beam can arrest the blowup, by showing numerically that the initial conditions that lead to singularity formation in the NLS result in focusing-defocusing oscillations in the NLH. In these simulations, however, they did not solve a true boundary-value problem for the NLH. Instead, they solved an initial-value problem for a "modified" NLH that only describes the right-propagating wave (while introducing several additional assumptions along the way). Akhmediev and collaborators [1, 2] analyzed an initial-value problem for a different "modified" NLH; their numerical simulations also suggested that nonparaxiality arrests the singularity formation. Neither numerical approach [4] nor [1, 2], however, accounted for the effect of backscattering. Fibich [5] applied asymptotic analysis to derive an ODE in z for self-focusing in the presence of small nonparaxiality. His analysis suggests that nonparaxiality indeed arrests the singularity formation, resulting instead in decaying focusing-defocusing oscillations. However, backscattering effects were neglected in this asymptotic analysis.

The aforementioned studies [1,2,4,5,15] have prompted a general belief that nonparaxiality arrests the collapse. However, no rigorous proof of global existence for the NLH has ever been provided. Moreover, all the simulations in the above studies neglected the backscattering and considered only the forward-propagating field. The first numerical solutions of the NLH as a true boundary-value problem, with backscattering effects fully included, have been obtained by Fibich and Tsynkov in [12], using a high-order discretization supplemented by a new two-way artificial boundary condition (ABC). In that study only the case of one transverse dimension was considered, in order to keep the computational costs low. The simulations in [12] were performed for the values of the input power of up to 90% of the threshold N_c, and they have captured the mild self-focusing of the corresponding solutions. In a subsequent paper [10], we have corroborated experimentally the prediction of the asymptotic analysis that the magnitude of the backscattered signal scales quadratically with the nonparaxiality parameter f (see section 2.2), and that the computed NLH solutions converge to the corresponding NLS solutions as f goes to zero.

The numerical methodology of [12] was obviously not free of limitations of its own. Foremost, we could not obtain converging solutions for initial powers equal

to or higher than the critical value N_c. In [12], we considered initial powers only up to 90% of N_c; in the current paper we computed the NLH solutions for up to $N(0) = 0.99N_c$; see section 4. In the course of these simulations we have noticed that, as $N(0)$ approaches the critical power from below, the convergence rate of the iterations slows down noticeably. This makes the simulations for higher subcritical values of $N(0)$ ($0.99N_c < N(0) < N_c$) difficult to conduct, although it is reasonable to assume that the NLH solution will converge for input powers all the way up to N_c. However, for the input power $N(0)$ exactly equal to N_c the convergence of nonlinear iterations of [12] is lost; see section 4.

The aforementioned slowdown of convergence for input powers slightly below N_c should be attributed either to deficiencies of the method itself, or to the limits that insufficient computer resources may impose on the parameters that control the quality of the discrete approximation, or to both. As concerns the iteration method of [12], it is the most straightforward approach based on simply freezing the nonlinearity; most likely, it can be improved or replaced by a more advanced technique, and we plan on looking into this issue in the future. As for the computer resource requirements, they are determined by the size of the computational domain, which should be sufficiently large so as to meet the condition of near-linear propagation in the far field (see [12]), and by the grid size, which should be sufficiently fine to resolve a given wavelength and the sharp near-blowup profile. These requirements become more stringent for higher input powers, which decay at larger distances and/or undergo stronger focusing. In other words, the higher the input power, the larger the domain and/or the finer the grid that one needs to use in order to maintain the same solution quality and/or convergence rate. In our previous simulations we have, indeed, seen examples of diverging NLH solutions with subcritical input powers which converged on a larger computational domain and/or at a finer resolution. It is still unclear, however, whether having more computer resources and/or a better nonlinear iteration scheme will allow one to solve the NLH for initial conditions that lead to collapse in the NLS, or whether the convergence breakdown at $N(0) > N_c$ is an indication of the loss of solvability of the NLH or loss of regularity of the solution.

As such, in the current paper we explore *an alternative approach* to the issue of solving the NLH in the blowup regime of the NLS, by considering *the linearly damped NLH* and the corresponding *linearly damped NLS*. The addition of linear damping is not an ad hoc procedure. Indeed, an electromagnetic wave is always partially absorbed by the medium through which it propagates, an effect neglected in the original undamped NLH and NLS, both of which model the propagation under "ideal transparency." A mathematical motivation to add linear damping comes from the so-called *limiting absorption principle* that is used for identifying the unique solutions of the linear Helmholtz equation; see, e.g., [27]. It is known that the classical constant-coefficient homogeneous Helmholtz equation

$$\Delta E + k_0^2 E = 0 \tag{1.3a}$$

has nontrivial solutions on the entire space even in the class of functions that vanish at infinity, which obviously amounts to nonuniqueness. To fix the problem, the additional Sommerfeld boundary conditions need to be introduced at infinity that basically distinguish between the incoming and outgoing waves. On the other hand, when a complex absorption coefficient is added, the new damped equation

$$\Delta E + k_0^2(1 + i\delta)E = 0 \tag{1.3b}$$

has only trivial solution. Consequently, its inhomogeneous counterpart will be uniquely solvable for any compactly supported right-hand side in rather wide classes of functions, such as tempered distributions; see [27]. Moreover, when $\delta \longrightarrow \pm 0$, the unique solution of the inhomogeneous damped equation will *converge uniformly on the entire space* to the solution of the respective undamped equation that corresponds to either the radiation of waves toward infinity (outgoing waves) or, conversely, the incidence of waves from infinity (incoming waves), where the distinction is rendered by the sign of δ. This, in particular, implies that if we decide to keep a small but finite damping in the equation, we may expect its solution to be uniformly close to the solution of the undamped equation that is driven by the same source terms and is composed of either only outgoing or only incoming waves in the far field. The latter consideration is especially important in the context of our iteration algorithm (see section 3 and [12] for detail), which basically reduces to a repeated solution of the constant-coefficient Helmholtz equation driven by a variety of compactly supported right-hand sides and subject to the radiation boundary conditions in the far field.

Solving the damped NLH numerically as a true boundary-value problem required only minor changes in the algorithm of [12] for the undamped NLH, which are described in section 3. At the same time, *the addition of damping allows us to better control the solution.* In particular, damping decreases the solution magnitude in the far field, which is a key requirement for the validity of the ABCs of [12]. As a result, *we have been able to consider initial conditions with powers well above* N_{c}.

Let us recall that, for a given initial condition that leads to the blowup in the undamped critical NLS, there is a threshold value $\delta_{\mathrm{th}}^{\mathrm{S}}$ of the damping parameter δ such that if $\delta > \delta_{\mathrm{th}}^{\mathrm{S}}$, then linear damping arrests the collapse, whereas when $\delta < \delta_{\mathrm{th}}^{\mathrm{S}}$, the solution of the NLS blows up; see [6].[1] In the numerical simulations of the damped NLH reported hereafter we found a similar threshold value $\delta_{\mathrm{th}}^{\mathrm{H}}$ such that for $\delta > \delta_{\mathrm{th}}^{\mathrm{H}}$ the solution exists and is regular everywhere, whereas when $\delta < \delta_{\mathrm{th}}^{\mathrm{H}}$ the iteration scheme diverges. As has been mentioned, in the latter case it is not clear whether the divergence indicates that there is no solution to the NLH or that our computational resources are insufficient (or the iteration scheme is suboptimal) to calculate the solution. Therefore, we can conclude that the actual (analytical) threshold value $\hat{\delta}_{\mathrm{th}}^{\mathrm{H}}$, such that regular solutions to the NLH exist for all $\delta > \hat{\delta}_{\mathrm{th}}^{\mathrm{H}}$, is *less than or equal to* the computed threshold $\delta_{\mathrm{th}}^{\mathrm{H}}$, which is determined from the simulations, i.e., that $0 \le \hat{\delta}_{\mathrm{th}}^{\mathrm{H}} \le \delta_{\mathrm{th}}^{\mathrm{H}}$.

The main result of the current study is that

$$\delta_{\mathrm{th}}^{\mathrm{H}} < \delta_{\mathrm{th}}^{\mathrm{S}}.$$

In other words, for a given initial condition that leads to the blowup in the undamped NLS, there is an entire range of values for the damping coefficient, $\delta_{\mathrm{th}}^{\mathrm{H}} < \delta < \delta_{\mathrm{th}}^{\mathrm{S}}$, for which the damped NLS solution will blow up, but the NLH solution will be regular everywhere. Therefore, we can conclude that nonparaxiality and backscattering arrest the collapse when the damping parameter is in the range $\delta_{\mathrm{th}}^{\mathrm{H}} < \delta < \delta_{\mathrm{th}}^{\mathrm{S}}$. Whether NLH solutions exist for infinitely small linear damping as well, i.e., in the limit $\delta \longrightarrow 0$, is a question that yet remains to be answered. We believe, however, that this question

[1] Self-focusing in the critical NLS is highly sensitive to the effect of small perturbations. Some perturbations can arrest the collapse even if they are initially infinitesimally small [11]. In contrast, an infinitesimally small linear damping does not arrest the collapse, and a sufficient amount of damping must be present to regularize the solution.

should be considerably easier to address, both numerically and analytically, than the question of solvability of the original undamped NLH.

2. Formulation of the problem.

2.1. The nonlinear Helmholtz equation. A typical setup for the propagation of electromagnetic waves in a Kerr medium is shown in Figure 2.1. An incoming laser beam with known characteristics impinges normally on the planar interface $z = 0$ between the linear and the nonlinear medium. The electric field $E = E(z, \boldsymbol{x})$ is governed by the NLH (1.2). For simplicity, we consider the cylindrically symmetric case,[2] where $E = E(z, r)$ and $r = \sqrt{x_1^2 + \cdots + x_d^2}$. The nonlinear medium occupies the semispace $z \geq 0$ (see Figure 2.1). Consequently, the NLH (1.2) has to be supplemented by boundary conditions at $z = 0$ and $z \longrightarrow +\infty$. We require that as $z \longrightarrow +\infty$, E have no left-traveling components and that the propagation be diffraction-dominated, with the field amplitude decaying to zero, i.e., $\lim_{z\to\infty} \max_{0\leq r<\infty} |E(z, r)| = 0$, which also means that the nonlinear wavenumber $k^2 \equiv k_0^2(1 + \epsilon|E|^{4/d})$ approaches its linear limit: $\lim_{z\to+\infty} k^2 = k_0^2$. In other words, at large z's the solution should be a linear superposition of right-traveling waves. Since the actual numerical simulation is carried out on a truncated domain $0 \leq z \leq z_{\max}$ (Figure 2.1), the desired behavior of the solution as $z \longrightarrow +\infty$ has to be captured by a far-field ABC at the artificial boundary $z = z_{\max}$. This boundary condition should guarantee a reflectionless propagation of all the waves traveling towards $z = +\infty$. Often, boundary conditions designed to ensure the transparency of the outer boundary to the outgoing waves are called *radiation boundary conditions* [24].

The situation is more complex at the interface $z = 0$, where the total field $E(0, r)$ is composed of a given incoming (right-traveling) component $E_{\text{inc}}(0, r)$ and an unknown backscattered (left-traveling) component $E_{\text{scat}}(0, r)$, i.e.,

$$E(0, r) = E_{\text{inc}}(0, r) + E_{\text{scat}}(0, r).$$

As such, the boundary condition at $z = 0$ has to guarantee the reflectionless propagation of any left-traveling wave through the interface and at the same time be able

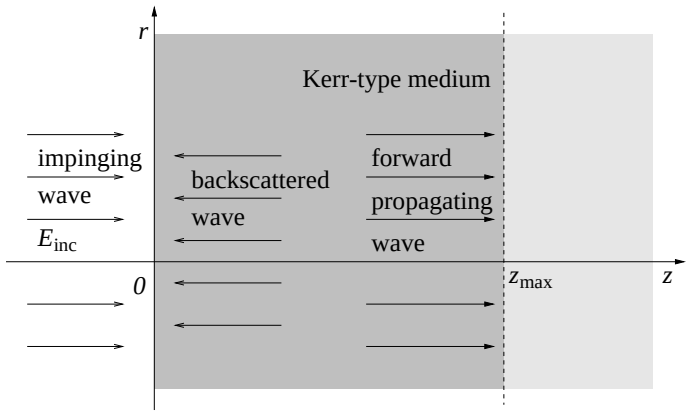


FIG. 2.1. *Schematic of propagation of waves in Kerr media.*

[2]This assumption is quite reasonable, since even when the initial conditions of the NLS are not cylindrically symmetric, near the singularity the solution becomes cylindrically symmetric [8].

to correctly prescribe the incoming signal. Implementation of such a *two-way ABC* was first carried out in [12] for the undamped NLS, and is extended to the damped case in section 3.3.

Finally, the electric field vanishes as $r \longrightarrow +\infty$. In practice, we truncate the domain at some large but finite $r_{\max}$ and require that $E(z, r_{\max}) = 0$. Similar approaches to the treatment of remote transverse artificial boundaries have been introduced and tested previously; see the discussion in the end of section 3.2. In order to avoid possible problems with reflections from the boundary $r = r_{\max}$, the computations of section 4 were conducted with $r_{\max}$ being 40 times larger than the radius of the impinging beam.

2.2. Paraxial approximation and the NLS. We first introduce the dimensionless quantities $\tilde{r}$, $\tilde{z}$, and ψ as

$$\tilde{r} = \frac{r}{r_0}, \quad \tilde{z} = \frac{z}{2L_{DF}}, \quad E = e^{ik_0 z}(\epsilon r_0^2 k_0^2)^{-d/4}\psi(z, r), \tag{2.1}$$

where r_0 is the transverse width of the input beam and $L_{DF} = k_0 r_0^2$ is the *diffraction length*. Then, by substituting the quantities (2.1) into the NLH (1.2) and dropping the tildes, we obtain

$$i\psi_z + \Delta_\perp \psi + |\psi|^{4/d}\psi = -4f^2\psi_{zz}, \tag{2.2}$$

where $f = 1/r_0 k_0 = \lambda/2\pi r_0$ is the *nonparaxiality parameter*.

The standard derivation of the NLS is motivated by the observation that $f \ll 1$, since typically $\lambda \ll r_0$. This suggests that one can neglect the ψ_{zz} term, i.e., apply the *paraxial approximation*, and obtain the NLS

$$i\psi_z(z, r) + \Delta_\perp \psi + |\psi|^{4/d}\psi = 0, \tag{2.3}$$

which is the same as the previously introduced (1.1), except that in (2.3) we use r instead of $\boldsymbol{x}$ for simplicity. The NLS (2.3) is supplemented by the initial condition at $z = 0$:

$$\psi(0, r) = (\epsilon r_0^2 k_0^2)^{d/4} E_{\text{inc}}(0, r).$$

Subsequently, it needs to be integrated by a "time"-marching algorithm, where the direction of propagation z plays the role of time. We reemphasize that *backscattering effects are not taken into account by the NLS* (2.3). Indeed, once (2.3) is solved, the overall solution, according to (2.1), is the slowly varying amplitude ψ times the forward-propagating oscillatory component $e^{ik_0 z}$.

2.3. Linear damping. When damping, i.e., linear absorption, is included, the NLH (1.2) becomes

$$\Delta E(z, \boldsymbol{x}) + k_0^2(1 + i\delta + \epsilon|E|^{4/d})E = 0, \tag{2.4}$$

where k_0 is the (real part of the) wavenumber,

$$\delta = \frac{\text{Im}(n_0^2)}{\text{Re}(n_0^2)},$$

and n_0 is the linear index of refraction of the medium. The corresponding NLS (2.3) becomes (see (2.1))

$$i\psi_z + \Delta_\perp \psi + |\psi|^{4/d}\psi + ir_0^2 k_0^2 \delta\psi = 0. \tag{2.5}$$

By definition, optical transparency of the medium means that the damping is small. For example, for water in the visible regime [14],

$$\frac{\mathrm{Im}(n_0^2)}{\mathrm{Re}(n_0^2)} \sim 10^{-7}.$$

Having small physical values of damping also agrees well with the mathematical reasoning behind the limiting absorption principle. As indicated in section 1 (see, e.g., [27] for detail), for a classical constant-coefficient Helmholtz operator of (1.3a), the introduction of a small complex absorption coefficient of the appropriate sign [as in (1.3b)] implies that there will be a unique solution for any compactly supported excitation, and that this solution will be uniformly close in the entire space $\mathbb{R}^{d+1}$ to the solution of the corresponding undamped linear Helmholtz equation driven by the same sources and subject to the radiation boundary conditions in the far field. In the following section 3, we show that for the formulation analyzed in this paper the proper sign of δ is positive.

As we have noted before, the physical case that corresponds to the propagation of laser beams in bulk Kerr media is $d = 2$. However, in order to reduce the complexity of the computations we rather consider a simpler case $d = 1$, as was previously done in [12]. Thus, the damped NLH for $E = E(z, r)$ and the damped NLS for $\psi = \psi(z, r)$ that are solved numerically in this study are

$$E_{zz}(z,r) + E_{rr} + k_0^2(1 + i\delta + \epsilon|E|^4)E = 0 \tag{2.6}$$

and

$$i\psi_z(z,r) + \psi_{rr} + ir_0^2k_0^2\delta\psi + |\psi|^4\psi = 0, \tag{2.7}$$

respectively.

3. Numerical methods. The damped NLH (2.6) is solved using fourth-order finite differences. The methodology of solution is outlined below in section 3.1; it is similar to the one that we have introduced in our previous work [12] for solving the undamped NLH. The choice of a higher-order method is motivated primarily by the necessity to resolve a small-scale phenomenon of backscattering at the background of the forward-propagating waves. Indeed, it is generally known that higher-order methods provide for a better resolution of waves. The damped NLS (2.7) is also solved using a fourth-order approximation in all coordinate directions. Since the Schödinger equation models the evolution of the slowly varying envelope, one can expect the magnitudes of the corresponding higher-order derivatives involved in the truncation error terms to be smaller for the NLS than for the NLH. This implies that on a grid of comparable size the accuracy of the numerical approximation for the NLS should be better than of those for the NLH. Moreover, in our simulations we typically employ finer grids for the NLS than those that we use for the NLH, thus obtaining an accurate numerical solution for the simpler model. It then serves as a natural reference point for the more "elaborate" NLH solution to be compared against.

3.1. Discretization of the NLH and solution methodology. We use a conventional fourth-order central-difference discretization for the Laplacian $\Delta = \partial_{zz} + \partial_{rr}$ of (2.6); thus the stencil is five nodes wide in each coordinate direction. As the equation is nonlinear, we implement a nested iteration scheme. On the outer loop, we freeze the nonlinearity, i.e., consider the coefficient $k^2 \equiv k_0^2(1 + i\delta + \epsilon|E|^4)$ as a given

function of the coordinates z and r, which is actually obtained by taking the quantity $|E|^4$ from the previous iteration; see (2.6). This way we arrive at a linear equation with variable coefficients. The latter is also solved by iterations on the inner loop of the nested scheme. Here, we leave the entire varying part of the equation, which is proportional to ϵ, on the lower level, and on the upper level need to invert only *the constant-coefficient linear damped Helmholtz operator* $\Delta + k_0^2(1+i\delta)\boldsymbol{I}$ (cf. (1.3b)).

Formally, our iteration scheme resembles the fixed-point approach; however, no rigorous convergence theory is available yet, and the convergence is assessed experimentally. The advantages of using these nested iterations are twofold. First, the method eventually reduces to the repeated solution of one and the same linear constant-coefficient equation driven by different source terms, which can be done efficiently at the discrete level. Second, the radiation boundary conditions at $z = z_{\max}$ and the two-way ABCs at $z = 0$ (see Figure 2.1) are most convenient to set on the upper time level of the iteration scheme already for the linear constant-coefficient operator.

To solve the linear constant-coefficient damped discrete Helmholtz equation

$$\Delta^{(h)}E + k_0^2(1+i\delta)E = g, \tag{3.1}$$

where g is the right-hand side generated on the previous iteration, we first separate the variables by implementing the discrete Fourier transform in the transverse direction r; the boundary conditions are symmetry at $r = 0$ and zero Dirichlet at $r = r_{\max}$ (see section 2.1). This yields a collection of fourth-order one-dimensional finite-difference equations (grid index n corresponds to the continuous variable z):

$$\frac{-\hat{E}_{n-2} + 16\hat{E}_{n-1} - 30\hat{E}_n + 16\hat{E}_{n+1} - \hat{E}_{n+2}}{12h_z^2} + (k_0^2(1+i\delta) - \lambda_m)\hat{E}_n = \hat{g}_n \tag{3.2}$$

parameterized by the dual Fourier variable λ_m; the latter is defined by formula (29) of [12]. Each equation (3.2) needs to be solved independently.[3] The two-way and radiation ABCs at $z = 0$ and $z = z_{\max}$, respectively, for the discrete equation (3.1) are set in the Fourier space, i.e., individually for each one-dimensional equation (3.2). This is done by first identifying the linearly independent eigenmodes for the homogeneous version of this equation. It is important to note that, even though the original differential equation is of the second order, we are using its fourth-order approximation, and thus each homogeneous discrete one-dimensional equation of type (3.2) has *four linearly independent solutions.* These solutions are q_1^n, q_1^{-n}, q_2^n, and q_2^{-n} (see [12]), where q_1, $1/q_1$, q_2, and $1/q_2$ are roots of the characteristic algebraic equation

$$-1 + 16q + (12h_z^2(k_0^2(1+i\delta) - \lambda_m) - 30)q^2 + 16q^3 - q^4 = 0. \tag{3.3}$$

3.2. Roots of the characteristic equation. It is indeed easy to see that (3.3) has two pairs of mutually inverse roots. We first notice that this equation originates from a central-difference, i.e., symmetric, discretization (3.2). Given that, if q is a root, then q^{-1} is obviously a root as well, which can be verified by direct substitution. Then, to actually find the roots we rewrite the polynomial on the left-hand side of (3.3) as

$$\begin{aligned}&(q-q_1)(q-q_1^{-1})(q-q_2)(q-q_2^{-1})\\ \equiv&-1 + (d_1+d_2)q - (2+d_1d_2)q^2 + (d_1+d_2)q^3 - q^4,\end{aligned}$$

[3]Note that the discrete equations (3.1) and (3.2) are very similar to the corresponding discrete equations studied in [12], except that previously we had no damping.

where

$$d_1 = q_1 + q_1^{-1}, \qquad d_2 = q_2 + q_2^{-1},$$

and match the coefficients. In so doing, we obtain

$$d_1 + d_2 = 16, \qquad -2 - d_1 d_2 = 12h_z^2(k_0^2(1+i\delta) - \lambda_m) - 30, \tag{3.4}$$

so that each pair of roots, q_1, q_1^{-1} and q_2, q_2^{-1}, can be found by solving the corresponding quadratic equation,

$$q^2 - d_1 q + 1 = 0 \tag{3.5a}$$

or

$$q^2 - d_2 q + 1 = 0, \tag{3.5b}$$

while the coefficients d_1 and d_2 are, in turn, determined by solving quadratic equations (3.4).

At this stage, the key difference between the current analysis for the damped equation and the previous analysis for the undamped equation of [12] needs to be emphasized. As shown in [12], when $\delta = 0$, the first pair of solutions of the homogeneous equation (3.2), q_1^n and q_1^{-n}, approximates the genuine "longitudinal," i.e., z-aligned, modes of the *undamped* homogeneous differential equation (1.3a):

$$\hat{E}_1 = e^{ik_c z} \quad \text{and} \quad \hat{E}_2 = e^{-ik_c z}, \tag{3.6}$$

respectively. The functions $\hat{E}_1 = \hat{E}_1(z)$ and $\hat{E}_2 = \hat{E}_2(z)$ are two linearly independent solutions of the ODE

$$\hat{E}_{zz} + (k_0^2 - \lambda)\hat{E} = 0 \tag{3.7}$$

obtained by Fourier transforming (1.3a) with respect to r; λ is the dual variable. In formulae (3.6), we have denoted $k_c = \sqrt{k_0^2 - \lambda}$, and a particular branch of the square root that we always take is $\sqrt{\rho e^{i\theta}} = \rho^{1/2} e^{i\theta/2}$. The two continuous modes (3.6) may be either traveling or evanescent waves, depending on whether the real quantity $k_c^2 = (k_0^2 - \lambda)$ is positive or negative, or in other words, whether the dual Fourier variable λ is less or greater than k_0^2. To demonstrate the aforementioned approximation property for the undamped ($\delta = 0$) discretization (3.2), we redefine $k_c = \sqrt{k_0^2 - \lambda_m}$, introduce $\alpha = h_z^2 k_c^2$, and show in [12] that if $\alpha > 0$, then q_1 and q_1^{-1} are complex conjugate roots of the characteristic equation (3.3). Both these roots have unit magnitude $|q_1| = |q_1^{-1}| = 1$, which indicates that q_1^n and q_1^{-n} are pure discrete traveling waves. Moreover, if $\alpha \ll 1$, then (see [12])

$$q_1 = e^{ik_c h_z} + \mathcal{O}((k_c h_z)^5), \qquad q_1^{-1} = e^{-ik_c h_z} + \mathcal{O}((k_c h_z)^5). \tag{3.8}$$

Equalities (3.8) imply that in the undamped case $\delta = 0$, q_1^n is a discrete counterpart of the right-traveling wave $\hat{E}_1$, and q_1^{-n} is a discrete counterpart of the left-traveling wave $\hat{E}_2$; the approximation is obviously fourth-order accurate because on the grid $z_n = h_z n$. If $\alpha < 0$ and still $\delta = 0$, then we again show in [12] that $|q_1| < 1$ and $|q_1^{-1}| > 1$, which indicates that q_1^n is a right-evanescent wave and q_1^{-n} is a left-evanescent wave.

The situation changes drastically with the introduction of damping. In contradistinction to the undamped case, when $\delta \neq 0$ the homogeneous differential equation no longer has pure propagating, i.e., constant-amplitude, longitudinal modes. Indeed, by Fourier transforming equation (1.3b) in the r direction, we arrive at the family of ODEs

$$\hat{E}_{zz} + (k_0^2(1+i\delta) - \lambda)\hat{E} = 0 \tag{3.9}$$

parameterized by the dual variable λ. Each of the equations (3.9) has two linearly independent solutions:

$$\begin{aligned} \hat{E}_1 &= e^{iz\sqrt{k_c^2+ik_0^2\delta}} = e^{ik_cz\sqrt{1+i\frac{k_0^2}{k_c^2}\delta}}, \\ \hat{E}_2 &= e^{-iz\sqrt{k_c^2+ik_0^2\delta}} = e^{-ik_cz\sqrt{1+i\frac{k_0^2}{k_c^2}\delta}}. \end{aligned} \tag{3.10}$$

Clearly, the second equality in each formula (3.10) is valid only if $k_c \neq 0$. Formulae (3.10) show that, as long as $\delta \neq 0$, there will always be a nontrivial real part in each exponent. Consequently, the amplitudes of the waves (3.10) *will always decrease or increase exponentially* for $z \longrightarrow \pm\infty$. In particular, if we analyze the traveling waves regime of the undamped equation, i.e., the case of small λ: $k_0^2 - \lambda > 0$, and additionally assume that $|\delta| \ll 1$, then formulae (3.10) yield (cf. formulae (3.6))

$$\begin{aligned} \hat{E}_1^{(\text{damped})} &\approx e^{ik_cz\left(1+i\frac{1}{2}\frac{k_0^2}{k_c^2}\delta\right)} = e^{ik_cz-\frac{1}{2}\frac{k_0^2}{k_c}\delta z} = \hat{E}_1^{(\text{undamped})} \cdot e^{-\frac{1}{2}\frac{k_0^2}{k_c}\delta z}, \\ \hat{E}_2^{(\text{damped})} &\approx e^{-ik_cz\left(1+i\frac{1}{2}\frac{k_0^2}{k_c^2}\delta\right)} = e^{-ik_cz+\frac{1}{2}\frac{k_0^2}{k_c}\delta z} = \hat{E}_2^{(\text{undamped})} \cdot e^{\frac{1}{2}\frac{k_0^2}{k_c}\delta z}. \end{aligned} \tag{3.11}$$

Since we identify $\hat{E}_1^{(\text{undamped})} = e^{ik_cz}$ of (3.6) as the right-traveling wave, and $\hat{E}_2^{(\text{undamped})} = e^{-ik_cz}$ of (3.6) as the left traveling wave, we can conclude that to have the propagation toward infinity (i.e., the radiation of waves) accompanied by *the decay of the amplitude* (as opposed to growth with no bound), we have to take *positive values of the damping factor* $\delta > 0$ (cf. section 1). In this case, the amplitude of $\hat{E}_1^{(\text{damped})}$ will decay exponentially for $z \longrightarrow +\infty$ (propagation to the right), and the amplitude of $\hat{E}_2^{(\text{damped})}$ will decay exponentially for $z \longrightarrow -\infty$ (propagation to the left). As one can easily see from (3.11), the rate of decay is controlled by the value of δ.

In connection to the aforementioned exponential behavior of the longitudinal modes, a more general fact is also worth mentioning. The full Fourier symbol of the undamped operator of (1.3a) obviously has real roots on the dual plane; these roots occupy the entire circle of radius k_0 centered at the origin. In contradistinction to that, the symbol of the damped operator of (1.3b) *does not have real roots* on the dual plane. As shown in [20], the damped operator will therefore have an exponentially decaying fundamental solution. In practical terms it means that the outgoing waves governed by the damped Helmholtz equation will decay exponentially toward infinity in all directions. For comparison we recall that the fundamental solution of the undamped operator is given by a zero-order Hankel function, which only decays at infinity as the inverse square root of the distance from the origin.

To establish the properties of the propagating modes for the discretization (3.2) in the presence of damping, and to demonstrate similarities to the continuous damped case, we first introduce and prove the following result.

PROPOSITION 3.1. *The characteristic equation* (3.3) *for* $\delta \neq 0$ *does not have roots with unit magnitude.*

Proof. Let us assume the opposite: There exists a unit magnitude root $q = e^{i\theta}$ to the algebraic characteristic equation (3.3). Then,

$$\begin{aligned} -1 + 16e^{i\theta} + (12h_z^2(k_0^2(1+i\delta) - \lambda_m) - 30)e^{2i\theta} + 16e^{3i\theta} - e^{4i\theta} \\ = \left[-e^{-2i\theta} + 16e^{-i\theta} + (12h_z^2(k_0^2(1+i\delta) - \lambda_m) - 30) + 16e^{i\theta} - e^{2i\theta}\right] \cdot e^{2i\theta} \\ = \left[-2\cos(2\theta) + 32\cos\theta + (12h_z^2(k_0^2(1+i\delta) - \lambda_m) - 30)\right] \cdot e^{2i\theta} = 0. \end{aligned}$$

As $e^{2i\theta} \neq 0$, the expression in rectangular brackets has to be equal to zero. Since the only imaginary contribution to this expression is proportional to δ, we conclude that it is only possible when $\delta = 0$. ∎

Proposition 3.1 implies that, similarly to the continuous case, there will be no constant-amplitude solutions to the homogeneous counterpart of the discrete equation (3.2). Each of the four corresponding modes, q_1^n, q_1^{-n}, q_2^n, and q_2^{-n}, will exponentially decrease in one direction and exponentially increase in the opposite direction. In particular, if we assume as before that $\alpha \ll 1$ in the undamped traveling waves regime,[4] and if we in addition let $\delta \ll 1$, then, solving (3.4) for d_1 first, then (3.5a) for q_1 and q_1^{-1}, and finally using the Taylor expansion, we obtain (cf. (3.8))

$$\begin{aligned} q_1 &= e^{ik_c h_z - \frac{1}{2}\frac{k_0^2}{k_c}\delta h_z} + \mathcal{O}\left(\left[k_c h_z\left(1 + i\frac{1}{2}\frac{k_0^2}{k_c^2}\delta\right)\right]^5\right), \\ q_1^{-1} &= e^{-ik_c h_z + \frac{1}{2}\frac{k_0^2}{k_c}\delta h_z} + \mathcal{O}\left(\left[k_c h_z\left(1 + i\frac{1}{2}\frac{k_0^2}{k_c^2}\delta\right)\right]^5\right). \end{aligned} \tag{3.12}$$

Equalities (3.12) mean that the damped discrete traveling waves q_1^n and q_1^{-n} approximate the damped continuous waves (3.11) with the fourth order of accuracy. This result is obviously similar to the one obtained in the undamped case; see formulae (3.8).

Thus far, our discussion has focused on the first pair of roots q_1 and q_1^{-1} of the characteristic equation (3.3), because these roots correspond to the genuine modes of the original differential equation. The second pair of roots q_2 and q_2^{-1} is obtained by solving (3.4) for d_2 and subsequently solving (3.5b). The corresponding pair of solutions q_2^n and q_2^{-n} is, of course, a pure numerical artifact. In [12] we have shown that for $\delta = 0$ the roots q_1 and q_1^{-1} cannot have unit magnitude: $|q_2| < 1$ and $|q_2^{-1}| > 1$, which means that the waves q_2^n and q_2^{-n} are always evanescent. In the damped case, Proposition 3.1 implies that these waves will remain evanescent as well. The presence of the second pair of waves, however, implies that the discrete equation requires two more boundary conditions compared to the original differential equation.

In section 1, we have outlined a general two-fold motivation behind the introduction of damping into the Helmholtz equation. One part was coming from physics because absorption by the medium always accompanies the propagation of electromagnetic waves in real-life settings. Moreover, from the standpoint of mathematics the introduction of damping helps select a unique solution using the limiting absorption principle. Besides these two key reasons, the presence of damping in the equation also positively affects the properties of the numerical algorithm.

[4] This would also imply $\frac{k_0^2}{k_c}h_z \ll 1$ because λ_m is small and $k_c \sim k_0$.

First, *having no roots of unit magnitude presents a significant advantage from the viewpoint of numerical stability.* In this case, every discrete system (3.2), supplemented by the boundary conditions that are discussed below in section 3.3, will be well posed in the classical sense of [13, 21]. In contrast to that, in the original undamped case existence of the roots with unit magnitude may, generally speaking, cause a weak polynomial growth of the error when the grid size is refined, although no major exponential instability will be possible.

We recall that the original formulation of the problem requires that $E(z, r)$ vanish as $|r| \longrightarrow \infty$. Instead, when solving the problem numerically, we set $E(z, r) = 0$ at a large but still finite distance $r = r_{\max}$. Of course, we expect that on some fixed bounded region of interest located next to the axis of the propagating beam our solution will converge to the original infinite-domain solution with the increase of $r_{\max}$. A general methodology for solving infinite-domain problems based on a similar idea was first introduced and studied in [22, 23, 25, 26] in the context of fluid flow. It was shown, in particular, that one may obtain the convergence rate inversely proportional to the square of the domain size (i.e., $\sim 1/r_{\max}^2$ using our particular notations). Besides, for a specific example that involves the Laplace equation that transforms into a Yukawa equation by introducing small "dissipation," Mishkov and Ryaben'kii have shown in [18] that one may expect a much faster convergence of the damped solution to the undamped one on a fixed-size domain rather than on the original unbounded domain. Even though the formulation of the problem in [18] is not quite the same as the one analyzed here, there are still similarities that allow us to consider the results of [18] as another argument for using the damped equation.

3.3. Boundary conditions. Apart from the foregoing key difference in the properties of the roots of (3.3) in the undamped and damped case (see section 3.2), the algorithm for solving the damped NLH remains basically the same as the undamped algorithm of [12]. Each equation (3.2) needs to be supplemented by the radiation boundary conditions at $z = z_{\max}$ and two-way ABCs at $z = 0$.

The *radiation boundary conditions* are constructed by requiring that on the right boundary $z = z_{\max}$ the solution of (3.2) be composed of only the waves that propagate/decay to the right, i.e., $\hat{E}_n = c_1 q_1^n + c_2 q_2^n$. The selection is rendered by the so-called one-way discrete Helmholtz equation [12], which is a linear homogeneous relation that defines the span of all the appropriate modes. Specifically, let us consider (3.2) on the grid $n = 0, 1, \ldots, N-1, N$, and assume that the right-hand side $\hat{g}_n$ is small and can therefore be neglected near the right boundary $n = N$, i.e., that *the propagation is almost linear in the far field.* Then, we require that the vector $[\hat{E}_{N-3}, \hat{E}_{N-2}, \hat{E}_{N-1}, \hat{E}_N]^T$ be a linear combination of the two vectors $[q_1^{N-3}, q_1^{N-2}, q_1^{N-1}, q_1^N]^T$ and $[q_2^{N-3}, q_2^{N-2}, q_2^{N-1}, q_2^N]^T$, which obviously translates into

$$\text{Rank} \begin{bmatrix} \hat{E}_{N-3} & \hat{E}_{N-2} & \hat{E}_{N-1} & \hat{E}_N \\ 1 & q_1 & q_1^2 & q_1^3 \\ 1 & q_2 & q_2^2 & q_2^3 \end{bmatrix} = 2. \tag{3.13}$$

Relation (3.13) is, in turn, equivalent to the two scalar equalities

$$q_1 q_2 \hat{E}_{N-3} - (q_1 + q_2)\hat{E}_{N-2} + \hat{E}_{N-1} = 0, \tag{3.14a}$$

$$q_1 q_2 \hat{E}_{N-2} - (q_1 + q_2)\hat{E}_{N-1} + \hat{E}_N = 0, \tag{3.14b}$$

which constitute *the one-way-to-the-right discrete Helmholtz equation.* Relations (3.14a) and (3.14b) supplement the scheme (3.2) at $n = N - 1$ and $n = N$, respec-

tively, i.e., at the two near-edge nodes of the grid where the regular five-point-wide stencil of (3.2) cannot be applied.

The *two-way ABC* at $z = 0$ also has to possess the capability of radiation boundary conditions, i.e., it has guarantee the transparency of the interface for all the waves that propagate/decay to the left. In other words, we require that at the left boundary the outgoing, i.e., scattered, waves be given by $\hat{E}_n^{(\mathrm{scat})} = c_1 q_1^{-n} + c_2 q_2^{-n}$. Assuming for a second the homogeneity $\hat{g}_n = 0$ near $n = 0$, we could obtain, similarly to (3.13),

$$\mathrm{Rank} \begin{bmatrix} \hat{E}_0^{(\mathrm{scat})} & \hat{E}_1^{(\mathrm{scat})} & \hat{E}_2^{(\mathrm{scat})} & \hat{E}_3^{(\mathrm{scat})} \\ 1 & q_1^{-1} & q_1^{-2} & q_1^{-3} \\ 1 & q_2^{-1} & q_2^{-2} & q_2^{-3} \end{bmatrix} = 2. \tag{3.15}$$

Relation (3.15), again, is equivalent to *the one-way-to-the-left discrete Helmholtz equation:*

$$\hat{E}_0^{(\mathrm{scat})} - (q_1 + q_2)\hat{E}_1^{(\mathrm{scat})} + q_1 q_2 \hat{E}_2^{(\mathrm{scat})} = 0, \tag{3.16a}$$

$$\hat{E}_1^{(\mathrm{scat})} - (q_1 + q_2)\hat{E}_2^{(\mathrm{scat})} + q_1 q_2 \hat{E}_3^{(\mathrm{scat})} = 0. \tag{3.16b}$$

Equations (3.16a), (3.16b), however, cannot be immediately used as the ABC at $z = 0$ because the foregoing assumption of homogeneity near the interface is, generally speaking, not correct, and moreover, (3.16a), (3.16b) do not account for the incoming wave at $z = 0$ (see section 2.1), i.e., do not have the important two-way capability. The analysis of [12] shows that to accurately address both issues, i.e., the inhomogeneity that comes from the previous iteration and the presence of the incoming wave, it is sufficient to introduce particular modifications to the right-hand side g_n at only two nodes: $n = 0$ and $n = 1$. The corresponding modification due to the incoming signal is obtained by simply substituting the right-traveling incoming wave $\hat{E}_0^{(\mathrm{inc})} q_1^n$ into the one-way-to-the-left Helmholtz equation (3.16a), (3.16b). Altogether, the two-way ABCs at $z = 0$ are given by (cf. formulae (3.16a), (3.16b))

$$\hat{E}_0 - (q_1 + q_2)\hat{E}_1 + q_1 q_2 \hat{E}_2 = \hat{g}_0', \tag{3.17a}$$

$$\hat{E}_1 - (q_1 + q_2)\hat{E}_2 + q_1 q_2 \hat{E}_3 = \hat{g}_1', \tag{3.17b}$$

where a prime denotes the aforementioned modification of the right-hand side; see [12]. Again, relations (3.17a) and (3.17b) supplement the scheme (3.2) at the near-edge nodes $n = 0$ and $n = 1$, respectively, where the regular five-point stencil cannot be applied. Straightforward considerations based on the linear superposition principle and uniqueness (see [12]) guarantee that inhomogeneous relations (3.17a), (3.17b) correctly specify the incoming signal at $z = 0$ and still ensure the reflectionless propagation of all the outgoing waves through $z = 0$ toward $z = -\infty$.

3.4. Computational complexity. The computational complexity of one solution of (3.1) is $\mathcal{O}(N_z N_r \ln N_r)$ operations, where N_z and N_r are the corresponding grid dimensions. Indeed, the cost of solving each of the N_r one-dimensional systems (3.2) is linear with respect to N_z, because each of these systems needs to be solved repeatedly for multiple right-hand sides. As such, the sparse LU decomposition can be performed only once ahead of time, and the cost of backward substitution is linear. Therefore, the overall complexity is dominated by the cost of N_z direct and inverse FFTs of length N_r, which is $\mathcal{O}(N_z N_r \ln N_r)$.

For the specific discretization parameters provided in section 4, the numbers of iterations could vary significantly. The borderline cases, i.e., those with the minimal damping necessary to allow the algorithm to converge, could take as many as a thousand iterations to reduce the initial relative difference between two successive iterations by three to four orders of magnitude. In contrast, the cases that involved damping substantially larger than the required minimum could converge to machine zero (fifteen orders of magnitude reduction) in as little as two hundred iterations.

4. Results. In this section we present simulation results for the Gaussian initial conditions $E^0_{\rm inc} = \exp(-r^2)$ and $\psi_0 = (\epsilon r_0^2 k_0^2)^{1/4} \exp(-r^2/r_0^2)$ for the NLH and NLS, respectively. Denoting, as before, the input power of the incoming wave by $N(0)$, we define *the fractional input power* as

$$p = \frac{N(0)}{N_{\rm c}}, \tag{4.1}$$

i.e., $p = 1$ when the input power is equal to the NLS critical power $N_{\rm c}$. For the Gaussian initial conditions used in our simulations, $p = k_0\sqrt{2\epsilon/3\pi}$ (see [12]). In all simulations we set $k_0 = 8$ and $r_0 = 1$.

In Table 4.1 we show the calculated threshold values $\delta^{\rm H}_{\rm th}$ and $\delta^{\rm S}_{\rm th}$. The quantity $\delta^{\rm H}_{\rm th}$ in Table 4.1 represents the smallest nonnegative value of δ for which we obtain a global solution of the NLH. By this we mean that the nonlinear iterations converge in the sense that the value of $\max_{z,r}(E^{(n+1)} - E^{(n)})/\max_{z,r} E^{(n+1)}$ drops by at least a factor of 10^{-6} in the course of iterations on the computational domain $0 \le z \le 40$ and $0 \le r \le 40$, with grid sizes $h_z = \lambda/20$ and $h_r = \lambda/8$, where $\lambda = 2\pi/k_0$. The particular choice of the domain size and grid resolution is "inherited" from our previous numerical experiments; see [10,12]. The values of $\delta^{\rm H}_{\rm th}$ in Table 4.1 are obtained with at least two significant digits by repeatedly running the code for a given ϵ and varying δ, which allows one to "close in" on the threshold. However, as discussed in section 1, with a larger computational domain and/or a finer grid it may be possible to obtain regular solutions for smaller values of δ, hence, to obtain a lower value of the threshold $\delta^{\rm H}_{\rm th}$. For example, using the same computational domain and twice as fine a grid, $h_z = \lambda/40$ and $h_r = \lambda/16$, we could obtain $\delta^{\rm H}_{\rm th} = 0.0133$ instead

TABLE 4.1
Threshold values of linear damping δ.

Case No.	ϵ	$p = N(0)/N_{\rm c}$	$\delta^{\rm H}_{\rm th}$	$\delta^{\rm S}_{\rm th}$
1	0.06	90%	0	0
2	0.07	97.5%	0	0
3	0.072165819	99%	0	0
4	$3\pi/128$	100%	$9.6 \cdot 10^{-5}$	0
5	0.075	100.9%	0.00023	0
6	0.08	104%	0.00071	0.00025
7	0.1	116%	0.0027	0.0025
8	0.125	130%	0.0049	0.0062
9	0.15	142%	0.0071	0.010
10	0.2	164%	0.0145	0.019
11	0.3	202%	0.030	0.035
12	0.4	233%	0.044	0.050
13	0.5	261%	0.058	0.065

of $\delta_{\rm th}^{\rm H} = 0.0145$ for the data in row 10 of Table 4.1 ($\epsilon = 0.2$). Likewise, using the original grid resolution $h_z = \lambda/20$ and $h_r = \lambda/8$ and a computational domain that was twice as large, $z_{\max} = 80$ and $r_{\max} = 80$, we could obtain $\delta_{\rm th}^{\rm H} = 0.0022$ instead of $\delta_{\rm th}^{\rm H} = 0.0027$ for the data in row 7 of Table 4.1 ($\epsilon = 0.1$). In other words, the values of $\delta_{\rm th}^{\rm H}$ from Table 4.1 should be considered *upper bounds* for the actual thresholds. However, the quantitative limits of pursuing this venue are still unexplored, i.e., it is not known how far down in $\delta_{\rm th}^{\rm H}$ one can go by increasing the domain size and/or grid resolution. Our ability to answer this question is obviously limited by computer resources, and as of yet *the question remains open.* In particular, it is unclear whether we can achieve $\delta_{\rm th}^{\rm H} = 0$ by choosing a sufficiently large domain and/or fine grid.

Similarly, the quantity $\delta_{\rm th}^{\rm S}$ in Table 4.1 represents the smallest nonnegative value of damping δ for which the NLS solution does not blow up. In our NLS simulations we use standard fourth-order finite-difference schemes for the spatial derivatives and explicit fourth-order Runge–Kutta for marching in z. As has recently been shown in [9], in finite-difference simulations of NLS solutions that are known analytically to become singular, the computed solution still remains bounded. Therefore, there is always an element of arbitrariness in selecting a numerical criterion for blowup in NLS simulations. In our NLH simulations the largest relative increase in amplitude due to self-focusing has never exceeded a factor of two. In order to make the blowup criteria in NLH and NLS simulations as close to one another as possible, we define the computed NLS solution as becoming singular once its amplitude increases by a factor of two. We checked that altering this NLS blowup criterion leads to only minor changes in the results for $\delta_{\rm th}^{\rm S}$. For example, using the blowup criterion of relative focusing by a factor of 4, rather than 2, for $\epsilon = 0.08$ (row 6 of Table 4.1) gives $\delta_{\rm th}^{\rm S} = 0.00021$ instead of $\delta_{\rm th}^{\rm S} = 0.00025$; and using this new criterion for $\epsilon = 0.15$ (row 9 of Table 4.1) yields $\delta_{\rm th}^{\rm S} = 0.0089$ instead of $\delta_{\rm th}^{\rm S} = 0.010$. In particular, this change does not affect our main finding of initial conditions for which $\delta_{\rm th}^{\rm H} < \delta_{\rm th}^{\rm S}$.

As expected, for both the NLS and the NLH the threshold values of δ increase with ϵ (i.e., a larger amount of damping is needed to arrest collapse of beams with higher input power). For $\epsilon = 0.06$ and $\epsilon = 0.07$ the input power is below critical. Therefore, both the NLS and the NLH have global solutions for $\delta = 0$. This behavior for the NLH holds (at least) until $\epsilon = 0.072165819$, which corresponds to the last subcritical value[5] that we have checked, $N(0)$ being equal to 99% of $N_{\rm c}$.

Starting from $\epsilon = 3\pi/2k_0^2 \approx 0.073631077$, which corresponds exactly to $N(0) = N_{\rm c}$, the NLH requires a certain positive amount of damping δ to maintain the regularity of the solution. For the NLS, the solution with no damping remains regular until $\epsilon = 0.75$, which corresponds to $p = N(0)/N_{\rm c} = 1.009$. Indeed, it is known that $N_{\rm c}$ is only a *lower bound* for the threshold power for NLS collapse, and that any initial condition which does blow up, and whose amplitude $|\psi_0|$ is not equal to the ground state profile $3^{1/4}\sqrt{\mathrm{sech}(2r)}$, has power strictly above $N_{\rm c}$ (see [7, 16, 17]). In our simulations we have discovered that for $\epsilon = 3\pi/2k_0^2$, which is the critical value for the NLH, as well as for the moderately supercritical values $\epsilon = 0.075$, $\epsilon = 0.08$, and $\epsilon = 0.1$, when the input power $N(0)$ is only slightly above $N_{\rm c}$, the threshold damping for the NLH is larger[6] than that for the NLS: $\delta_{\rm th}^{\rm H} > \delta_{\rm th}^{\rm S}$. However, for input powers that are equal to or higher than $1.30N_{\rm c}$ (which corresponds to $\epsilon = 0.125$) this trend

[5] As mentioned in section 1, for larger subcritical values of $N(0)$ the convergence of nonlinear iterations becomes prohibitively slow.

[6] We recall that the values of $\delta_{\rm th}^{\rm H}$ in Table 4.1 are only upper bounds for the threshold; lower values may be obtained by refining the grid and/or enlarging the computational domain.

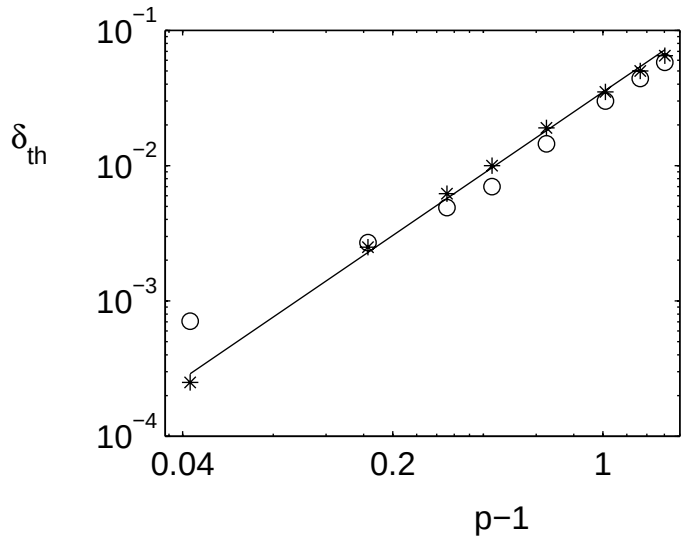


FIG. 4.1. *Threshold values* $\delta_{\rm th}^{\rm H}$ *(open circles "∘") and* $\delta_{\rm th}^{\rm S}$ *(asterisks "∗") as a function of* $(p-1)$ *for the data in Table* 4.1. *The solid line* $0.035(p-1)^{1.517}$ *is the best fit to the values of* $\delta_{\rm th}^{\rm S}$.

reverses (see Table 4.1), and we obtain $\delta_{\rm th}^{\rm H} < \delta_{\rm th}^{\rm S}$. Thus, for $N(0) \geq 1.30N_{\rm c}$,[7] there must be other mechanisms in the NLH not present in the NLS that help suppress the formation of singularity in the solution. Therefore, we may conclude that in this regime *nonparaxiality and backscattering help arrest collapse of nonlinear waves.*[8]

In [6] Fibich has used asymptotic analysis to show that

$$\delta_{\rm th}^{\rm S} \sim c(p-1)^{3/2}, \tag{4.2}$$

where p is the fractional critical power (4.1). In Figure 4.1 we put this theoretical prediction to a test by plotting the values of $\delta_{\rm th}^{\rm S}$ and $\delta_{\rm th}^{\rm H}$ as a function of $(p-1)$. When we computed the best fit of the values of $\delta_{\rm th}^{\rm S}$ with the two-parameter family of curves $\delta_{\rm th} = c(p-1)^{\alpha}$, we obtained $\alpha = 1.517$, which is in excellent agreement with formula (4.2). Relation (4.2) also provides a good approximation to the data points $\delta_{\rm th}^{\rm H}$; see Figure 4.1. The only exception is the lowest-power NLH data point in Figure 4.1 that corresponds to $\epsilon = 0.08$ (row 6 in Table 4.1), for which the value of $\delta_{\rm th}^{\rm H}$ has most likely been overpredicted numerically because of the computational constraints discussed previously.

In Figure 4.2 we plot the on-axis ($r = 0$) amplitudes of the NLH and NLS solutions for $\epsilon = 0.2$ and various values of δ. The on-axis behavior is most representative of the physical processes that we are studying, because for symmetric beams this is the location of the peak intensity. When $\delta = \delta_{\rm th}^{\rm H} = 0.0145$, the NLH solution exists globally, but the NLS solution becomes singular at a finite propagation distance. As the value of damping increases, both the NLS and the NLH solutions undergo less

[7] More precisely, $N(0)$ higher than some value between $1.16N_{\rm c}$ and $1.30N_{\rm c}$.

[8] The fact that for the input values just above the critical power we do not observe nonparaxiality and backscattering helping arrest the collapse is apparently the "continuation" of the fact that we have been unable yet to solve the undamped NLH for $N(0) \geq N_{\rm c}$, even though the NLS solution exists globally for $N(0) \leq 1.009N_{\rm c}$. The reasons are probably the same in both cases; those for the latter have already been discussed in the section 1.

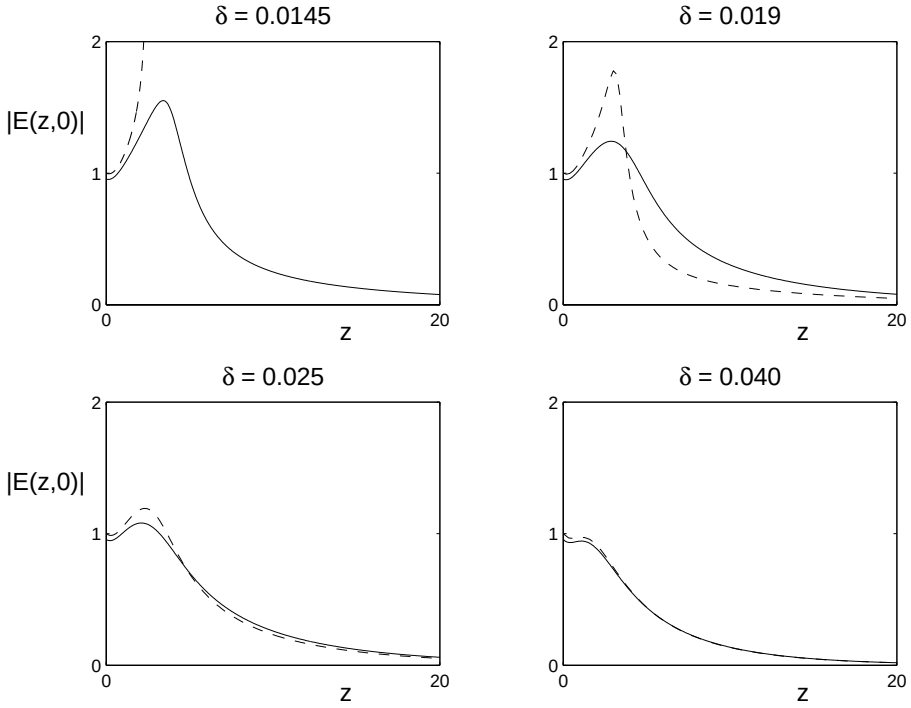


FIG. 4.2. *On-axis amplitude of NLH (solid) and NLS (dashes) solutions for* $\epsilon = 0.2$ *and various values of* δ.

focusing. For all the cases for which both solutions remain regular, the NLS solution curve is higher than the NLH one from $z = 0$ until its maximum, i.e., the point of the arrest of collapse. This provides additional support to the foregoing conclusion that nonparaxiality and backscattering arrest collapse of nonlinear waves. Note that after the collapse has been arrested, the NLS solution becomes lower than that of the NLH. One possible explanation for this is that the NLS solution is undergoing higher focusing, and hence it loses more power due to damping.

We emphasize that at $z = 0$ the NLH solution is not equal to $E^0_{\rm inc}$; see Figure 4.2. The difference between the two is due to backscattering and can be used to quantify the level of backscattering for a particular setting; see [10, 12].[9] In Table 4.2 we provide the values of maximum self-focusing and maximum backscattering in the NLH, defined as $\max_{r,z} |E(z,r)|$ and $\max_r |E(0,r) - E^0_{\rm inc}(r)|$, respectively, for various values of ϵ and δ. The dash "—" in a particular cell of Table 4.2 means that the level of damping was insufficient to guarantee the convergence of the numerical algorithm. As expected, for a given level of damping δ, the NLH solution undergoes stronger self-focusing as the nonlinearity coefficient ϵ increases. The level of backscattering also increases with the increase of ϵ. As also expected, for a given input power ϵ, when the damping δ increases, the NLH solution undergoes weaker self-focusing (see Figure 4.2).

[9] There are, in fact, two phenomena that account for the discrepancy between the NLH and NLS curves: nonparaxiality of the forward propagating wave and backscattering. Because the problem is nonlinear, these two mechanisms cannot be easily and explicitly told apart inside the domain. The only location where we can clearly say that the difference is purely due to backscattering is the "inflow" interface $z = 0$; see [10].

TABLE 4.2
Maximum absolute levels of self-focusing and backscattering in the NLH for a variety of ϵ and δ.

	Maximum self-focusing			Maximum backscattering		
	$\delta = 0.0145$	$\delta = 0.0175$	$\delta = 0.0210$	$\delta = 0.0145$	$\delta = 0.0175$	$\delta = 0.0210$
$\epsilon = 0.15$	1.1179	1.0601	1.0162	0.0372	0.0373	0.0373
$\epsilon = 0.175$	1.2718	1.1538	1.0761	0.0420	0.0421	0.0421
$\epsilon = 0.2$	1.5515	1.3158	1.1716	0.0465	0.0466	0.0467
$\epsilon = 0.225$	—	—	1.3242	—	—	0.0509

TABLE 4.3
Maximum absolute levels of self-focusing and backscattering in the NLH for $\epsilon = 0.2$.

Case No.	Damping δ	Max. self-focusing	Max. backscattering
1	0.0145	1.5515	0.0465
2	0.0147	1.5296	0.0465
3	0.0150	1.4992	0.0465
4	0.0155	1.4538	0.0465
5	0.0160	1.4135	0.0466
6	0.0165	1.3776	0.0466
7	0.0170	1.3451	0.0466
8	0.0175	1.3158	0.0466
9	0.0180	1.2892	0.0466
10	0.0190	1.2428	0.0466
11	0.0200	1.2041	0.0466
12	0.0210	1.1716	0.0467

Surprisingly, however, *changing the value of damping δ has very little or no effect on the level of backscattering.* To further corroborate this observation, we picked a particular value of the nonlinearity coefficient, $\epsilon = 0.2$, and ran an additional series of numerical tests with a substantially finer sampling for δ. These results, which are presented in Table 4.3, confirm that backscattering is not affected by linear damping. This phenomenon certainly cannot be explained by saying that linear damping has an overall negligible effect, since its effect on the focusing dynamics can be clearly seen through the changing values of the maximum focusing both in Table 4.2 and in Figure 4.2. At present, we have no good explanation for this surprising observation.

5. Concluding remarks. The question of whether nonparaxiality and backscattering may arrest collapse of nonlinear waves has been open for many years. While the answer to this question is probably positive, no conclusive argument toward it, whether analytical or numerical, has been previously available in the literature. In this study we addressed this question within the framework of the linearly damped NLH and NLS. As has been mentioned, the addition of linear damping is not ad hoc, because it has both physical and mathematical motivation. Methodologically, linear damping provides a very useful "extra dimension" that allows us to efficiently control the solutions of the NLH and NLS. Specifically, the variation along this extra dimension has helped us to numerically identify the regimes for which the NLS solution blows up, while the NLH solution remains regular. In other words, our results furnish the first ever definite numerical evidence that *nonparaxiality and backscattering can arrest collapse.* The question of whether regular solutions to the NLH still exist in the absence of damping remains open. However, we hope that the arguments based on linear damping and the limiting absorption principle may be useful for proving global existence and uniqueness, both for the damped NLH and for the undamped NLH.

REFERENCES

[1] N. Akhmediev, A. Ankiewicz, and J. M. Soto-Crespo, *Does the nonlinear Schrödinger equation correctly describe beam propagation?*, Optics Letters, 18 (1993), pp. 411–413.

[2] N. Akhmediev and J. M. Soto-Crespo, *Generation of a train of three-dimensional optical solitons in a self-focusing medium*, Phys. Rev. A, 47 (1993), pp. 1358–1364.

[3] R. W. Boyd, *Nonlinear Optics*, Academic Press, Boston, 1992.

[4] M. D. Feit and J. A. Fleck, *Beam nonparaxiality, filament formation, and beam breakup in the self-focusing of optical beams*, J. Opt. Soc. Amer. B Opt. Phys., 5 (1988), pp. 633–640.

[5] G. Fibich, *Small beam nonparaxiality arrests self-focusing of optical beams*, Phys. Rev. Lett., 76 (1996), pp. 4356–4359.

[6] G. Fibich, *Self-focusing in the damped nonlinear Schrödinger equation*, SIAM J. Appl. Math., 61 (2001), pp. 1680–1705.

[7] G. Fibich and A. Gaeta, *Critical power for self-focusing in bulk media and in hollow waveguides*, Optics Letters, 25 (2000), pp. 335–337.

[8] G. Fibich and B. Ilan, *Self focusing of elliptic beams: An example of the failure of the aberrationless approximation*, J. Opt. Soc. Amer. B Opt. Phys., 17 (2000), pp. 1749–1758.

[9] G. Fibich and B. Ilan, *Discretization effects in the nonlinear Schrödinger equation*, Appl. Numer. Math., 44 (2003), pp. 63–75.

[10] G. Fibich, B. Ilan, and S. V. Tsynkov, *Computation of nonlinear backscattering using a high-order numerical method*, J. Sci. Comput., 17 (2002), pp. 351–364.

[11] G. Fibich and G. Papanicolaou, *Self-focusing in the perturbed and unperturbed nonlinear Schrödinger equation in critical dimension*, SIAM J. Appl. Math., 60 (1999), pp. 183–240.

[12] G. Fibich and S. V. Tsynkov, *High-order two-way artificial boundary conditions for nonlinear wave propagation with backscattering*, J. Comput. Phys., 171 (2001), pp. 632–677.

[13] S. K. Godunov and V. S. Ryaben'kii, *Canonical forms of systems of ordinary linear differential equations with constant coefficients*, U.S.S.R. Comput. Math. and Math. Phys., 3 (1963), pp. 281–295.

[14] J. D. Jackson, *Classical Electrodynamics*, Wiley, New-York, 1975.

[15] P. L. Kelley, *Self-focusing of optical beams*, Phys. Rev. Lett., 15 (1965), pp. 1005–1008.

[16] F. Merle, *On uniqueness and continuation properties after blow-up time of self-similar solutions of nonlinear Schrödinger equation with critical exponent and critical mass*, Comm. Pure Appl. Math., 45 (1992), pp. 203–254.

[17] F. Merle, *Determination of blow-up solutions with minimal mass for nonlinear Schrödinger equation with critical power*, Duke Math. J., 69 (1993), pp. 427–454.

[18] M. N. Mishkov and V. S. Ryaben'kii, *Investigation of artificial boundary conditions constructed by periodization and the introduction of a small parameter for subsonic flow problems*, Mathematical Modeling, 10 (1998), pp. 87–98 (In Russian).

[19] A. C. Newell and J. V. Moloney, *Nonlinear Optics*, Addison-Wesley, Redwood City, Calif., 1992.

[20] V. P. Palamodov, *Conditions at infinity for correct solvability of a certain class of equations of the form* $p(i\frac{\partial}{\partial x})u = f$, Dokl. Akad. Nauk SSSR, 129 (1959), pp. 740–743.

[21] V. S. Ryaben'kii, *Necessary and sufficient conditions for good definition of boundary value problems for systems of ordinary difference equations*, Comput. Math. Math. Phys., 4 (1964), pp. 43–61.

[22] V. S. Ryaben'kii and S. V. Tsynkov, *Artificial boundary conditions for the numerical solution of external viscous flow problems*, SIAM J. Numer. Anal., 32 (1995), pp. 1355–1389.

[23] S. V. Tsynkov, *An application of nonlocal external conditions to viscous flow computations*, J. Comput. Phys., 116 (1995), pp. 212–225.

[24] S. V. Tsynkov, *Numerical solution of problems on unbounded domains. A review*, Appl. Numer. Math., 27 (1998), pp. 465–532.

[25] S. V. Tsynkov, *External boundary conditions for three-dimensional problems of computational aerodynamics*, SIAM J. Sci. Comput., 21 (1999), pp. 166–206.

[26] S. V. Tsynkov, E. Turkel, and S. Abarbanel, *External flow computations using global boundary conditions*, AIAA J., 34 (1996), pp. 700–706.

[27] V. S. Vladimirov, *Equations of Mathematical Physics*, Marcel Dekker, New York, 1971.

[28] M. I. Weinstein, *Nonlinear Schrödinger equations and sharp interpolation estimates*, Comm. Math. Phys., 87 (1983), pp. 567–576.

SIAM J. APPL. MATH.
Vol. 63, No. 5, pp. 1737–1763

THE RADIO-FREQUENCY DRIVEN PLASMA SHEATH: ASYMPTOTICS AND ANALYSIS*

M. SLEMROD†

Abstract. This paper considers the dynamics of a radio-frequency driven plasma consisting of ions and electrons. The method of matched asymptotic expansions is used to derive the dynamics in bulk quasi-neutral plasma, transition, and sheath regions. Furthermore, a constructive existence theorem is presented for solutions of the system governing sheath dynamics.

Key words. plasma, sheath

AMS subject classifications. 35Q35, 34E10, 34B16

DOI. 10.1137/S0036139902411831

1. Introduction. The purpose of this paper is to provide a unified discussion of the dynamics of a bounded plasma consisting of ions and electrons sustained by a radio-frequency (rf) current. Such configurations occur naturally in reactive ion etching [1]. Various models have been introduced to describe the plasma behavior [2], [3], [4]; a rather complete discussion is found in [1]. In fact the main interest, both physical and mathematical, for the rf driven plasma is the formation of a space charge sheath boundary layer near the bounding wall. In this boundary layer the plasma exhibits time periodic motion, with the same period as the driving current and with an electric potential $\varphi(\xi,\tau)$ satisfying to leading order in a small parameter the nonlocal (in time τ)-local (in space ξ) system

$$\begin{aligned} &\frac{\partial^2\varphi}{\partial\xi^2} = (1+2\overline{\varphi})^{-1/2} - e^{-\varphi}, \quad -\infty<\xi<0,\ 0\le\tau\le 1,\\ &\frac{\partial\varphi}{\partial\xi} = F(\tau) \quad \text{at } \xi=0,\\ &\varphi\to 0,\ \frac{\partial\varphi}{\partial\xi}\to 0 \quad \text{as } \xi\to-\infty, \end{aligned} \tag{1.1}$$

where $\overline{\varphi}(\xi)=\int_0^1\varphi(\xi,\tau)d\tau$, and $F(\tau)>0$ is continuously differentiable with period 1.

System (1.1) is rather intriguing. It possesses no time τ derivatives, yet temporal behavior is nontrivial due to the appearance of the time average $\overline{\varphi}$ in the differential equation for φ. Hence two questions come immediately to mind:

i. Where does the nonlocal system (1.1) come from?

ii. Does (1.1) have a solution, and how do we find it?

This paper answers both questions. It is shown that the nonlocal system arises from the method of matched asymptotic expansions applied to the classical two fluid model for a collisional plasma undergoing ionization.

Modulo the matching, the physics is well understood and can be found in references [2], [3], [4], and the matching itself without details has been suggested in [4]

*Received by the editors July 22, 2002; accepted for publication (in revised form) January 21, 2003; published electronically July 26, 2003. This research was sponsored in part by grants DMS-9803223 and DMS-00711463 from the National Science Foundation.

http://www.siam.org/journals/siap/63-5/41183.html

†Mathematics Department, University of Wisconsin, Madison, WI 53706 (slemrod@math.wisc.edu).

and is similar to that given in the paper of Franklin and Ockendon [5], where the plasma was undriven. Nevertheless, the matching is provided so that the uninitiated reader will have a straightforward self-contained derivation of (1.1). (While this paper considers the case of an essentially collisionless sheath, the interested reader can also consult the paper of Gegick and Young [6], which considers the opposite limit of a collisional sheath.)

As to solvability of (1.1), the answer is that (1.1) does indeed possess a time periodic solution. More valuable, however, is that the proof of existence is constructive. Simply put, the proof is done in two stages as follows.

Step 1. *Regularization.* The problem (1.1) is regularized in the form

$$\begin{aligned} &\frac{\partial^2\varphi}{\partial\xi^2} = (1+2\overline{\varphi})^{-1/2} - e^{-\varphi} + \mu\varphi, \qquad -\frac{1}{\mu} < \xi < 0,\ 0 \le \tau \le 1, \\ &\frac{\partial\varphi}{\partial\xi} = F(\tau) \quad \text{at } \xi = 0, \\ &\varphi = 0 \quad \text{at } \xi = -\frac{1}{\mu}. \end{aligned} \tag{1.2}$$

Here $\mu > 0$ represents a small parameter.

Of course (1.2) is still a nonlocal problem, but it admits a *variational principle*. If

$$J(\varphi) \stackrel{\text{def}}{=} \int_0^1 \int_{-\frac{1}{\mu}}^0 \frac{1}{2}\left(\frac{\partial\varphi}{\partial\xi}\right)^2 + (1+2\overline{\varphi})^{1/2} + e^{-\varphi} + \mu\frac{\varphi^2}{2} - F(\tau)\frac{\partial\varphi}{\partial\xi}\,d\xi\,d\tau,$$

then the Euler–Lagrange equations corresponding to minimizing $J(\varphi)$ over a suitable class of functions φ, vanishing at $\xi = -\frac{1}{\mu}$, are precisely (1.2). Hence the existence of solutions of (1.2) and their explicit computation in practice can be accomplished by minimizing $J(\varphi)$.

Step 2. *Passage to the limit* $\mu \to 0$. In fact, it is this process that is the most delicate. As in all such limiting arguments, precompactness of a sequence of solutions to (1.2) must be established; this is accomplished with straightforward arguments based on the behavior of solutions of (1.2) as functions of ξ for fixed τ values.

In practice, of course, one cannot numerically resolve either Step 1 or Step 2 directly. However, it seems that finite-dimensional minimization problems approximating Step 1 implemented on larger and larger spatial ξ domains would generate a sequence of approximations. The results given here guarantee a convergent subsequence.

The paper is divided into two sections after this Introduction. Section 2 uses the method of matched asymptotic expansions, similar in spirit to [5], to derive the dynamics of the plasma in (A) the bulk quasi-neutral plasma regime, (B) a transition regime between bulk plasma and the space charge sheath, and (C) the sheath itself. In subsection (D) a short discussion of higher order matching and the need for the transition layer (B) is provided. Section 3 is devoted to analysis of (1.1). Steps 1 and 2 as outlined above are carried out, and existence of a periodic-in-τ solution of (1.1) is proven.

2. Derivation of the basic equations. Let n_i denote ion density, n_e electron density, Φ electric potential, n_{ch} characteristic charged particle density, u_i ion velocity, u_e electron velocity, θ_e electron temperature. We consider a one-dimensional spatial

domain $-R < X < X_w$, where X_w is the location of the wall. T denotes time. For small electron mass, we assume the Boltzmann relation

$$n_e = n_{ch} \exp\left(\frac{e\Phi}{k\theta_e}\right),$$

where e is the electron charge and k is Boltzmann's constant. Denote by m_i the mass of the ions. Then the conservation laws of mass for the ions and electrons and of momentum for the ions are

$$\begin{aligned}
\frac{\partial n_i}{\partial T} + \frac{\partial}{\partial X}(n_i u_i) &= Z n_e, \\
\frac{\partial n_e}{\partial T} + \frac{\partial}{\partial X}(n_e u_e) &= Z n_e, \\
m_i \frac{\partial (n_i u_i)}{\partial T} + m_i \frac{\partial}{\partial X}(n_i u_i^2) &= -e n_i \frac{\partial \Phi}{\partial X} - \nu(u_i) m_i u_i n_i,
\end{aligned}$$

where ν is the ion friction coefficient and Z is rate of ionization. (We assume cold ions and hence the ion pressure is zero.) For simplicity we take $\nu(u_i) = \frac{|u_i|}{\lambda}$; $\lambda > 0$ is the constant ion collision mean free path. In addition, Φ satisfies Poisson's equation

$$-\frac{\epsilon_0}{e}\frac{\partial^2 \Phi}{\partial X^2} = n_i - n_e,$$

where ϵ_0 is the permittivity of free space.

The above equations for n_i, n_e, u_i, Φ may be simplified if we introduce the quantities

$$c_s = \sqrt{\frac{k\theta_e}{m_i}}, \qquad \lambda_D = \sqrt{\frac{\epsilon_0 k \theta_e}{n_{cn} e^2}}$$

representing ion sound velocity and Debye length,

$$\begin{aligned}
&t = \frac{T c_s}{\lambda},\ x = \frac{X}{\lambda},\ \varphi = \frac{-e\Phi}{k\theta_e},\ n_+ = \frac{n_i}{n_{ch}},\ n_- = \frac{n_e}{n_{ch}}, \\
&u_+ = \frac{u_i}{c_s},\ u_- = \frac{u_e}{c_s},\ \epsilon = \frac{\lambda_D}{\lambda},\ z = \frac{Z\lambda}{c_s},
\end{aligned}$$

and set $n_- = e^{-\varphi}$ by adjusting the zero point of the potential Φ.

In the new dependent variables $n_+, n_-, u_+, u_-, \varphi$ the balance laws are now

$$\frac{\partial n_+}{\partial t} + \frac{\partial}{\partial x}(n_+ u_+) = z n_-, \tag{2.1}$$

$$\frac{\partial n_-}{\partial t} + \frac{\partial}{\partial x}(n_- u_-) = z n_-, \tag{2.2}$$

$$\frac{\partial u_+}{\partial t} + u_+ \frac{\partial u_+}{\partial x} = \frac{\partial \varphi}{\partial x} - |u_+| u_+ - z u_+ \left(\frac{n_-}{n_+}\right), \tag{2.3}$$

$$\epsilon^2 \frac{\partial^2 \varphi}{\partial x^2} = n_+ - n_-, \tag{2.4}$$

$$n_- = e^{-\varphi}. \tag{2.5}$$

The system is now considered on a bounded domain $-L < x < x_w$, and at $x = -L$ ($L = \frac{R}{\lambda}, x_w = \frac{X_w}{\lambda}$) we impose boundary conditions

$$\varphi = \varphi_L, \quad u_+ = u_L, \quad u_- = u_L^e\left(\frac{t}{\epsilon^2 p}\right), \quad n_+ = u_L^{-1}, \quad n_- = u_L^{-1}, \tag{2.6}$$

where $u_L = e^{\varphi_L}$ and $\varphi_L \ll 0$ so that $0 < u_L \ll 1$. The boundary value u_L^e will be determined in subsection A to follow. At the wall $x = x_w$ we impose the electron velocity

$$u_- = u_{-_{\text{wall}}}, \tag{2.7}$$

where $u_{-_{\text{wall}}}$ is independent of ϵ.

One of the delicate points in this asymptotic derivation is that the wall position x_w is itself to be determined. This is done in subsection C.

Differentiation of the Poisson equation (2.4) with respect to t yields

$$\frac{\epsilon^2 \partial^3 \varphi}{\partial t \partial x^2} = \frac{\partial n_+}{\partial t} - \frac{\partial n}{\partial t} = -\frac{\partial}{\partial x}(n_+ u_+) + \frac{\partial}{\partial x}(n_- u_-),$$

and hence

$$\epsilon^2 \frac{\partial^2 \varphi}{\partial x \partial t} + n_+ u_+ - n_- u_- = \frac{f_1(t)}{\epsilon} \tag{2.8}$$

for all x. Here $\frac{f_1(t)}{\epsilon}$ is the prescribed rf current. We take $f_1(t)$ to be a periodic function in t with period $\epsilon^2 p$, $\int_0^{\epsilon^2 p} f_1(t)dt = 0$. Set $t = \epsilon^2 \tau p$, $f(\tau) \overset{\text{def}}{=} f_1(\epsilon^2 \tau p)$ so that $f(\tau)$ is periodic with period 1.

In the new $\tau = t/\epsilon^2 p$ time variable, the Euler–Poisson equations become

$$\frac{1}{\epsilon^2 p}\frac{\partial n_+}{\partial \tau} + \frac{\partial}{\partial x}(n_+ u_+) = z n_-, \tag{2.9}$$

$$\frac{1}{\epsilon^2 p}\frac{\partial u_+}{\partial \tau} + u_+ \frac{\partial u_+}{\partial x} = \frac{\partial \varphi}{\partial x} - |u_+| u_+ - z u_+ \left(\frac{n_-}{n_+}\right), \tag{2.10}$$

$$\epsilon^2 \frac{\partial^2 \varphi}{\partial x^2} = n_+ - n_-, \tag{2.11}$$

$$\frac{1}{p}\frac{\partial^2 \varphi}{\partial x \partial \tau} + n_+ u_+ - n_- u_- = \frac{f(\tau)}{\epsilon}, \tag{2.12}$$

$$n_- = e^{-\varphi}. \tag{2.13}$$

Notice that the current equation (2.12) is now used instead of (2.2) since (2.9), (2.11), (2.12) imply (2.2).

A. Bulk plasma outer solution. First consider a region away from the wall. Write $n_+(x, \tau)$, etc., as asymptotic expansions in ϵ:

$$\begin{aligned} n_+ &= n_0 + \epsilon n_1 + \cdots, \\ u_+ &= u_0 + \epsilon u_1 + \cdots, \\ \varphi &= \varphi_0 + \epsilon \varphi_1 + \cdots, \\ u_- &= \frac{u_0^-}{\epsilon} + u_1^- + \epsilon u_2^-, \end{aligned} \tag{2.14}$$

where periodicity in τ is assumed.

Substitution of (2.14) into the rescaled Euler–Poisson system (2.9)–(2.12) and balancing powers of ϵ yields

$$\frac{\partial n_0}{\partial \tau} = \frac{\partial n_1}{\partial \tau} = \frac{\partial u_0}{\partial \tau} = \frac{\partial u_1}{\partial \tau} = 0,$$

and hence n_0, n_1, u_0, u_1 are independent of τ and depend only on x. Furthermore, (2.9) implies

$$\frac{1}{p}\frac{\partial n_2}{\partial \tau} + \frac{\partial}{\partial x}(n_0 u_0) = z e^{-\varphi}. \tag{2.15}$$

The Poisson equation (2.11) implies $n_0 = e^{-\varphi_0}$, and φ_0 is also independent of τ. Additionally, integration of (2.15) from $\tau = 0$ to $\tau = 1$ and the periodicity of n_2 in τ imply

$$\frac{d}{dx}(n_0 u_0) = z n_0, \tag{2.16}$$

which combined with (2.10) and $n_0 = e^{-\varphi_0}$ yields

$$\frac{du_0}{dx} = \frac{z(1+u_0^2) + u_0^3}{1 - u_0^2}. \tag{2.17}$$

Hence u_0 is monotone increasing in x and

$$\frac{du_0}{dx} \to \infty \quad \text{as} \quad u_0 \to 1. \tag{2.18}$$

Let x_B (the Bohm point) denote that value of x for which $\lim_{x \to x_B} u_0(x) = 1$.

From (2.16), once $u_0(x)$ is determined, n_0 and hence φ_0 are determined as well on $(-L, x_B)$. The current equation (2.12) determines

$$u_0^- = -e^{\varphi_0} f(\tau). \tag{2.19}$$

Returning to boundary condition (2.6), we see that (2.19) implies $u_L^e(\tau) = e^{\varphi_L} f(\tau)$. *Thus in fact the current f is determined by the boundary condition or vice versa.*

To determine the nature of the singularity at x_B, set $u_0 = 1 + U$ and substitute into (2.17). Then to leading order in U we find

$$\frac{d}{dx} U^2 = -(2z+1),$$

and, setting $U(x_B) = 0$, we see

$$U(x) = -((2z+1)(x_B - x))^{1/2}$$

and

$$u_0(x) \sim 1 - ((2z+1)(x_B - x))^{1/2} \quad \text{as} \quad x \to x_B. \tag{2.20}$$

From (2.16) we then see

$$\frac{dn_0}{dx} \sim -\frac{n_0}{2}(2z+1)^{-1/2} \quad \text{as} \quad x \nearrow x_B,$$

and since $\frac{d\varphi_0}{dx} = -\frac{1}{n_0}\frac{dn_0}{dx}$ we see

$$\frac{d\varphi_0}{dx} \sim \frac{(2z+1)^{-1/2}}{2}(x_B - x)^{1/2} \quad \text{as} \quad x \nearrow x_B.$$

Hence

$$\varphi_0(x) \sim \varphi_0(x_B) - (2z+1)^{1/2}(x_B - x)^{1/2} \quad \text{as} \quad x \nearrow x_B, \tag{2.21}$$

where $\varphi_B = \varphi_0(x_B) = -\ln n_0(x_B)$ and $n_B \doteq n_0(x_B)$ are determined by solving (2.16) subject to the boundary conditions (2.6).

Let us employ the normalization

$$\varphi_B = 0, \qquad n_B = 1,$$

which will provide a simplification of the algebraic computation to follow. Of course this will shift the value of φ and hence φ_L by a constant.

B. Transition layer. Introduce a new space variable $\zeta = \frac{x - x_B}{\delta}, \delta = \epsilon^{4/5}, \zeta < 0$, so that the Euler–Poisson system (2.9)–(2.12) becomes

$$\frac{1}{\delta^{3/2}p}\frac{\partial n_+}{\partial \tau} + \frac{\partial}{\partial \zeta}(n_+ u_+) = \delta z e^{-\varphi}, \tag{2.22}$$

$$\frac{1}{\delta^{3/2}p}\frac{\partial u_+}{\partial \tau} + u_+\frac{\partial u_+}{\partial \zeta} = \frac{\partial \varphi}{\partial \zeta} - \delta |u_+| u_+ - \delta z \frac{u_+ e^{-\varphi}}{n_+}, \tag{2.23}$$

$$\delta^{1/2}\frac{\partial^2 \varphi}{\partial \zeta^2} = n_+ - e^{-\varphi}, \tag{2.24}$$

$$\frac{1}{\delta p}\frac{\partial^2 \varphi}{\partial \zeta \partial \tau} + n_+ u_+ - n_- u_- = \frac{f(\tau)}{\epsilon}. \tag{2.25}$$

Again expand

$$\begin{aligned} n_+ &= 1 + \delta^{1/2} n_1 + \cdots, \\ u_+ &= 1 + \delta^{1/2} u_1 + \cdots, \\ \varphi &= \delta^{1/2}\varphi_1 + \cdots, \\ u_- &= \frac{u_0^-}{\delta^{5/4}} + \frac{u_1^-}{\delta} + \cdots, \\ x_w &= x_B + \delta x_{w_1} + \cdots, \end{aligned}$$

where all the indicated terms except those for x_w are now functions of ζ, τ and are periodic in τ with period 1. The expansion of the wall location x_w will become crucial in subsection C.

Substitute the expansions into (2.22)–(2.25). We then see, equating powers of δ, that

$$\frac{\partial n_1}{\partial \tau} = \frac{\partial n_2}{\partial \tau} = \frac{\partial n_3}{\partial \tau} = \frac{\partial u_1}{\partial \tau} = \frac{\partial u_2}{\partial \tau} = \frac{\partial u_3}{\partial \tau} = 0$$

and $n_1, n_2, n_3, u_1, u_2, u_3$ are independent of τ. Also (2.22) and the periodicity of n_4, n_5 in τ imply

$$\frac{\partial}{\partial \zeta}(n_1 + u_1) = 0, \qquad \frac{\partial}{\partial \zeta}(n_1 u_1 + n_2 + u_2) = z, \tag{2.26}$$

whereas (2.23) and the periodicity of u_4, u_5 in τ imply

$$\frac{\partial}{\partial \zeta}(u_1 - \overline{\varphi}_1) = 0, \qquad \frac{\partial u_2}{\partial \zeta} + u_1\frac{\partial u_1}{\partial \zeta} = \frac{\partial \overline{\varphi}_2}{\partial \zeta} - 1 - z. \tag{2.27}$$

Equating order $\delta^{1/2}$ and δ terms in the Poisson equation (2.24) yields

$$n_1 + \varphi_1 = 0, \qquad n_2 + \varphi_2 - \frac{\varphi_1^2}{2} = \frac{\partial^2 \varphi_1}{\partial \zeta^2}. \tag{2.28}$$

Hence φ_1, φ_2 are independent of τ as well, and $\overline{\varphi}_2 = \varphi_2$ in (2.27).

For x in a transition matching regime (say, $x - x_B = -\epsilon^\alpha, 0 < \alpha < 4/5$) we have $x \to x_B-$ and $\zeta \to -\infty$ as $\epsilon \to 0$. Hence the matching condition

$$\lim_{\zeta \to -\infty} n_+ u_+ = 1$$

is inherited from the bulk plasma solution of subsection A. Since $n_+ = 1 + \delta^{1/2} n_1 + \cdots$, $u_+ = 1 + \delta^{1/2} u_1 + \cdots$, in this middle transition layer we see

$$\lim_{\zeta \to -\infty} n_1 + u_1 = 0;$$

however, (2.26) tells us that $n_1 + u_1 = \text{const}$, and hence

$$n_1 + u_1 = 0. \tag{2.29}$$

Also from (2.27), (2.28), $u_1 - \overline{\varphi}_1 = \text{const}$, $n_1 = -\varphi_1$, and since $\varphi_1 = \overline{\varphi}_1$ we have

$$u_1 = \varphi_1 = -n_1. \tag{2.30}$$

Next differentiate (2.28) with respect to ζ, and to the resulting expression add (2.27). This yields

$$\frac{\partial}{\partial \zeta}(u_2 + n_2) - \frac{\partial^3 \varphi_1}{\partial \zeta^3} + u_1 \frac{\partial u_1}{\partial \zeta} = -1 - z + \varphi_1 \frac{\partial \varphi_1}{\partial \zeta}.$$

However, from (2.26), $\frac{\partial}{\partial \zeta}(u_2 + n_2) = z - \frac{\partial}{\partial \zeta}(n_1 u_1)$, and so

$$2z - \frac{\partial}{\partial \zeta}(u_1 u_1) - \frac{\partial^3 \varphi_1}{\partial \zeta^3} + u_1 \frac{\partial u_1}{\partial \zeta} = -1 + \varphi_1 \frac{\partial \varphi_1}{\partial \zeta}.$$

Finally, use (2.30) to obtain

$$\frac{d^3 \varphi_1}{d\zeta^3} = \frac{d}{d\zeta}(\varphi_1^2) + 1 + 2z. \tag{2.31}$$

Recall that in subsection A we have shown for the bulk plasma potential $\varphi(x) \sim -(2z+1)^{1/2}(x_B - x)^{1/2}$ as $x \to x_B-$. Hence the matching condition for $\varphi(\zeta) = \delta^{1/2}\varphi_1(\zeta) + \cdots$ as $\zeta \to -\infty$ is

$$\varphi_1(\zeta) \to -(2z+1)^{1/2}(-\zeta)^{1/2} \quad \text{as} \quad \zeta \to -\infty. \tag{2.32}$$

The integration of (2.31) and use of (2.32) yield the Painlevé 1 equation

$$\frac{d^2 \varphi_1}{d\zeta^2} = \varphi_1^2 + (1 + 2z)\zeta \quad \text{for} \quad -\infty < \zeta < 0. \tag{2.33}$$

If in addition we wish to match the electric field $\frac{\partial \varphi}{\partial x}$ of bulk plasma of subsection A with the transition layer solution of this section, we need to match the additional condition

$$\frac{\partial \varphi}{\partial x}(x) - \left(\frac{(2z+1)^{1/2}}{2}\right)(x - x_B)^{-1/2} \to 0 \quad \text{as} \quad x \to x_B,$$

which implies

$$\varphi_1'(\zeta) - \frac{(2z+1)^{1/2}}{2}(-\zeta)^{-1/2} \to 0 \quad \text{as} \quad \zeta \to -\infty. \tag{2.34}$$

In [7], [8] it is shown there is only one solution of (2.33) that satisfies (2.31) and (2.34), i.e., the unique monotone increasing solution of the Painlevé 1 equation.

Hence $\varphi_1 = u_1 = -n_1$, where φ_1 is the unique monotone increasing solution of the Painlevé 1 equation (2.33). Finally, a simple substitution of our asymptotic expansions into (2.25) yields $u_0^- = f(\tau)$.

C. Sheath layer. For the study of the sheath boundary layer near $x = x_w$ we introduce yet another space variable $\xi = \frac{x-x_w}{\epsilon}, \xi \le 0$. In this scaling, (2.9)–(2.12) become

$$\frac{1}{\epsilon p}\frac{\partial n_+}{\partial \tau} + \frac{\partial}{\partial \xi}(n_+ u_+) = \epsilon z e^{-\varphi}, \tag{2.35}$$

$$\frac{1}{\epsilon p}\frac{\partial u_+}{\partial \tau} + u_+\frac{\partial u_+}{\partial \xi} = \frac{\partial \varphi}{\partial \xi} - \epsilon |u_+| u_+ - \epsilon u_+ \frac{e^{-\varphi}}{n_+}, \tag{2.36}$$

$$\frac{\partial^2 \varphi}{\partial \xi^2} = n_+ - e^{-\varphi}, \tag{2.37}$$

$$\frac{1}{\epsilon p}\frac{\partial^2 \varphi}{\partial \xi \partial \tau} + n_+ u_+ - n_- u_- = \frac{f(\tau)}{\epsilon}. \tag{2.38}$$

Again we expand the dependent variables in asymptotic expansions:

$$\begin{aligned} n_+ &= n_0 + \epsilon n_1 + \cdots, \\ u_+ &= u_0 + \epsilon u_1 + \cdots, \\ \varphi &= \varphi_0 + \epsilon \varphi_1 + \cdots, \\ u_- &= \frac{u_0^-}{\epsilon} + u_1^- + \cdots. \end{aligned} \tag{2.39}$$

Again all terms depend on τ, ξ, where periodicity in τ with period 1 is assumed.

Substituting (2.39) into (2.35)–(2.38) and equating powers of ϵ, we find

$$\frac{\partial n_0}{\partial \tau} = \frac{\partial u_0}{\partial \tau} = 0,$$

and hence u_0, n_0 are independent of τ. Also balancing terms of order one in (2.36) gives

$$\frac{1}{p}\frac{\partial n_1}{\partial \tau} + \frac{\partial}{\partial \xi}(n_0 u_0) = 0,$$

and integration in τ from 0 to 1 and the periodicity of n_1 imply

$$\frac{d}{d\xi}(n_0 u_0) = 0.$$

In fact we also see that n_1 is independent of τ.

Recall that the middle region is defined by $\zeta = \frac{x-x_B}{\delta}$ and the sheath region by $\xi = \frac{x-x_w}{\epsilon}$. Hence in a matching regime, e.g., $x - x_w = -\epsilon^\alpha, \frac{4}{5} < \alpha < 1$, we have $\zeta = -\epsilon^{\alpha-4/5} + x_{w_1}$, $\xi = -\epsilon^{\alpha-1}$, and $\zeta \to x_{w_1}$, $\xi \to -\infty$ as $\epsilon \to 0$.

Substitution into (2.36) of our asymptotic expansions also yields

$$\frac{1}{p}\frac{\partial u_1}{\partial \tau} + u_0\frac{\partial u_0}{\partial \xi} = \frac{\partial \varphi_0}{\partial \xi}, \tag{2.40}$$

and again, since u_1 is periodic in τ with period one, we see

$$u_0\frac{du_0}{d\xi} = \frac{d}{d\xi}\overline{\varphi}_0$$

and hence

$$\frac{u_0^2}{2} - \overline{\varphi}_0 = \text{ const.}$$

Recall in the middle region (subsection B)

$$\varphi(\zeta,\tau) = 1 + \delta^{1/2}\varphi_1 + \cdots,$$
$$u_+(\zeta,\tau) = 1 + \delta^{1/2}u_1 + \cdots;$$

with $x - x_w = O(\epsilon^\alpha)(\frac{4}{5} < \alpha < 1)$ we know $\zeta \to x_{w_1}$, $\xi \to -\infty$.

Thus matching the middle and sheath regions must be done as $\zeta \to x_{w_1}$ from the middle region and $\xi \to -\infty$ from the sheath region.

Since $n_+u_+ = 1$ in the middle region trivially, we see that matching requires $n_0u_0 = 1$ for all ξ in the sheath region.

Matching the ion velocity u_0 and potential φ_0 to the middle region requires

$$u_0 \to 1,\ \varphi_0 \to 0 \quad \text{as} \quad \xi \to -\infty,$$

and hence the above const. $= \frac{1}{2}$ and we have

$$\frac{u_0^2}{2} - \overline{\varphi}_0 = \frac{1}{2}.$$

But since $n_0u_0 = 1$, we have trivially

$$n_0 = (1 + 2\overline{\varphi}_0)^{-1/2}. \tag{2.41}$$

Substitution of the asymptotic expansions into the Poisson equation (2.37) gives

$$\frac{\partial^2 \varphi_0}{\partial \xi^2} = n_0 - e^{-\varphi_0},$$

and use of the relation (2.41) then gives

$$\frac{\partial^2 \varphi_0}{\partial \xi^2} = (1 + 2\overline{\varphi}_0)^{-1/2} - e^{-\varphi_0}. \tag{2.42}$$

Next substitute the asymptotic expansion into the current equation (2.38) to obtain

$$\frac{1}{p}\frac{\partial^2 \varphi_0}{\partial \xi \partial \tau} - e^{-\varphi_0}u_0^- = f(\tau). \tag{2.43}$$

But recall

$$u_- = \frac{u_0^-}{\epsilon} + u_1^- + \cdots,$$

and u_- at $x = x_w$ is prescribed independent of ϵ. Hence at the boundary $x = x_w$, i.e., $\xi = 0$, we must have $u_0^-(0,\tau) = 0$, and thus at the boundary $\xi = 0$ we have

$$\frac{1}{p}\frac{\partial^2\varphi_0}{\partial\xi\partial\tau} = f(\tau).$$

Integrating with respect to τ, we find

$$\frac{\partial\varphi_0}{\partial\xi} = F(\tau) \quad \text{at} \quad \xi = 0, \tag{2.44}$$

where

$$F(\tau) = p\int_0^\tau f(\tau)d\tau + \frac{\partial\varphi_0}{\partial\xi}(0,0).$$

(Notice that $\frac{\partial\varphi_0}{\partial\xi}(0,0)$ is the prescribed initial (rescaled) electric field evaluated at the boundary $\xi = 0$.) *We will assume* $F(\tau) > 0$ *on* $[0,1]$.

As noted above, matching the potential in the sheath with the potential determined by the middle region requires

$$\varphi_0 \to 0 \quad \text{as} \quad \xi \to -\infty. \tag{2.45}$$

Matching $\frac{\partial\varphi}{\partial\xi}$ requires revisiting the expansion for the potential φ in the middle region.

Recall in the middle region

$$\varphi(\xi,\tau) = \delta^{1/2}\varphi_1(\zeta) + \cdots.$$

Hence

$$\frac{\partial\varphi}{\partial\xi}(\xi,\tau) = \delta^{1/2}\frac{d\varphi_1}{d\zeta}(\xi)\frac{d\zeta}{d\xi} + \cdots,$$

and since $\zeta = \epsilon^{1/5}\xi + \epsilon^{-4/5}x_w$, $\frac{d\zeta}{d\xi} = \epsilon^{1/5}$ and

$$\frac{\partial\varphi}{\partial\xi}(\xi,\tau) = \epsilon^{3/5}\frac{d\varphi_1}{d\zeta}(\zeta) + \cdots.$$

Hence again for a typical transition layer (say, $x = -\epsilon^\alpha$, $\frac{4}{5} < \alpha < 1$), $\xi \to -\infty$, $\zeta \to x_{w_1}$ as $\epsilon \to 0$, and

$$\frac{\partial\varphi}{\partial\xi} \to 0 \quad \text{as} \quad \zeta \to x_{w_1}.$$

Thus the matching condition for $\frac{\partial\varphi_0}{\partial\xi}$ is

$$\frac{\partial\varphi_0}{\partial\xi} \to 0 \quad \text{as} \quad \xi \to -\infty. \tag{2.46}$$

In summary, $\varphi_0(\xi,\tau)$ will satisfy

$$\frac{\partial^2\varphi_0}{\partial\xi^2} = (1 + 2\overline{\varphi}_0)^{-1/2} - e^{-\varphi_0}, \quad -\infty < \xi < 0, \tag{2.47}$$

$$\frac{\partial\varphi_0}{\partial\xi} = F(\tau) \quad \text{at} \quad \xi = 0, \tag{2.48}$$

$$\varphi_0, \frac{\partial\varphi_0}{\partial\xi} \to 0, 0 \quad \text{as} \quad \xi \to -\infty. \tag{2.49}$$

There is still one more step in matching the middle and sheath regions, i.e., the determination of x_{w_1}. Recall that from the middle region

$$\varphi(\zeta) = \epsilon^{2/5}\varphi_1(\zeta) + \cdots,$$

where φ_1 satisfies the Painlevé 1 equation (2.33), which classically [9] satisfies the formula

$$\varphi_1(\zeta) = \frac{6}{(\zeta - \zeta_0)^2}(1 + \text{h.o.t.} \quad \text{in} \ \ (\zeta - \zeta_0))$$

near ζ_0. Here $\zeta_0 > 0$ is the left-most pole of the unique monotone increasing first Painlevé transient. On the other hand, the sheath solution satisfies (2.47). As we shall see in subsection D, $\varphi_0 - \overline{\varphi}_0 \to 0$ exponentially as $\xi \to -\infty$, and hence (2.47) implies

$$\frac{d^2\varphi_0}{d\xi^2} \sim \varphi_0^2 \quad \text{as} \quad \xi \to -\infty$$

and hence

$$\varphi_0(\xi) \sim \frac{6}{\xi^2} \quad \text{as} \quad \xi \to -\infty.$$

Thus in a matching region, φ has representations $\varphi(\zeta) \sim \frac{6\epsilon^{2/5}}{(\zeta-\zeta_0)^2}$ from the middle region, $\varphi(\xi) \sim \frac{6}{\xi^2}$ from the sheath region. However, since $\zeta = \frac{x-x_B}{\epsilon^{4/5}}, \xi = \frac{x-x_w}{\epsilon}$, we see $\zeta = \epsilon^{-4/5}(x_w - x_B) + \epsilon^{1/5}\xi$, and substitution into the above expression for φ in the middle region shows that equality of the two representations of φ occurs when

$$x_w = x_B + \epsilon^{4/5}\zeta_0 + \cdots, \quad \text{i.e.,} \quad x_{w_1} = \zeta_0.$$

D. Higher order sheath asymptotics. From subsection C we know that n_0, u_0, n_1 are independent of τ, and balancing terms of order ϵ yields

$$\frac{d}{d\xi}(n_1u_0 + n_0\overline{u}_1) = \overline{ze^{-\varphi_0}}, \tag{2.50}$$

$$\frac{d}{d\xi}(u_0\overline{u}_1) = \frac{d\overline{\varphi}_1}{d\xi} - u_0^2 - \frac{zu_0}{n_0}\overline{e^{-\varphi_0}}, \tag{2.51}$$

$$\frac{\partial^2\varphi_1}{\partial\xi^2} = n_1 + e^{-\varphi_0}\varphi_1. \tag{2.52}$$

It is possible to eliminate $n_1, \overline{u}_1$ from (2.50)–(2.52) to find a single higher order equation for φ_1. However, the goal here to compute an asymptotic relation for φ_1 as $\xi \to -\infty$. The computation is in fact elementary.

Equation (2.50) implies

$$n_1u_0 + n_0\overline{u}_1 \sim z\xi + c_1(\tau) \quad \text{as} \quad \xi \to -\infty. \tag{2.53}$$

Equation (2.51) implies

$$u_0\overline{u}_1 - \overline{\varphi}_1 \sim -(1+z)\xi + c_2 \quad \text{as} \quad \xi \to -\infty. \tag{2.54}$$

Thus if we multiply (2.53) by u_0, (2.54) by n_0, and subtract the resultant equations, we find

$$n_1 u_0^2 + n_0 \overline{\varphi}_1 \sim u_0(z\xi + c_1(\tau)) + n_0((1+z)\xi - c_2) \quad \text{as} \quad \xi \to -\infty, \tag{2.55}$$

and hence

$$n_1 \sim -\frac{n_0}{u_0^2}\overline{\varphi}_1 + \frac{1}{u_0}(z\xi + c_1(\tau)) + \frac{n_0}{u_0^2}((1+z)\xi - c_2) \quad \text{as} \quad \xi \to -\infty. \tag{2.56}$$

Since $n_0 u_0 = 1$, $u_0^2 = 1 + 2\overline{\varphi}_0$, and $n_0(\xi) \to 1$, $u_0(\xi) \to 1$ as $\xi \to -\infty$, (2.56) can be simplified by writing

$$n_1 \sim -(1 + 2\overline{\varphi}_0)^{-3/2}\overline{\varphi}_1 + (2z+1)\xi \quad \text{as} \quad \xi \to -\infty. \tag{2.57}$$

Inserting (2.57) into (2.52), we see

$$\frac{\partial^2 \varphi_1}{\partial \xi^2} \sim (e^{-\varphi_0} - (1 + 2\overline{\varphi}_0)^{-3/2})\overline{\varphi}_1 + (2z+1)\xi \quad \text{as} \quad \xi \to -\infty. \tag{2.58}$$

Now average (2.58) to see that $\overline{\varphi}_1$ will have asymptotic behavior as $\xi \to -\infty$ given by solutions $y(\xi)$ of the equation

$$\frac{d^2 y}{d\xi^2} = \big(\overline{e^{-\varphi_0}} - (1 + 2\overline{\varphi}_0)^{-3/2}\big)y + (2z+1)\xi. \tag{2.59}$$

Next note that $\varphi_0 - \overline{\varphi}_0$ satisfies the equation

$$\frac{\partial^2}{\partial \xi^2}(\varphi_0 - \overline{\varphi}_0) = -e^{-\varphi_0} + \overline{e^{-\varphi_0}} = (\varphi_0 - \overline{\varphi}_0) + \cdots,$$

and hence

$$\varphi_0 - \overline{\varphi}_0 \sim a_1 e^{\xi} + a_2 e^{-\xi} \quad \text{as} \quad \xi \to -\infty.$$

Since $\varphi_0, \overline{\varphi}_0$ are bounded as $\xi \to -\infty$, we must have $a_1 = 0$, $\overline{\varphi}_0 = \varphi_0$ plus exponentially small terms as $\xi \to -\infty$. Hence in (2.47) we may replace $\overline{\varphi}_0$ by φ_0 as $\xi \to -\infty$ to see

$$\frac{\partial^2 \varphi_0}{\partial \xi^2} \sim (1 + 2\varphi_0)^{-1/2} - e^{-\varphi_0} \sim \varphi_0^2 \quad \text{as} \quad \xi \to -\infty$$

and hence

$$\varphi_0 \sim \frac{6}{\xi^2} \quad \text{as} \quad \xi \to -\infty. \tag{2.60}$$

Substitution of (2.60) into (2.59) shows that (2.59) is asymptotically equivalent to

$$\frac{d^2 y}{d\xi^2} = \frac{12}{\xi^2} y + (2z+1)\xi, \tag{2.61}$$

which has the explicit solution

$$y(\xi) = b_1 \xi^4 + b_2 \xi^{-3} - \frac{(2z+1)}{6}\xi^3.$$

We enforce minimal growth as $\xi \to -\infty$ and set $b_1 = 0$ so that

$$\overline{\varphi}_1 \sim \frac{2z+1}{6}\xi^3 \quad \text{as} \quad \xi \to -\infty.$$

Recall that our expansion for the potential φ in the sheath was written as

$$\varphi(\xi,\tau) = \varphi_0(\xi,\tau) + \epsilon\varphi_1(\xi,\tau) + \cdots,$$

and hence

$$\overline{\varphi}(\xi) = \overline{\varphi_0(\xi)} + \epsilon\overline{\varphi_1(\xi)} + \cdots.$$

Thus on a matching overlap domain between the middle region and the sheath domain, say,

$$x - x_w = -\epsilon^\beta,$$

$\frac{4}{5} < \beta < 1$, we see $\xi = \frac{x-x_w}{\epsilon} = -\epsilon^{(\beta-1)}$ and

$$\overline{\varphi}_0(\xi) + \epsilon\overline{\varphi}_1(\xi) \sim -\frac{(2z+1)}{6}\epsilon^{3\beta-2} \quad \text{as} \quad \epsilon \to 0$$
$$\to 0 \quad \text{since} \quad 3\beta - 2 > \frac{2}{5}.$$

Analysis of (2.52) shows that

$$\frac{\partial^2}{\partial\xi}(\varphi_1 - \overline{\varphi}_1) \sim (\varphi_1 - \overline{\varphi}_1),$$

and hence φ_1 and $\overline{\varphi}_1$ are exponentially close as $\xi \to -\infty$. Thus on a matching regime $\varphi_0 + \epsilon\varphi_1$ may be matched to the middle solution.

On the other hand, in the absence of a middle regime, an attempt to directly match plasma and sheath fails. For example, if $x - x_w = -\epsilon^\beta$, $0 < \beta < \frac{2}{3}$, we have

$$\overline{\varphi_0(\xi)} + \epsilon\overline{\varphi_1(\xi)} \sim -\frac{(2z+1)}{6}\epsilon^{3\beta-2} \to -\infty \quad \text{as} \quad \epsilon \to 0,$$

which does not match the limit from the quasi-neutral plasma

$$\lim_{x \to x_{B^-}} \varphi(x) \sim \lim_{x \to x_B^-} -(x - x_B)^{1/2} = 0.$$

We note that a similar argument is given in the paper of Franklin and Ockendon [5] in the case when the rf-current and friction are not considered.

3. Existence of solutions for the "exact" sheath system. In this section we will prove existence of solutions to the exact sheath system (1.1) ((2.47)–(2.49) of subsection C). It is rewritten here with the subscript zero deleted, i.e., as

$$\frac{\partial^2\varphi}{\partial\xi^2} = (1 + 2\overline{\varphi})^{-1/2} - e^{-\varphi}, \quad -\infty < \xi < 0, \tag{3.1}$$

$$\frac{\partial\varphi}{\partial\xi} = F(\tau) \quad \text{at} \quad \xi = 0, \tag{3.2}$$

$$\varphi, \frac{\partial\varphi}{\partial\xi} \to 0, 0 \quad \text{as} \quad \xi \to -\infty. \tag{3.3}$$

Again recall that F is a given positive C^1 periodic function of τ with period 1, and the overbar denotes the τ average over interval $0 \le \tau \le 1$.

Step 1. *Regularization.* Define

$$g(\varphi) = \begin{cases} e^{-\varphi}, & \varphi \ge 0, \\ 1 - \varphi, & \varphi \le 0, \end{cases}$$

so that $g(\varphi)$ is $C^1(\mathbb{R})$, convex, and $|g'(\varphi)| \le 1$. It is the boundedness of $g'(\varphi)$ that proves useful.

We now consider the regularized problem

$$\frac{\partial^2 \varphi}{\partial \xi^2} = (1 + 2\overline{\varphi})^{-1/2} + g'(\varphi) + \mu\varphi, \quad -L < \xi < 0, \tag{3.4}$$

$$\frac{\partial \varphi}{\partial \xi} = F(\tau) \quad \text{at} \quad \xi = 0, \quad 0 \le \tau \le 1, \tag{3.5}$$

$$\varphi = 0 \quad \text{at} \quad \xi = -L, \quad 0 \le \tau \le 1, \tag{3.6}$$

where $L = \mu^{-1}, 0 < \mu \le 1$.

We shall prove existence of solutions to (3.4)–(3.6) via the direct method of the calculus of variations [10].

Set

$$J(\varphi) = \int_0^1 \int_{-L}^0 \left\{ \frac{1}{2} \left(\frac{\partial \varphi}{\partial \xi} \right)^2 - \frac{\partial \varphi}{\partial \xi} F(\tau) + (1 + 2\overline{\varphi})^{1/2} + g(\varphi) + \frac{\mu \varphi^2}{2} \right\} d\xi d\tau. \tag{3.7}$$

Let

$$\tilde{H}^1(-L, 0) = \{\varphi \in H^1(-L, 0); \ \varphi = 0 \quad \text{at} \quad \xi = -L\},$$

where $H^1(-L, 0) \subset C[-L, 0]$ denotes the usual Sobolev space of square integrable functions with square integrable generalized derivatives on $(-L, 0)$ endowed with inner product

$$(\varphi, \psi)_{H^1(-L,0)} = \int_{-L}^0 \{\varphi\psi + \varphi_\xi \psi_\xi\} d\xi.$$

Of course $\tilde{H}^1$ inherits the H^1 inner product. Let $\mathbb{H} = L^2((0, 1); \tilde{H}^1(-L, 0))$ so that $\mathbb{H}$ is a Hilbert space endowed with inner product

$$(\varphi, \psi)_{\mathbb{H}} = \int_0^1 \int_{-L}^0 \{\varphi\psi + \varphi_\xi \psi_\xi\} d\xi d\tau.$$

For convenience we recall Jensen's inequality [10] within the context we will use here: Let $f : \mathbb{R} \to \mathbb{R}$ be convex and $u : [0, 1] \to \mathbb{R}$ be summable; then

$$f(\overline{u}) \le \overline{f(u)}. \tag{3.8}$$

Notice that Jensen's inequality implies

$$\|\varphi\|_{\mathbb{H}} \ge \|\overline{\varphi}\|_{\tilde{H}^1}, \tag{3.9}$$

and hence $\varphi \in \mathbb{H}$ implies $\overline{\varphi} \in C$ via the usual Sobolev embedding theorem [10]. Hence the admissible set

$$A \stackrel{\text{def}}{=} \{\varphi \in \mathbb{H}; \overline{\varphi} \ge 0 \quad \text{on} \quad [-L, 0]\} \tag{3.10}$$

is a well defined closed subset of $\mathbb{H}$.

Application of the direct method requires three basic estimates [10]:

(i) The functional to be minimized is bounded from below in the class of admissible functions, so that the infimum and therefore a minimizing sequence exists.

(ii) The functional is weakly lower semicontinuous with respect to weak convergence in the class of admissible functions.

(iii) The minimizing sequence possesses a weakly convergent subsequence, which converges to an admissible function.

If these estimates can be satisfied, existence of a minimizer can be ascertained.

We will now check (i)–(iii) for our functional $J(\varphi)$ given by (3.7). For $\varphi \in A$ note

$$J(\varphi) \geq \frac{\mu}{2}\|\varphi\|_{\mathbb{H}}^2 - \max|F(\tau)|\mu^{-1/2}\|\varphi\|_{\mathbb{H}},$$

and hence $J(\varphi)$ is bounded from below for $\varphi \in \mathbb{H}$. Thus a minimizing sequence $\{\varphi^{(n)}\}$ exists so that $m = \inf_{\varphi\in A} J(\varphi)$, $J(\varphi^{(n)}) \to m$ as $n \to \infty$, and $\|\varphi^{(n)}\|_{\mathbb{H}}$ is bounded for all $n \geq 1$. Furthermore, by (3.9), const. $\geq \|\overline{\varphi}^{(n)}\|_{\tilde{H}^1}$ for all n. Hence there is a subsequence of $\{\varphi^{(n)}\}$ also denoted by $\{\varphi^{(n)}\}$ and $\varphi \in \mathbb{H}$ so that $\varphi^{(n)} \to \varphi$ weakly in $\mathbb{H}$ and $\overline{\varphi}^{(n)} \to w$ strongly in $L^2(-L,0)$. In fact $w = \overline{\varphi}$. To see this, let $\theta \in L^2(-L,0)$. By weak convergence of $\{\varphi^{(n)}\}$ in $\mathbb{H}$,

$$\int_{-L}^{0} \theta(\xi) \int_0^1 (\varphi^{(n)} - \varphi) d\tau d\xi \to 0 \quad \text{as} \quad n \to \infty,$$

and hence

$$\int_{-L}^{0} \theta(\xi)(\overline{\varphi}^{(n)} - \overline{\varphi}) d\xi \to 0 \quad \text{as} \quad n \to \infty,$$

and so $w = \overline{\varphi}$.

Finally, by the usual Sobolev embedding theorem, $\overline{\varphi}^{(n)} \to \overline{\varphi}$ uniformly on $[-L,0]$, $\overline{\varphi}(-L) = 0$, $\overline{\varphi}$ is continuous in $[-L,0]$, $\overline{\varphi} \geq 0$. We summarize the above information in the following lemma.

LEMMA 3.1. *$J(\cdot)$ has a minimizing sequence $\{\varphi^{(n)}\} \subset \mathbb{H}$, $\varphi^{(n)} \to \varphi$ weakly in $\mathbb{H}$, $\overline{\varphi}^{(n)} \to \overline{\varphi}$ strongly in $L^2(-L,0)$ and uniformly on $[-L,0]$, $\overline{\varphi}$ is continuous nonnegative with $\overline{\varphi}(-L) = 0$ and $\varphi \in A$.*

The first and third estimates (i), (iii) in our agenda for applying the direct method are now complete, and we move on to (ii), i.e., lower semicontinuity.

LEMMA 3.2. *Let $\{\varphi^{(n)}\}$ be the convergent minimizing sequence of $J(\cdot)$ given in Lemma* 3.1. *Then*

$$\liminf_{n\to\infty} J(\varphi^{(n)}) \geq J(\varphi).$$

Proof. We will examine pieces of $J(\varphi^{(n)})$ separately. First note that the uniform convergence of $\{\overline{\varphi}^{(n)}\}$ on $[-L,0]$ implies

$$\int_{-L}^{\infty} (1 + 2\overline{\varphi}^{(n)})^{1/2} d\xi \to \int_{-L}^{0} (1 + 2\overline{\varphi})^{1/2} d\xi \quad \text{as} \quad n \to \infty.$$

Next note that since g is convex $\int_0^1 \int_{-L}^0 g(\varphi_n) - g(\varphi) d\xi d\tau \geq \int_0^1 \int_{-L}^0 g'(\varphi)(\varphi_n - \varphi) d\xi d\tau$. Since $|g'(\varphi)| \leq 1$ we have $g' \in L^2((-L,0)\times(0,1))$, and via weak convergence $\lim_{n\to\infty} \int_0^1 \int_{-L}^0 g(\varphi_n) d\xi d\tau \geq \int_0^1 \int_{-L}^0 g(\varphi) d\xi d\tau$. Thus

$$\int_0^1 \int_{-L}^0 \left\{ \frac{1}{2}\left(\frac{\partial\varphi}{\partial\xi}\right)^2 + \frac{\mu}{2}\varphi^2 + g(\varphi) - \frac{\partial\varphi}{\partial\xi} F(\tau) \right\} d\xi d\tau$$

is weakly lower semicontinuous on $\mathbb{H}$. Now put the two pieces of $J(\cdot)$ together, and the lemma is proven. (Notice that if we had used the original function $e^{-\varphi}$ and not g we would not know a priori that $e^{-\varphi}$ was in $L^2((-L,0)\times(0,1))$.) □

Now all three elements of the application of the direct method have been satisfied, and we can assert the following theorem.

THEOREM 3.3. *$J(\cdot)$ has a minimizer $\varphi \in A$, and φ satisfies the weak form of the Euler–Lagrange equation*

$$\int_0^1 \int_{-L}^0 \left\{ \frac{\partial \varphi}{\partial \xi}\frac{\partial \psi}{\partial \xi} - \frac{\partial \psi}{\partial \xi} F(\tau) + (1+2\overline{\varphi})^{-1/2}\overline{\psi} + g'(\varphi)\psi - \mu\varphi\psi \right\} d\xi d\tau = 0 \tag{3.11}$$

for all $\psi \in A$.

Proof. We know that $m \geq J(\varphi)$, and hence $J(\varphi) = m$ and φ is a minimizer. The derivation of the weak form of the Euler–Lagrange equation is classical.

Now set $\psi(s,\xi) = \chi_{[0,\tau]}\psi_1(\xi)$ in (3.11), where

$$\chi_{[0,\tau]}(s) = \begin{cases} 1, & 0 \leq s \leq \tau, \\ 0, & \tau < s < 1, \end{cases}$$

and $\psi_1 \in \tilde{H}^1(-L,0), \psi_1 \geq 0$. Then $\psi \in A$, and the weak form of the Euler–Lagrange equation can be written

$$\int_0^\tau h(\tau)d\tau = 0, \quad 0 \leq \tau \leq 1, \tag{3.12}$$

where

$$h(\tau) \stackrel{\text{def}}{=} \int_{-L}^0 \left(\frac{\partial \varphi}{\partial \xi} - F(\tau) \right) \psi_1'(\xi) + ((1+2\overline{\varphi})^{-1/2} + g'(\varphi) + \mu\varphi)\psi_1(\xi)d\xi.$$

It is a straightforward application of the Cauchy–Schwarz inequality to see

$$|h(\tau)| \leq \text{ const.} \left(\left\| \frac{\partial \varphi}{\partial \xi} \right\|_{L^2(-L,0)} + \mu^{-1}|F(\tau)| + 2\mu^{-1} + \mu\|\varphi\|_{L^2(-L,0)} \right)$$

and hence $h \in L^1(0,1)$. Thus (3.12) holds for all $\tau \in [0,1]$ and $h \in L^1(0,1)$, so $h(\tau) = 0$ a.e. in $[0,1]$. □

We summarize in the following lemma.

LEMMA 3.4.

$$\int_{-L}^0 \left\{ \left(\frac{\partial \varphi}{\partial \xi} - F(\tau) \right) \psi_1'(\xi) + ((1+2\overline{\varphi})^{-1/2} + g'(\varphi) + \mu\varphi)\psi_1(\xi) \right\} d\xi = 0 \tag{3.13}$$

for all $\psi_1 \in \tilde{H}^1(-L,0), \psi_1 \geq 0$, for all $\tau \in S \subseteq [0,1]$, meas $S = 1$; i.e., φ is a weak solution of our regularized problem (3.4)–(3.6) *for τ a.e. in* $[0,1]$.

LEMMA 3.5. *There is a set $S_1 \subset S$, meas $S_1 = 1$, so that φ is a classical $C^2[-L,0]$ solution of* (3.4)–(3.6) *for all $\tau \in S_1$.*

Proof. First $\varphi(\tau,\cdot) \in \tilde{H}^1(-L,0)$ for $\tau \in S_1 \subset S$, where meas $S_1 = 1$. For if meas $S_1 < 1$, then set $S_2 = [0,1]/S_1$, where meas $S_2 > 0$, and the inequality

$$\int_{S_2} \|\varphi(\tau,\cdot)\|^2_{\tilde{H}^1(-L,0)} d\tau \leq \int_0^1 \|\varphi(\tau,\cdot)\|^2_{\tilde{H}^1(-L,0)} = \|\varphi\|^2_{\mathbb{H}}$$

will be violated. Thus $\varphi(\tau,\xi)$ is both a weak solution for (3.4)–(3.6) and a continuous function of ξ on $[-L,0]$ for all $\tau \in S_1$, $S_1 \subset [0,1]$, meas $S_1 = 1$. Of course $\overline{\varphi} \in \tilde{H}^1(-L,0)$. Hence the right-hand side of (3.4) is continuous in ξ for $\tau \in S_1$, and hence φ must be in $C^2[-L,0]$, and φ is a classical $C^2[-L,0]$ solution of (3.4)–(3.6) for $\tau \in S_1$. ∎

LEMMA 3.6. *Assume $F(\tau) > 0$ for all $\tau \in [0,1]$; then $\varphi \geq 0$ for $-L \leq \xi \leq 0$ and almost all $\tau \in [0,1]$.*

Proof. Since φ is a classical solution of (3.4)–(3.6) for $\tau_1 \in S_1$, if $\varphi < 0$ at some $\xi \in (-L,0]$, then boundary conditions $\varphi(-L,\tau) = 0$, $\frac{\partial\varphi}{\partial\xi}(0,\tau) > 0$, imply that φ must have a negative minimum at ξ^* for this value of τ_1. Hence $\overline{\varphi}(\xi^*) > 0$, $\frac{\partial^2\varphi}{\partial\xi^2}(\xi^*,\tau_1) \geq 0$, and (3.4) implies

$$0 \leq (1 + 2\overline{\varphi}(\xi^*))^{-1/2} + g'(\varphi(\xi^*,\tau_1)) + \mu\varphi(\xi^*,\tau_1).$$

But $g(\varphi) = 1 - \varphi$ when $\varphi < 0$, and hence $g'(\varphi(\xi^*,\tau_1)) = -1$ and

$$0 \leq (1 + 2\overline{\varphi}(\xi^*))^{-1/2} - 1 + \mu\varphi(\xi^*,\tau_1) \leq \mu\varphi(\xi^*,\tau_1),$$

which is a contradiction. ∎

THEOREM 3.7. *Assume $F(\tau) > 0$ for all $\tau \in [0,1]$. There is a classical nonnegative solution φ of the regularized system*

$$\frac{\partial^2\varphi}{\partial\xi^2} = (1 + 2\overline{\varphi})^{-1/2} - e^{-\varphi} + \mu\varphi, \tag{3.14}$$

$$\frac{\partial\varphi}{\partial\xi} = F(\tau) \quad at \quad \xi = 0, \tag{3.15}$$

$$\varphi = 0 \quad at \quad \xi = -L, \tag{3.16}$$

which is continuous in τ and twice continuously differentiable in ξ on $[0,1]\times[-L,0]$.

Proof. By Lemma 3.6 we know $\varphi \geq 0$ and hence $g'(\varphi) = -e^{-\varphi}$. Hence all that remains to be shown is continuity in τ. Let $\tau_1, \tau_2 \in S_1$ so that we know that $\varphi(\cdot,\tau_1), \varphi(\cdot,\tau_2)$ are solutions of (3.14)–(3.16). Set $w(\xi) = \varphi(\xi,\tau_1) - \varphi(\xi,\tau_2)$. Then from (3.14) we know

$$w''(\xi) = -e^{-\varphi(\xi,\tau_1)} + e^{-\varphi(\xi,\tau_2)} + \mu w(\xi),$$

and, by the mean-value theorem,

$$w''(\xi) = e^{\hat{\varphi}(\xi)} w(\xi) + \mu w(\xi), \tag{3.17}$$

where $0 < \varphi(\xi,\tau_1) \leq \hat{\varphi}(\xi) \leq \varphi(\xi,\tau_2)$ (or vice versa with τ_1, τ_2 interchanged). Thus, by integration by parts,

$$\begin{aligned}\int_{-L}^{0} w(\xi)w''(\xi)d\xi &= w(\xi)w'(\xi)\,|_{\xi=-L}^{\xi=0} - \int_{-L}^{0} w'(\xi)^2 d\xi \\ &= w(0)(F(\tau_1) - F(\tau_2)) - \int_{-L}^{0} w'(\xi)^2 d\xi, \end{aligned} \tag{3.18}$$

and by (3.17),

$$\int_{-L}^{0} (w'(\xi)^2 + e^{\hat{\varphi}(\xi)} w^2(\xi) + \mu^2 w^2(\xi))d\xi \leq |w(0)||F(\tau_1) - F(\tau_2)|.$$

By the embedding of the Sobolev space $H^1(-L,0)$ into $C[-L,0]$, we have

$$\sup_{0\le\xi\le L}|w(\xi)|^2 \le \text{ const. } \sup_{0\le\xi\le L}|w(\xi)||F(\tau_1)-F(\tau_2)|$$

and

$$\sup_{\xi}|w(\xi)| \le |F(\tau_1)-F(\tau_2)|. \tag{3.19}$$

Hence for $\tau \notin S_1$ let $\{\tau_n\} \subset S_1$, $\tau_n \to \tau$. Then by (3.19)

$$\sup_{\xi}|\varphi(\xi,\tau_n)-\varphi(\xi,\tau_m)| \le |F(\tau_n)-F(\tau_m)| \to 0 \quad \text{as} \quad m,n\to\infty.$$

Thus $\varphi(\cdot,\tau_n)$ is a Cauchy sequence in $C[-L,0]$ and converges to an element of $C[-L,0]$ which we call $\varphi(\xi,\tau)$. By construction, $\varphi(\xi,\cdot)$ is continuous at τ. Finally, represent the solution of (3.14) as integrals (simply by integrating (3.14) with respect to ξ twice); then passing to the same limit shows that $\varphi(\xi,\tau)$ is a classical solution to (3.14)–(3.16). □

Thus Step 1 is completed: A classical solution of the regularized problem (3.14)–(3.16) has been obtained. We now proceed to Step 2, passage to the limit as $\mu \to 0+$ to obtain a classical solution of (3.1)–(3.3).

Step 2. *Passage to the limit as* $\mu \to 0$.

LEMMA 3.8. *If* $\varphi(\cdot,\tau)$ *has a local maximum at* $\xi^* \in (-L,0)$, *then* $\overline{\varphi}(\xi^*) \ge \varphi(\xi^*,\tau)$.

Proof. If φ has a local maximum at ξ^*, then $\frac{\partial^2\varphi}{\partial\xi^2}(\xi^*,\tau) \le 0$ and, from (3.14), $0 \ge (1+2\overline{\varphi}(\xi^*))^{-1/2} - e^{-\varphi(\xi^*,\tau)} + \mu\varphi(\xi^*,\tau)$. Since φ is nonnegative, $e^{-\varphi(\xi^*,\tau)} \ge (1+2\overline{\varphi}(\xi^*))^{-1/2}$, and hence $1+2\overline{\varphi}(\xi^*) \ge e^{2\varphi(\xi^*,\tau)} = 1+2\varphi(\xi^*,\tau) + \text{positive terms}$, and the lemma is proven. □

LEMMA 3.9. *Define* $v(\xi,\tau) = \varphi(\xi,\tau) - \overline{\varphi}(\xi)$. *If* $v(\cdot,\tau)$ *has a local maximum at* $\xi^* \in (-L,0)$, *then* $v(\xi^*,\tau) \le 0$, *i.e.,* $\varphi(\xi^*,\tau) \le \overline{\varphi}(\xi^*)$.

Proof. From (3.14),

$$\frac{\partial^2\varphi}{\partial\xi^2} = (1+2\overline{\varphi})^{-1/2} - e^{-\varphi} + \mu\varphi, \qquad \frac{\partial^2\overline{\varphi}}{\partial\xi^2} = (1+2\overline{\varphi})^{-1/2} - \overline{e^{-\varphi}} + \mu\overline{\varphi},$$

and hence by Jensen's inequality

$$\frac{\partial^2 v}{\partial\xi^2} = -e^{-\varphi} + \overline{e^{-\varphi}} + \mu v \ge -e^{-\varphi} + e^{-\overline{\varphi}} + \mu v.$$

This in turn implies

$$\frac{\partial^2 v}{\partial\xi^2} \ge e^{-\overline{\varphi}}(1-e^{-v}) + \mu v.$$

Now if v has a local maximum at $\xi^* \in (-L,0)$, then $\frac{\partial^2 v}{\partial\xi^2}(\xi^*,\tau) \le 0$ and $0 \ge e^{-\overline{\varphi}}(1-e^{-v}) + \mu v$. Since $v(\xi^*,\tau) > 0$ would yield a contradiction, we must have $v(\xi^*,\tau) \le 0$. □

LEMMA 3.10. (i) *If* $\varphi(0,\tau) - \overline{\varphi}(0) > 0$, *then the graph of* $\varphi(\cdot,\tau)$ *can intersect the graph of* $\overline{\varphi}$ *at most once on* $(-L,0)$.

(ii) *If* $\varphi(0,\tau) - \overline{\varphi}(0) \le 0$, *then the graph of* $\varphi(\cdot,\tau)$ *can never intersect the graph of* $\overline{\varphi}$ *on* $(-L,0)$.

Proof. (i) In this case, $v(0,\tau) = \varphi(0,\tau) - \overline{\varphi}(0) > 0$, $v(-L,\tau) = \varphi(-L,\tau) - \overline{\varphi}(-L) = 0$. If $\varphi(\cdot,\tau)$ intersects $\overline{\varphi}$ twice on $(-L,0)$, then $v(\cdot,\tau)$ has a positive local maximum in $(-L,0)$, which contradicts Lemma 3.9.

(ii) In this case, $v(0,\tau) = \varphi(0,\tau) - \overline{\varphi}(0) \le 0$, $v(-L,\tau) = 0$, and again if $\varphi(\cdot,\tau)$ intersects $\overline{\varphi}$, then $v(\cdot,\tau)$ has a positive local maximum in $(-L,0)$, which again contradicts Lemma 3.9. ∎

From Lemma 3.10 we see that there are three possible cases for behavior of the graph $\varphi(\cdot,\tau)$:

(a) $\varphi(0,\tau) > \overline{\varphi}(0)$, and the graph of $\varphi(\cdot,\tau)$ is always above the graph of $\overline{\varphi}$ on $(-L,0)$.

(b) $\varphi(0,\tau) > \overline{\varphi}(0)$, and the graph of $\varphi(\cdot,\tau)$ intersects the graph of $\overline{\varphi}$ once on $(-L,0)$.

(c) $\varphi(0,\tau) \le \overline{\varphi}(0)$, and the graph of $\varphi(\cdot,\tau)$ is always below the graph of $\overline{\varphi}$ on $(-L,0)$.

LEMMA 3.11. *In case* (a), $\varphi(\cdot,\tau)$ *is monotone increasing in* $[-L,0]$.

Proof. Since $\frac{\partial\varphi}{\partial\xi}(0,\tau) = F(\tau) > 0$, φ is monotone increasing near $\xi = 0$. Thus for φ to lose monotonicity it would have to possess a local minimum, say at $\xi_1 < 0$, where $\varphi(\xi_1,\tau) > 0$. Since $\varphi(-L,\tau) = 0$, there would have been a local maximum of $\xi_2 \in (-L,\xi_1)$. By Lemma 3.8, $\overline{\varphi}(\xi_2) \ge \varphi(\xi_2,\tau)$, which is a contradiction, and the lemma is proven. ∎

LEMMA 3.12. φ *is bounded on* $[-L,0]\times[0,1]$, $L = \mu^{-1}$, *uniformly in* τ,ξ, *and* μ.

Proof. If case (a) occurs for some value of $\tau \in [0,\tau]$, then $\varphi(\xi,\tau) > \overline{\varphi}(\xi)$ on $(-L,0)$, and hence $(1+2\overline{\varphi}(\xi))^{-1/2} > (1+2\varphi(\xi,\tau))^{-1/2}$. Hence (3.14) implies

$$\frac{\partial^2\varphi}{\partial\xi^2} > (1+2\varphi)^{-1/2} - e^{-\varphi} + \mu\varphi.$$

Multiplication by $\frac{\partial\varphi}{\partial\xi}$ (which is nonnegative, by Lemma 3.11) and integration from $\xi = -L$ to $\xi = 0$ yields

$$\left[\frac{1}{2}\left(\frac{\partial\varphi}{\partial\xi}\right)^2 - (1+2\varphi)^{1/2} - e^{-\varphi} - \frac{\mu\varphi^2}{2}\right]_{\xi=-L}^{\xi=0} \ge 0.$$

Evaluation of φ at $\xi = 0$, $\xi = -L$ and use of (3.15) imply

$$\frac{F^2(\tau)}{2} \ge (1+2\varphi(0,\tau))^{1/2} + e^{-\varphi(0,\tau)} + \mu\frac{\varphi^2}{2}(0,\tau) - 2,$$

which gives

$$2 + \frac{F^2(\tau)}{2} > (1+2\varphi(0,\tau))^{1/2}.$$

Thus in case (a), $\varphi(0,\tau)$ is bounded independently of μ, and by monotonicity of $\varphi(\xi,\tau)$ in ξ, so is $\varphi(\cdot,\tau)$. Moreover, if case (a) occurs for all $\tau \in [0,1]$, the above inequality shows that φ is bounded on $[-L,0], L = \mu^{-1}$, uniformly in τ and μ.

Now note that case (a) must occur for some value of $\tau \in [0,1]$. (If not, then we would have only cases (b), (c) and hence a value of ξ where the τ average of φ is less

than $\overline{\varphi}(\xi)$, which is of course impossible.) Hence, since φ is above $\overline{\varphi}$ in case (a), we have $\overline{\varphi}$ bounded independently of μ, and hence trivially case (c) is now covered; i.e., φ is bounded independently of μ in case (c).

Thus we need consider only case (b). In case (b) denote by ξ_* the crossing value of ξ, i.e., $\varphi(\xi_*,\tau)=\overline{\varphi}(\xi_*)$. (Of course ξ_* will in general depend on τ.) Hence $\varphi(\xi,\tau)>\overline{\varphi}(\xi)$ on $(\xi_*,0]$, $0<\varphi(\xi,\tau)<\overline{\varphi}(\xi)$ on $(-L,\xi_*)$. Thus we need consider only the interval $(\xi_*,0)$ since the boundedness of $\overline{\varphi}$ implies boundedness of $\varphi(\cdot,\tau)$ on $(-L,\xi_*]$.

By Lemma 3.8, $\varphi(\cdot,\tau)$ cannot have a local maximum on $(\xi_*,0)$. Hence $\varphi(\cdot,\tau)$ can be monotone increasing, be monotone decreasing, or have a local minimum. If $\varphi(\cdot,\tau)$ is monotone decreasing, we trivially have an upper bound, i.e., $\overline{\varphi}(\xi_*)$. On the other hand, if $\varphi(\cdot,\tau)$ is monotone increasing in $(\xi_*,0)$, then again

$$\frac{\partial^2\varphi}{\partial\xi^2}>(1+2\varphi)^{-1/2}-e^{-\varphi}+\mu\varphi,$$

and multiplication by $\frac{\partial\varphi}{\partial\xi}>0$ and integration from ξ_* to 0 yields

$$\begin{aligned}\frac{F^2(\tau)}{2}+(1+2\overline{\varphi}(\xi_*))^{1/2}+e^{-\overline{\varphi}(\xi_*)}+\frac{\mu\overline{\varphi}^2(\xi_*)}{2}&\\ >(1+2\varphi(0,\tau))^{1/2}+e^{-\varphi(0,\tau)}+\frac{\mu\varphi(0,\tau)^2}{2}&\\ >(1+2\varphi(0,\tau))^{1/2}.&\end{aligned}$$

Since $\overline{\varphi}$ is bounded independently of μ, we see that $\varphi(0,\tau)$ is bounded independently of μ for all $\tau\in[0,1]$. Thus the monotonicity of φ on $(\xi_*,0)$ yields φ bounded independently of μ on $(\xi_*,0)$ as well as $(-L,\xi_*]$.

Finally, if φ has a minimum on $(\xi_*,0)$, then φ is monotone decreasing on (ξ_*,ξ_{**}) and monotone increasing on $(\xi_{**},0)$, where ξ_{**} is the point of local minimum. On (ξ_*,ξ_{**}), φ is trivially less than $\overline{\varphi}(\xi_*)$, whereas on $(\xi_{**},0)$, φ is monotone increasing, and the same argument given above when φ was monotone increasing on $(\xi_*,0)$ applies. Hence φ is bounded on $(-L,0)$ uniformly in ξ,τ,μ. ∎

LEMMA 3.13. *In case* (b), $0\le\frac{\partial\varphi}{\partial\xi}(\xi,\tau)\le F(\tau)$, $-L\le\xi\le 0$, *and* $\frac{\partial\varphi}{\partial\xi}$ *is bounded uniformly in* ξ,τ,μ.

Proof. In case (b), we know from Lemma 3.11 that $\varphi(\cdot,\tau)$ is monotone, $\varphi>\overline{\varphi}$. Hence

$$\frac{\partial^2\varphi}{\partial\xi^2}=(1+2\overline{\varphi})^{-1/2}-e^{-\varphi}+\mu\varphi\ge(1+2\varphi)^{-1/2}-e^{-\varphi}+\mu\varphi\ge\mu\varphi>0$$

on $(-L,0)$. Hence $\frac{\partial\varphi}{\partial\xi}(\xi,\tau)$ is monotone increasing in ξ, and hence $\frac{\partial\varphi}{\partial\xi}(-L,\tau)\le\frac{\partial\varphi}{\partial\xi}(\xi,\tau)\le F(\tau)$. But $\frac{\partial\varphi}{\partial\xi}(-L,\tau)\ge 0$ since if $\frac{\partial\varphi}{\partial\xi}(-L,\tau)<0$, then the fact that $\varphi(-L,\tau)=0$ implies that φ would take on negative values for $\xi>-L$, which is impossible. Hence $0\le\frac{\partial\varphi}{\partial\xi}(\xi,\tau)\le F(\tau)$ in case (a). ∎

LEMMA 3.14. *In case* (c), $\frac{\partial\varphi}{\partial\xi}$ *is bounded uniformly in* ξ,τ,μ.

Proof. Let τ_1,τ_3 be values of τ so that $\varphi(\tau_1,\xi)$ is in case (a) and $\varphi(\tau_3,\xi)$ is case (c). Let $\psi(\xi)=\varphi(\xi,\tau_1)-\varphi(\xi,\tau_3)$. Then

$$\psi''(\xi)=-e^{-\varphi(\xi,\tau_1)}+e^{-\varphi(\xi,\tau_3)}+\mu(\varphi(\xi,\tau_1)-\varphi(\xi,\tau_3))\ge 0\quad\text{on}\quad(-L,0)$$

since $\varphi(\xi,\tau_1) > \overline{\varphi}(\xi) > \varphi(\xi,\tau_3)$ on $(-L,0)$. Furthermore, $\psi'(-L) \geq 0$ since $\varphi(\xi,\tau_1) \to 0$ as $\xi \to -L$ from values above $\overline{\varphi}$, while $\varphi(\xi,\tau_3) \to 0$ as $\xi \to -L$ from values below $\overline{\varphi}$. Hence $\psi'(\xi) \geq 0$ on $[-L,0]$ and $\frac{\partial\varphi}{\partial\xi}(\xi,\tau_1) \geq \frac{\partial\varphi}{\partial\xi}(\xi,\tau_3)$ on $[-L,0]$, and from Lemma 3.13, $F(\tau_1) \geq \frac{\partial\varphi}{\partial\xi}(\xi,\tau_3)$. Finally, integrate the inequality $\psi''(\xi) \geq 0$ from ξ to 0 to see $\psi'(0) \geq \psi'(\xi)$ and hence $F(\tau_1) - F(\tau_3) \geq \frac{\partial\varphi}{\partial\xi}(\xi,\tau_1) - \frac{\partial\varphi}{\partial\xi}(\xi,\tau_3)$. Since $\frac{\partial\varphi}{\partial\xi}(\xi,\tau_1) \geq 0$, we conclude $F(\tau_1) \geq \frac{\partial\varphi}{\partial\xi}(\xi,\tau_3) \geq F(\tau_3) - F(\tau_1)$ on $[-L,0]$, and the lemma is proven. ∎

LEMMA 3.15. *$\overline{\varphi}'$ is bounded uniformly in μ,ξ.*

Proof. From (3.14) we have

$$\frac{\partial^2\varphi}{\partial\xi^2} = (1+2\overline{\varphi})^{-1/2} + e^{-\varphi} + \mu\varphi \quad \text{on} \quad (-L,0).$$

Multiply by $\frac{\partial\varphi}{\partial\xi}$ and integrate in τ from 0 to 1 to obtain

$$\frac{1}{2}\frac{\partial}{\partial\xi}\overline{\left(\left(\frac{\partial\varphi}{\partial\xi}\right)^2\right)} = (1+2\overline{\varphi})^{-1/2}\overline{\varphi}'(\xi) + \frac{\partial}{\partial\xi}\overline{\left(e^{-\varphi} + \frac{\mu\varphi^2}{2}\right)}.$$

Now integrate from $-L$ to ξ and we see

$$\frac{1}{2}\overline{\left(\frac{\partial\varphi}{\partial\xi}\right)^2}\Bigg|_{-L}^{\xi} = (1+2\overline{\varphi})^{1/2} + \overline{e^{-\varphi} + \frac{\mu\varphi^2}{2}}\Bigg|_{-L}^{\xi}. \tag{3.20}$$

From Lemma 3.12 the right-hand side of (3.20) is bounded uniformly in μ,ξ. Hence by Jensen's inequality

$$\left(\frac{\partial\overline{\varphi}}{\partial\xi}(\xi)\right)^2 \leq \overline{\left(\frac{\partial\varphi}{\partial\xi}(\cdot,-L)\right)^2} + \text{ const.} \tag{3.21}$$

Since $\frac{\partial\varphi}{\partial\xi}(\tau,-L) \geq 0$, the greatest value of $\frac{\partial\varphi}{\partial\xi}(\tau,-L)$ must occur in case (a) when $\varphi \to 0$ as $\xi \to -L$ from above $\overline{\varphi}$. But in case (a) we already know from Lemma 3.13 that $\frac{\partial\varphi}{\partial\xi}$ is bounded uniformly in ξ,τ,μ. Therefore (3.21) implies the statement of the lemma. ∎

LEMMA 3.16. *In case* (b), *$\frac{\partial\varphi}{\partial\xi}$ is bounded uniformly in ξ,τ,μ.*

Proof. We use the same solution as in the proof of Lemma 3.12 and let ξ_* denote the point where $\varphi(\xi_*,\tau) = \overline{\varphi}(\xi_*)$ (where of course ξ_* will generally depend on τ). Recall that on $(-L,\xi_*)$, $\varphi(\xi,\tau)$ is below $\overline{\varphi}$, and on $(\xi_*,0]$, $\varphi(\xi,\tau)$ is above $\overline{\varphi}$.

We first consider $(-L,\xi_*)$ and let τ_1,τ_2 be values of τ for which cases (a) and (b) occur. (Recall that case (a) must occur since otherwise we would have $\overline{\varphi} < 0$ at some values of ξ on $(-L,0)$.) Hence $\varphi(\xi,\tau_2) < \overline{\varphi}(\xi) < \varphi(\xi,\tau_1)$ on $(-L,\xi_*(\tau_2))$. Set $\psi(\xi) = \varphi(\xi,\tau_1) - \varphi(\xi,\tau_2)$. Then

$$\frac{\partial^2\psi}{\partial\xi^2} = e^{-\varphi(\xi,\tau_2)} - e^{-\varphi(\xi,\tau_1)} + \mu(\varphi(\xi,\tau_1) - \varphi(\xi_2,\tau_2)) \geq 0$$

on $(-L,\xi_*(\tau_2))$. Since $\varphi(\xi,\tau_1)$ approaches 0 as $\xi \to -L$ from above $\overline{\varphi}$, while $\varphi(\xi,\tau_2)$ approaches 0 as $\xi \to -L$ from below $\overline{\varphi}$, we have $\psi'(-L) \geq 0$ and hence $\psi'(\xi) \geq 0$ on $[-L,\xi_*(\tau_2)]$. Thus $\frac{\partial\varphi}{\partial\xi}(\xi,\tau_1) > \frac{\partial\varphi}{\partial\xi}(\xi,\tau_2)$ on $[-L,\xi_*(\tau_2)]$, and by Lemma 3.13, $F(\tau_1) \geq \frac{\partial\varphi}{\partial\xi}(\xi,\tau_2)$.

Thus on $(-L, \xi_*(\tau_2)]$ we have $\frac{\partial\varphi}{\partial\xi}$ bounded from above uniformly in ξ, τ, μ. When $\frac{\partial\varphi}{\partial\xi}(\xi, \tau_2) \geq 0$, then the bound from below is trivial. Otherwise, a bound from below can be obtained by considering two cases.

Case 1. $\frac{\partial\varphi}{\partial\xi}(\xi_*(\tau_2), \tau_2) \geq 0$. In this case, $\frac{\partial\varphi}{\partial\xi}(\xi, \tau_2)$ can be negative but only on subintervals, say (ξ_1, ξ_2), for which $\frac{\partial\varphi}{\partial\xi}(\xi_1, \tau_2) = \frac{\partial\varphi}{\partial\xi}(\xi_2, \tau_2) = 0$. This is because at the end points of the region $(-L, \xi_*(\tau_2))$ we have $\frac{\partial\varphi}{\partial\xi}(-L, \tau_2) \geq 0$ and $\frac{\partial\varphi}{\partial\xi}(\xi_*(\tau_2), \tau_2) \geq 0$. Thus on a region (ξ_1, ξ_2), for which $\frac{\partial\varphi}{\partial\xi}(\xi, \tau_2) < 0$, we have

$$\frac{\partial^2\varphi}{\partial\xi^2} \leq (1 + 2\varphi)^{-1/2} - e^{-\varphi} + \mu\varphi$$

(since $\varphi < \overline{\varphi}$), and multiplication by $\frac{\partial\varphi}{\partial\xi}(\xi, \tau_2)$ and integration from ξ to ξ_2 yields

$$\frac{1}{2}\left(\frac{\partial\varphi}{\partial\xi}\right)^2\Bigg|_{\xi}^{\xi_2} \geq \left[(1 + 2\varphi)^{1/2} + e^{-\varphi} + \frac{\mu\varphi^2}{2}\right]\Bigg|_{\xi}^{\xi_2}$$

and hence

$$\sup_{\xi}\left\{(1 + 2\varphi)^{1/2} + e^{-\varphi} + \frac{\mu\varphi^2}{2}\right\} \geq \frac{1}{2}\left(\frac{\partial\varphi}{\partial\xi}(\xi, \tau_2)\right)^2. \tag{3.22}$$

Since Lemma 3.12 implies that the left-hand side of (3.21) is bounded uniformly in μ, τ, we have $\frac{\partial\varphi}{\partial\xi}(\xi, \tau_2)$ uniformly bounded in ξ, μ, τ when $\frac{\partial\varphi}{\partial\xi}(\xi, \tau_2) < 0$ and $-L \leq \xi \leq \xi_*(\tau_2)$.

Case 2. $\frac{\partial\varphi}{\partial\xi}(\xi_*(\tau_2), \tau_2) < 0$. In this case, since $\varphi < \overline{\varphi}$ on $(-L, \xi_*(\tau_2))$ and $\varphi > \overline{\varphi}$ on $(\xi_*(\tau_2), 0)$, we must have

$$\overline{\varphi}'(\xi_*(\tau_2)) < \frac{\partial\varphi}{\partial\xi}(\xi_*(\tau_2), \tau_2) < 0. \tag{3.23}$$

Thus on an interval $(\xi_3, \xi_*(\tau_2))$, where

$$\frac{\partial\varphi}{\partial\xi}(\xi_3, \tau_2) = 0, \qquad \frac{\partial\varphi}{\partial\xi}(\xi_*(\tau_2), \tau_2) < 0,$$

and $\frac{\partial\varphi}{\partial\xi}(\xi, \tau) < 0$ on $(\xi_3, \xi_*(\tau_2))$, we have again have from (3.14)

$$\frac{\partial^2\varphi}{\partial\xi^2} \leq (1 + 2\varphi)^{-1/2} - e^{-\varphi} + \mu\varphi,$$

and multiplication by $\frac{\partial\varphi}{\partial\xi} < 0$ and integration from ξ to $\xi_*(\tau_2)$ yields

$$\frac{1}{2}\left(\frac{\partial\varphi}{\partial\xi}\right)^2\Bigg|_{\xi}^{\xi_*(\tau_2)} \geq \left[(1 + 2\varphi)^{1/2} + e^{-\varphi} + \frac{\mu\varphi^2}{2}\right]\Bigg|_{\xi}^{\xi_*(\tau_2)}$$

and

$$\frac{1}{2}\left(\frac{\partial\varphi}{\partial\xi}(\xi_*(\tau_2), \tau_2)\right)^2 + \sup_{\xi}\left((1 + 2\varphi)^{1/2} + e^{-\varphi} + \frac{\mu\varphi^2}{2}\right) \geq \frac{1}{2}\left(\frac{\partial\varphi}{\partial\xi}(\xi, \tau_2)\right)^2. \tag{3.24}$$

Lemma 2.12 shows that the second term on the left-hand side of (3.24) is bounded uniformly in μ, τ, while inequality (3.23) and Lemma 3.15 cover the first term, and Case 2 is complete.

To complete the proof of the lemma we must consider the interval $(\xi_*(\tau_2), 0]$. Now define $\Gamma(\xi) = \varphi(\xi, \tau_2) - \varphi(\xi, \tau_3)$, where τ_2, τ_3 are values for which cases (b), (c) occur, respectively. Again note that if the case (b) occurs, where $\varphi > \overline{\varphi}$ on $(\xi_*(\tau_2), 0]$ since case (a) must occur, then case (c) must also occur. Otherwise, φ would be greater than $\overline{\varphi}$ on some ξ interval for all $\tau \in [0, 1]$. Thus $\Gamma(\xi)$ is well defined and $\Gamma(\xi) > 0$. From (3.14)

$$\Gamma''(\xi) = -e^{-\varphi(\xi,\tau_2)} + e^{-\varphi(\xi,\tau_3)} + \mu\Gamma(\xi) > 0 \tag{3.25}$$

since $\varphi(\xi, \tau_2) > \overline{\varphi} > \varphi(\xi, \tau_3)$ on $(\xi_*(\tau_2), 0]$. Integrate (3.25) from ξ to 0. We obtain

$$\Gamma'(\xi) < \Gamma'(0) = \frac{\partial\varphi}{\partial\xi}(0, \tau_2) - \frac{\partial\varphi}{\partial\xi}(0, \tau_3) = F(\tau_2) - F(\tau_3)$$

and hence

$$\frac{\partial\varphi}{\partial\xi}(\xi, \tau_2) < \frac{\partial\varphi}{\partial\xi}(\xi, \tau_3) + F(\tau_2) - F(\tau_3). \tag{3.26}$$

Now on $(\xi_*(\tau_2), 0)$ we have $\varphi(\xi, \tau_2) > \overline{\varphi}(\xi)$, and hence by (3.14)

$$\frac{\partial^2\varphi}{\partial\xi^2} > (1 + 2\varphi)^{-1/2} - e^{-\varphi} + \mu\varphi > 0. \tag{3.27}$$

Hence $\varphi(\xi, \tau_2)$ can have a local minimum but no local maximum. Hence either $\varphi(\xi, \tau_2)$ is monotone increasing or $\varphi(\xi, \tau_2)$ has a local minimum at $\xi_{**}(\tau_2), \xi_*(\tau_2) < \xi_{**}(\tau_2) < 0$. If $\varphi(\xi, \tau_2)$ is monotone increasing, then (3.26) and Lemma 3.14 imply that $\frac{\partial\varphi}{\partial\xi}(\xi, \tau_2)$ is bounded uniformly in ξ, μ, τ, and we are finished. If $\varphi(\xi, \tau_2)$ has a local minimum at $\xi_{**}(\tau_2)$, then on $(\xi_{**}(\tau_2), 0)$, $\frac{\partial\varphi}{\partial\xi}(\xi, \tau_2) \geq 0$, and again (3.26) yields the uniform bound. Finally on $(\xi_*(\tau_2), \xi_{**}(\tau_2))$ we have $\frac{\partial\varphi}{\partial\xi}(\xi, \tau_2) < 0$. Multiply (3.27) by $\frac{\partial\varphi}{\partial\xi}(\xi, \tau_2) < 0$. Then integration from $\xi_*(\tau_2)$ to ξ yields

$$\frac{1}{2}\left(\frac{\partial\varphi}{\partial\xi}\right)^2 \Bigg|_{\xi_*(\tau_2)}^{\xi} < \left[(1 + 2\varphi)^{1/2} + e^{-\varphi} + \frac{\mu\varphi^2}{2}\right] \Bigg|_{\xi_*(\tau_2)}^{\xi}. \tag{3.28}$$

The right-hand side of (3.28) is bounded uniformly in ξ, μ, τ by Lemma 3.12, and $\frac{\partial\varphi}{\partial\xi}(\xi_*(\tau_2), \tau_2)$ is also bounded uniformly in ξ, μ, τ by Cases 1 and 2 above. Hence $\frac{\partial\varphi}{\partial\xi}(\xi, \tau_2)$ is uniformly bounded as well, and the lemma is proven. ∎

LEMMA 3.17. *Let $\tau, \sigma \in [0, 1]$. Then*

$$\sup_{\xi} |\varphi(\tau, \xi) - \varphi(\sigma, \xi)| \leq \text{ const.} |F(\tau) - F(\sigma)|,$$

where the const. is independent of τ, σ, μ.

Proof. Set $w(\xi) = \varphi(\tau, \xi) - \varphi(\sigma, \xi)$. Then from (3.14)

$$w''(\xi) = -e^{-\varphi(\tau,\xi)} + e^{-\varphi(\sigma,\xi)} + \mu w,$$

and by the mean value theorem

$$e^{-\varphi(\sigma,\xi)} - e^{-\varphi(\tau,\xi)} = -e^{-\gamma(\xi,\sigma,\tau)}(\varphi(\sigma, \xi) - \varphi(\tau, \xi)) = e^{-\gamma(\xi,\sigma,\tau)} w(\xi),$$

where either

$$0 < \varphi(\sigma,\xi) < \gamma(\xi,\sigma,\tau) < \varphi(\tau,\xi) \quad \text{or} \quad 0 < \varphi(\tau,\xi) < \gamma(\xi,\sigma,\tau) < \varphi(\sigma,\xi). \tag{3.29}$$

Hence we see

$$w''(\xi) = e^{-\gamma(\xi,\sigma,\tau)} w(\xi) + \mu w(\xi), \tag{3.30}$$

and multiplying (3.30) by $w(\xi)$ and integrating from $-L$ to ξ we find

$$-\int_{-L}^{0} w'(\xi)^2 d\xi + w(\xi)w'(\xi)\,|_{-L}^{0} = \int_{-L}^{0} e^{-\gamma(\xi,\sigma,\tau)} w^2(\xi) + \mu w^2(\xi) d\xi. \tag{3.31}$$

Since $w(-L) = 0, w'(0) = F(\tau) - F(\sigma)$, γ is nonnegative, and by Lemma 3.12 the function γ is uniformly bounded in τ, σ, μ, ξ, we see

$$\begin{aligned} -\int_{-L}^{0} w'(\xi)^2 d\xi + w(0)(F(\tau) - F(\sigma)) \\ \geq \text{const}_1 \int_{-L}^{0} w^2(\xi) d\xi, \end{aligned} \tag{3.32}$$

where

$$e^{-\gamma(\xi,\sigma,\tau)} \geq \text{const}_1 > 0$$

and the const_1 is independent of μ, ξ, τ. From (3.32) we easily see

$$\begin{aligned} \sup_{\xi} |w(\xi)||F(\tau) - F(\sigma)| &\geq \text{const}_1 \int_{-L}^{0} w'(\xi)^2 + \int_{-L}^{0} w(\xi)^2 d\xi \\ &\geq \text{const}_2 \left(\sup_{\xi} |w(\xi)| \right)^2, \end{aligned}$$

where const_2 is independent of μ, ξ, τ. The lemma is proven. ∎

Now define the extended function φ_e:

$$\varphi_e(\xi,\tau;\mu) \stackrel{\text{def}}{=} \begin{cases} \varphi(\xi,\tau), & -L \leq \xi \leq 0, \\ 0, & \xi < -L. \end{cases}$$

LEMMA 3.18. *$\{\varphi_e(\cdot,\cdot;\mu)\}$ is uniformly bounded and equicontinuous on $(-\infty, 0] \times [0,1]$. Furthermore, $\{\varphi_e(\cdot,\cdot;\mu)\}$ has a subsequence which converges as $\mu \to 0$ uniformly on compact subsets of $(-\infty, 0] \times [0,1]$ to a function $\varphi_\#$. The function $\varphi_\#(\xi,\tau)$ is periodic in τ with period 1.*

Proof. Lemmas 3.12, 3.13, 3.14, 3.16, 3.17 prove the uniform boundedness and equicontinuity. Existence of a convergent subsequence on compact subsets of $(-\infty, 0] \times [0,1]$ follows from the Ascoli–Arzela theorem. Finally, the limit function $\varphi_\#$ inherits the inequality of Lemma 3.17, and hence $\varphi_\#$ is periodic in τ with period 1. ∎

LEMMA 3.19. *The limit function $\varphi_\#(\cdot,\tau)$ is nonnegative on $(-\infty, 0]$ and satisfies (3.1), (3.2) for every $\tau \in [0,1]$.*

Proof. Since $\varphi(\cdot,\cdot;\mu)$ is a smooth solution of (3.14)–(3.16), it solves the weak form of the equations (3.14)–(3.16):

$$\int_{-L}^{0} \left(\frac{\partial \varphi}{\partial \xi} - F(\tau) \right) \psi'(\xi) + ((1 + 2\overline{\varphi})^{-1/2} + e^{-\varphi} + \mu\varphi)\psi(\xi) d\xi = 0,$$

$0 \le \tau \le 1$, and all $\psi \in C_0^\infty(-\infty, 0]$. Integration by parts and use of the definition of φ_e imply

$$\int_{-\infty}^{0} -\varphi_e \psi''(\xi) + ((1 + 2\overline{\varphi}_e)^{-1/2} + e^{-\varphi_e} + \mu\varphi_e)\psi(\xi) d\xi + (\varphi_e(0,\tau) - F(\tau))\psi'(0) = 0. \tag{3.33}$$

Now let $\mu \to 0$. By Lemma 3.18, $\{\varphi_e(\cdot,\cdot,\mu)\}$ has a convergent subsequence which approaches $\varphi_\#$ uniformly on compact subsets of $(-\infty, 0] \times [0, 1]$. By the uniform convergence of φ_e to $\varphi_\#$, the limit function $\varphi_\#$ satisfies (3.33) and is a weak solution of (3.1), (3.2). Since the right-hand side of (3.1) is continuous in ξ, we see that $\varphi_\#$ is C^2 in ξ, and $\varphi_\#$ is a classical solution of (3.1), (3.2), twice continuously differentiable in ξ and continuous in τ. The nonnegativity of $\varphi_\#$ follows from the nonnegativity of φ_e. ∎

THEOREM 3.20. *Assume that* $F(\tau) = p\int_0^t f(\tau)d\tau + \frac{\partial\varphi_0}{\partial\xi}(0,0) > 0$ *is given, continuously differentiable, and periodic with period* 1. *Then the limit function* $\varphi_\#$ *is a classical solution of* (3.1)–(3.3)*, twice continuously differentiable in* ξ*, continuous in* τ*, on* $(-\infty, 0] \times [0, 1]$*, and periodic in* τ *with period* 1.

Proof. From Lemma 3.19 we know that $\varphi_\#$ satisfies (3.1)–(3.2). Thus we need only verify the boundary condition (3.3). Since $\varphi(\cdot,\cdot,\mu)$ satisfied cases (a), (b), (c), we see that $\varphi_e(\cdot,\cdot,\mu)$ satisfies the following cases:

(a_e). $\varphi_e(0,\tau;\mu) > \overline{\varphi}_e(0;\mu)$, and the graph of φ_e is always above the graph of $\overline{\varphi}_e$ on $(-L, 0)$ and $\varphi_e(\xi,\tau;\mu) = \overline{\varphi}_e(\xi,\mu) = 0$ on $(-\infty, -L)$.

(b_e). $\varphi_e(0,\tau;\mu) > \overline{\varphi}_e(0;\mu)$, and the graph of φ_e intersects the graph of $\overline{\varphi}_e$ once on $(-L, 0)$ and $\varphi_e(\xi,\tau;\mu) = \overline{\varphi}_e(\xi,\mu) = 0$ on $(-\infty, -L)$.

(c_e). $\varphi_e(0,\tau;\mu) \le \overline{\varphi}_e(0;\mu)$, and the graph of φ_e is always below the graph of $\overline{\varphi}_e$ on $(-L, 0)$ and $\varphi_e(\xi,\tau;\mu) = \overline{\varphi}_e(\xi,\mu) = 0$ on $(-\infty, -L)$.

Since $\varphi_e(\cdot,\tau;\mu_k) \to \varphi_\#(\cdot,\tau)$ uniformly on compact subsets of $(-\infty, 0]$ for some sequence $\mu_k \to 0$ as $k \to \infty$, the inequalities $\varphi_e \ge \overline{\varphi}_e$ and $\varphi_e \le \overline{\varphi}_e$ of cases (a_e), (c_e) are preserved in the limit.

Case (b_e) has $\varphi_e > \overline{\varphi}_e$ on $(\xi_*(\mu), 0]$, $\varphi_e < \overline{\varphi}_e$ on $(-L, \xi_*(\mu))$, $\varphi_e = \overline{\varphi}_e$ on $(-\infty, -L]$. Consider the sequence $\{\xi_*(\mu_k)\}$. If $\xi_*(\mu_k) \to \infty$ as $\mu_k \to 0$, then the limit function $\varphi_\#$ satisfies $\varphi_\# \ge \overline{\varphi}_\#$ on $(-\infty, 0]$. On the other hand, if $\{\xi_*(\mu_k)\}$ is bounded as $\mu_k \to 0$, then, possibly extracting a convergent subsequence if necessary, we have $\xi_*(\mu_k) \to \xi_0 \le 0$ and, by uniform convergence, $\varphi_\# \ge \overline{\varphi}_\#$ on $(\xi_0, 0]$, $\varphi_\# \le \overline{\varphi}_\#$ on $(-\infty, \xi_0]$. Hence the limit function $\varphi_\#$ satisfies

($a_\#$). $\varphi_\#(0,\tau) \ge \overline{\varphi}_\#(0)$, and the graph of $\varphi_\#$ is above or touching the graph of $\overline{\varphi}_\#$ on $(-\infty, 0]$.

($b_\#$). $\varphi_\#(0,\tau) \ge \overline{\varphi}_\#(0)$, and the graph of $\varphi_\#$ intersects the graph of $\varphi_\#$ once on $(-\infty, 0]$.

($c_\#$). $\varphi_\#(0,\tau) \le \overline{\varphi}_\#(0)$, and the graph of $\varphi_\#$ is below or touching the graph of $\overline{\varphi}_\#$ on $(-\infty, 0]$.

In case ($a_\#$), $\varphi_\#$ is obtained as the uniform limit of monotone increasing functions (via Lemma 3.13). Hence $\varphi_\#$ is nondecreasing in ξ, and $\frac{\partial\varphi_\#}{\partial\xi} \ge 0$ on $(-\infty, 0)$. Also since $\overline{\varphi}_\# \ge \varphi_\#$, (3.1) implies

$$\frac{\partial^2\varphi_\#}{\partial\xi^2} \ge (1 + 2\varphi_\#)^{-1/2} - e^{-\varphi_\#} \ge 0, \tag{3.34}$$

and hence both $\varphi_\#$ and $\frac{\partial\varphi_\#}{\partial\xi}$ are nondecreasing and bounded from below on $(-\infty, 0]$. Hence $\lim_{\xi\to-\infty}\varphi_\#(\xi,\tau)$ and $\lim_{\xi\to-\infty}\frac{\partial\varphi_\#}{\partial\xi}(\xi,\tau)$ exist.

Now integrate (3.34) from $-\infty$ to 0 in ξ. We see

$$F(\tau) - \frac{\partial\varphi_\#}{\partial\xi}(-\infty,\tau) > \int_{-\infty}^{0}((1+2\varphi_\#)^{-1/2} - e^{-\varphi_\#})d\xi. \tag{3.35}$$

Since the integrand in (3.35) is nonnegative and $\lim_{\xi\to\infty}\varphi_\#$ exists, (3.35) implies that the limit is zero. Similarly we have

$$\varphi_\#(0,\tau) - \varphi_\#(-\infty,\tau) = \int_{-\infty}^{0}\frac{\partial\varphi_\#}{\partial\xi}(\xi,\tau)d\xi. \tag{3.36}$$

Again the left-hand side of (3.36) is bounded and the integrand is nonnegative; hence $\lim_{\xi\to-\infty}\frac{\partial\varphi_\#}{\partial\xi}(\xi,\tau) = 0$.

Cases ($b_\#$) and ($c_\#$): Since case ($a_\#$) must always occur, and since for ξ sufficiently negative the graphs of $\varphi_\#$ in cases ($b_\#$) and ($c_\#$) lie below the graph $\varphi_\#$ of case ($a_\#$), we have $\lim_{\xi\to\infty}\varphi_\#(\xi,\tau) = 0$ in cases ($b_\#$) and ($c_\#$) as well.

Now set $w(\xi) = \varphi_\#(\tau_1,\xi) - \varphi_\#(\tau_2,\xi)$, where $\varphi_\#(\tau_1,\xi)$ is as in case ($a_\#$) and $\varphi(\tau_2,\xi)$ is as in case ($b_\#$). Then since $\varphi_\#(\tau_1,\xi) > \varphi_\#(\tau_2,\xi)$ for ξ sufficiently negative, say, $-\infty < \xi < \xi_1 < 0$, we have

$$w(\xi) > 0, \tag{3.37}$$

$$w''(\xi) = -e^{-\varphi_\#(\tau_1,\xi)} + e^{-\varphi_\#(\tau_2,\xi)} > 0 \tag{3.38}$$

on $(-\infty,\xi_1)$.

If $w'(\xi_1) < 0$, then $w'(\xi) < 0$ on $(-\infty,\xi_1)$ by (3.38). Since $w(\xi)\to 0$ as $\xi\to-\infty$, we have

$$0 > \int_{-\infty}^{\xi_1} w'(\xi)d\xi = w(\xi_1),$$

which is impossible. Hence the only possibility is $w'(\xi_1)\geq 0$. We must have $w'(\xi)\geq 0$ on $(-\infty,\xi_1)$ since if $w'(\xi_2) < 0$ for some $-\infty < \xi_2 < \xi_1$, then the preceding argument with ξ_1 replaced by ξ_2 would show $w(\xi_2) < 0$, a contradiction. Thus for decreasing ξ, $w'(\xi)$ is monotone decreasing and bounded from below by zero. Hence $\lim_{\xi\to-\infty}w'(\xi)$ exists. Since

$$w(\xi) = \int_{-\infty}^{\xi} w'(\xi)d\xi$$

and the left-hand side is bounded and the integrand is nonnegative, $\lim_{\xi\to-\infty}w'(\xi) = 0$. Hence

$$\lim_{\xi\to-\infty}\frac{\partial\varphi_\#}{\partial\xi}(\tau_2,\xi) = \lim_{\xi\to-\infty}\left(\frac{\partial\varphi_\#}{\partial\xi}(\tau_1,\xi) - w'(\xi)\right) = 0,$$

and case ($b_\#$) is complete. Fortunately the same argument that was used for case ($b_\#$) applies to case ($c_\#$), and the theorem is proven. □

REFERENCES

[1] M. A. Lieberman and A. J. Lichtenberg, *Principles of Plasma Discharges and Material Processing*, John Wiley, New York, 1994.

[2] V. A. Godyak and N. Sternberg, *Dynamic model of the electrode sheaths in symmetrically driven discharges*, Phys. Rev. A, 42 (1990), pp. 2299–2312.

[3] N. Sternberg and V. A. Godyak, *Solving the mathematical model of the electrode sheath in symmetrically driven rf discharges*, J. Comput. Phys., 111 (1994), pp. 347–353.

[4] J. Gierling and K.-U. Riemann, *Comparison of a consistent theory of radio frequency sheaths with step models*, J. Appl. Phys., 83 (1988), pp. 3521–3528.

[5] R. N. Franklin and J. R. Ockendon, *Asymptotic matching of plasma and sheath in an active low pressure discharge*, J. Plasma Phys., 4 (1970), pp. 371–385.

[6] M. Gegick and G. W. Young, *Plasma carburization of an axisymmetric steel sample*, SIAM J. Appl. Math., 54 (1994), pp. 877–906.

[7] M. Slemrod, *Monotone increasing solutions of the Painlevé* 1 *equation* $y'' = y^2 + x$ *and their role in the stability of the plasma-sheath transition*, European J. Appl. Math., 13 (2002), pp. 663–680.

[8] N. Joshi and A. V. Kitaev, *On Boutroux's tritronquée solutions of the first Painlevé equation*, Stud. Appl. Math., 107 (2001), pp. 253–291.

[9] E. L. Ince, *Ordinary Differential Equations*, Dover, New York, 1944.

[10] L. C. Evans, *Partial Differential Equations*, AMS, Providence, 1998.

SIAM J. APPL. MATH.
Vol. 63, No. 5, pp. 1764–1788

ACOUSTIC PULSE SPREADING IN A RANDOM FRACTAL*

KNUT SØLNA†

Abstract. Fractal medium models are used to model, for instance, the heterogeneous earth and the turbulent atmosphere. A wave pulse propagating through such a medium will be affected by multiscale medium fluctuations. For a class of one-dimensional fractal random media defined in terms of fractional Brownian motion we show how the wave interacts with the medium fluctuations. The modification in the pulse shape depends on the roughness of the medium and can be described in a *deterministic* way when the pulse is observed at its *random* arrival time. For very rough media the coherent wave is confined to a surface layer.

Key words. wave propagation, random medium, fractional Brownian motion, homogenization, anomalous diffusion

AMS subject classifications. 34F05, 34E10, 37H10, 60H20

DOI. 10.1137/S0036139902404657

1. Introduction.

Propagation of wave pulses in a *smooth* medium is well understood, but propagation in a *rough* or *multiscale* medium is not as well understood. We will look at how a propagating pulse interacts with rough variations in the medium.

Given the importance and long history of wave propagation and scattering problems, a multitude of approaches have been developed to analyze them. In the homogenization or effective media regime, rapidly varying properties of the medium average out when the width of the propagating pulse is large compared to the scale of the medium fluctuations. However, over long propagation distances the accumulated effect of the scattering, associated with the medium microstructure, gradually changes the pulse *beyond* the geometrical effects of the high frequency analysis in the smooth homogenized medium. In the 1960s and early 1970s, *mean* pulse propagation over long distances was analyzed. More recently a mathematical theory has been developed that gives a more precise description [1, 8, 16] of pulse propagation. It deals with pulses in a particular realization of the random medium and explains why in many cases the evolution of the *pulse shape* is to leading order *deterministic*. We refer to this phenomenon as pulse *stabilization*. So far, two salient features of this "pulse shaping" theory have been that it assumes a one-dimensional medium and a separation of scales for the medium heterogeneities; that is, the medium has features on microscales which are well separated from the macroscale. However, several studies [9, 13, 14, 21] suggest that, for instance, the earth's crust should be modeled as containing fluctuations on a continuum of length scales. Multiscale medium models are also used for the turbulent atmosphere [20] and, moreover, to model the transition zone between different parts of tissues or the zone between different parts of certain devices, for instance, the zone associated with a large change in the dielectric permittivity. Burridge, Papanicolaou, and White give a nice derivation of pulse shaping in periodic and stationary random media in [6]. Here, we generalize the pulse shaping theory for a two scale medium, as presented in [6], to the multiscale case.

*Received by the editors March 27, 2002; accepted for publication (in revised form) November 5, 2002; published electronically July 26, 2003.
http://www.siam.org/journals/siap/63-5/40465.html
†Department of Mathematics, University of California at Irvine, Irvine, CA 92697 (ksolna@math.uci.edu). Supported by NSF grant DMS-009399.

The stabilization phenomenon has been shown to hold true also for waves propagating in three spatial dimensions in the case with layered media. This problem has been analyzed in detail in [7] and more recently also in [12], where it is discussed in the context of time-reversal of waves. Stabilization and pulse shaping in the case with slow lateral variations in the medium have been analyzed in [15] and [25].

The analysis of the interaction of a wave pulse with a medium varying on many length scales is an interesting but largely open question from a mathematical viewpoint, despite its importance in applications. We analyze this problem for acoustic waves propagating in a one-dimensional discrete medium, modeled in terms of fractional Brownian motion. Fractional Brownian motion is a Gaussian (self-similar) stochastic process and is often used as a model for processes containing fluctuations on a continuum of length scales, for instance, for modeling of turbulent environments. The discretization assumes that the medium has a smallest scale. In turbulence theory this is the *inner* scale. Below, we refer to media defined in terms of fractional Brownian motion as "fractal" media. The Hurst exponent H characterizes the roughness of the fractional Brownian motion, and the value $H = 1/2$ gives standard Brownian motion. In the simplest case with $H = 1/2$ the medium model that we consider satisfies a separation of scales assumption. For $H \neq 1/2$ the medium contains long-range interactions and variations on many scales. We show that in the limit of small inner scale relative to the travel distance the transformation of the pulse shape becomes *deterministic*; thus the classic pulse shaping theory for media satisfying a separation of scales assumption generalizes in this sense. However, now the *scale* on which the spreading of the pulse happens depends on the roughness of the medium and does *not* in general correspond to the inner scale as in the classic theory. In fact, a pulse supported on the inner scale is trapped in a surface layer if the medium is rougher than the standard model. If the medium is smoother than the standard model, i.e., $H > 1/2$, the shape of such a pulse is *not* affected by the random medium fluctuations.

Most previous work on wave interaction with a fractal object deals with scattering caused by fractal interfaces. However, some authors have explored wave-interaction with deterministic fractal media using numerical simulations [4, 17, 26]. Reflections from a random fractal and how they depend on the fractal exponent is explored by numerical experiments in [4]. In [17], Konotop, Fei, and Vazquez examine the wave reflections from a fractal devil's staircase and introduce a heuristic scheme for computing effective parameters of such a medium. Sun and Jaggard [26] numerically explore wave propagation in a similar medium and observe strong resonance effects. Here, we analyze acoustic pulse transmission through a random fractal and illustrate our theoretical results with numerical simulations.

In section 2 we state the governing equations for the acoustic pulse and in section 3 the models for the fractal media that we consider. We summarize how the pulse shaping theory generalizes to these media in section 4. In section 5 we derive the general averaging result that can be used for fractal media. Finally, in section 6 we apply this averaging result to the fractal media that we consider and also illustrate our theoretical results with numerical simulations.

2. Governing equations. We follow the notation set forth in [3] and [6]. The governing equations for the continuum are the Euler equations giving conservation of momentum and mass:

$$\begin{aligned} \rho\, u_t + p_z &= 0, \\ K^{-1} p_t + u_z &= 0, \end{aligned} \tag{1}$$

with t being time and z measuring depth into the medium. The dependent variables are the pressure p and the (z-component) of the particle velocity u. The medium parameters are the density ρ and the bulk-modulus K, which is the reciprocal of the compressibility. We next make a change of variables from depth z to the first arrival time from the surface to this depth:

$$x = x(z) \;=\; \int_0^z \frac{1}{c(s)}\, ds, \tag{2}$$

with the local speed of sound being $c \;=\; \sqrt{K/\rho}$. The first arrival time gives the travel time for the first arriving disturbances. An important aspect of the propagating pulse is the travel time of its *coherent* part and this differs in general from the first arrival time. In travel time coordinates (1) transforms into

$$\begin{aligned} \zeta u_t + p_x &= 0, \\ p_t + \zeta u_x &= 0. \end{aligned} \tag{3}$$

The characteristic impedance ζ is

$$\zeta = \zeta(x) \;=\; \sqrt{\rho(z(x))\; K(z(x))} \;=\; \rho(z(x))\; c(z(x)), \tag{4}$$

where $z(x)$ is the inverse of the map defined in (2).

We model ζ as being piecewise constant; thus, within each medium section the wave propagation can be described as a pure translation of "up"- and "down"-propagating wave components. We decompose the wavefield in terms of up- and down-propagating wave components as

$$\begin{aligned} u &= \frac{1}{\sqrt{\zeta}}\; (D - U), \\ p &= \sqrt{\zeta}\; (D + U). \end{aligned} \tag{5}$$

The positive x direction defines the downward direction, and D is the wave propagating in this direction. Our objective is to describe a down-propagating pressure pulse somewhere deep into the medium and examine how the multiscale random fluctuations in ζ affect this pulse. In section 3 we give the particular models that we consider for the medium fluctuations and in section 4 discuss their impact on the transmitted pulse.

3. Modeling of the medium. The discrete medium is defined by a *uniform* discretization in the travel time coordinate x as

$$\zeta = \begin{cases} 1 & \text{for } \; x < h, \\ \zeta_k^h & \text{for } \; (k-1)\; h \le x < k\; h, \end{cases} \tag{6}$$

with $k \in \{1, 2, \ldots\}$. Therefore, the time it takes a pulse to traverse a medium section is constant and equal to h. Such a medium is sometimes referred to as a Goupillaud medium; it has been discussed in, for instance, [6, 22, 24]. Here, we consider finely layered media, and h is the small parameter in our modeling. In the next section we describe our choices for the impedance sequence ζ_k^h, the sequence that defines the medium.

3.1. A standard one-scale medium model. We first consider a medium model where the fluctuations form a stationary process, with the impedances in the different medium sections being independent and identically distributed. Such a medium model is used in [5, 6, 16]. Let the discrete impedance sequence be given by

$$\zeta_k^h = 1 + Z_k^h, \tag{7}$$

with Z_k^h a sequence of independent mean zero Gaussian random variables with variance $\mathcal{O}(h)$. The medium fluctuations are therefore relatively weak. Note that in practice we truncate the fluctuations in the above model such that the impedance is positive and bounded. If β denotes standard Brownian motion, then a version of (7) can be constructed as

$$\zeta_k^h = 1 + \beta(kh) - \beta((k-1)h). \tag{8}$$

This formulation serves to motivate the medium model we introduce next, a model that incorporates fluctuations on many scales.

3.2. Multiscale medium from fractional Brownian noise. We aim to formulate a simple medium model that incorporates long range interactions or correlations, that is, a model that is not limited to one intrinsic scale as is the one in (8). A standard stochastic model process that incorporates long-range interactions and variations on a continuum of length scales is fractional Brownian motion (fBm), $\{\beta_H(x); x \geq 0\}$. This process was introduced by Mandelbrot and Van Ness in [18]. We define the medium model in terms of this process.

First, consider the following generalization of (8):

$$\zeta_k^h = 1 + \beta_H(kh) - \beta_H((k-1)h), \tag{9}$$

with β_H being fBm with Hurst exponent H. Thus, the fluctuations in the impedance form a fBm *noise* sequence. Note that $\beta_{1/2}$ is standard Brownian motion, and then the models (8) and (9) coincide. In general, fBm is a Gaussian process with mean zero, stationary increments, and with covariance and structure functions

$$E[\beta_H(x)\beta_H(y)] = \frac{\sigma^2}{2}(|x|^{2H} + |y|^{2H} - |x-y|^{2H}), \tag{10}$$

$$E[(\beta_H(x) - \beta_H(x-\Delta x))^2] = \sigma^2|\Delta x|^{2H}, \tag{11}$$

where $0 < H < 1$, σ is a scaling parameter, and $\beta_H(0) = 0$. The Hurst exponent H determines the correlation of the increments. The covariance of a future increment with the past increment is

$$E[(\beta_H(x) - \beta_H(x-\Delta x))(\beta_H(x+\Delta x) - \beta_H(x))] = \sigma^2(2^{2H-1} - 1)|\Delta x|^{2H}$$

and is independent of the location index x. When $H > 1/2$ this quantity is positive, so if the past increment is positive, then on average the future increment will be positive. Feder [10] calls this persistence. When $H < 1/2$ we have an antipersistent process, with a positive increment in the past making a positive increment in the future less likely. The paths of fBm in the persistent case will be associated with larger excursions, but will be "smoother" than the paths in the antipersistent case. The quadratic variation of the process in the persistent case is almost surely zero,

whereas it is almost surely infinite in the antipersistent case [23]. Below we show how this entails that wave propagation through a medium defined in terms of antipersistent fBm be qualitatively very different from wave propagation in the persistent case.

In the model (8) the impedance is piecewise constant and ζ_k^h are uncorrelated with ζ_{k+m}^h unless $m = 0$. We refer to this model as a one-scale model whose scale of variation corresponds to the discretization scale.

In this section we consider a model for the impedance that has long-range correlations; the covariance is now

$$C_\zeta(m; h, H, \sigma) := Cov[\zeta_k^h, \zeta_{k+m}^h] \ \sim \ \frac{\sigma^2 h^{2H} H(2H-1)}{m^{2(1-H)}} \quad \text{as } m \ \to \ \infty,$$

and the medium therefore exhibits correlations also over long scales. Note that if we observe the medium on a coarser scale, then we see the same decay of correlations for $a > 0$:

$$C_\zeta(am; h, H, \sigma) \sim a^{-2(1-H)} C_\zeta(m; h, H, \sigma) \quad \text{as } m \ \to \ \infty.$$

Since the medium has similar and nontrivial correlation structure over many scales, we refer to it as a multiscale model.

In fact, fBm itself is self-similar since $\beta_H(x)$ and $a^H \beta_H(x/a)$ have the same finite-dimensional distributions for all $a > 0$. This property illustrates how this process incorporates variations on all scales.

3.3. A fractal medium model. The model (9) is defined in terms of fractional Gaussian noise, and the medium fluctuations are therefore stationary. Next, we define a medium model where the fluctuations are defined by the fBM process itself. In this case the fluctuations are nonstationary; moreover, they are strong $\mathcal{O}(1)$ and not weak as in the above two models. We consider the medium model

$$\zeta_k^h = \exp(\beta_H(kh)). \tag{12}$$

The value $H = 1/3$ is of particular interest since the fBm process then corresponds to Kolmogorov turbulence, a standard medium model in the context of wave propagation in the turbulent atmosphere. We will see below that the same theorem, Theorem 5.1, that characterizes the transformation of the pulse shape for the models in the previous two subsections with weak or small medium fluctuations applies in this case with relatively strong medium fluctuations. Below, in (16), we introduce the interface reflection coefficients associated with the sequence ζ_k^h. The theorem characterizes the way in which the decay of the correlations in these interface reflection coefficients determines how the medium affects the shape of the propagating wave pulse. The important parameter that determines this decay is the Hurst exponent H, and the pulse shaping thus depends sensitively on the value of this. We give the decay of correlations for the interface reflection coefficients in (41). Note that even though the fluctuations of the impedance in (12) are large, the magnitude of the fluctuations of the interface reflection coefficients are actually small.

Observe finally that the analysis we present below holds for more general media models than those discussed above.

4. Summary of results. In this section we characterize the wave pulse that has propagated through the multiscale medium. We assume the model (9) and in addition that the density ρ in (4) is constant. This allows us to characterize the travel time to a given depth. The general case is considered in (5.1). We give a more detailed account of the results and how they are derived in sections 5 and 6.

The pulse impinging on the half-space $z > 0$ has shape p_0, a compactly supported function. In the *random* medium the transmitted pulse at depth L can be characterized in terms of (i) $\chi_h(L)$, a *random* travel time correction, and (ii) $\mathcal{G}$, a *deterministic* pulse shaping function. The support of $\mathcal{G}$ is $\mathcal{O}(\sqrt{L})$. Let $\tau(L)$ be the travel time to depth $z = L$ in the deterministic (homogenized) medium; then we have the following result for the transmitted pulse.

LEMMA 4.1. *Let $1/4 < H \leq 1/2$ and $p(0,t) = p_0(t/h^{H+1/2})$ be the impinging pulse at the surface. Then for every $\varepsilon > 0$ and $M > 0$,*

$$\mathbb{P}\left(\sup_{|s|<M}\left|p(L,\tau(L)+\chi_h(L)+h^{H+1/2}s) - \int p_0(s-u)\mathcal{G}(u;L)\,du\right| > \varepsilon\right) \to 0 \quad \text{as} \quad h \to 0. \tag{13}$$

The random variable χ_h is a Gaussian random variable with magnitude $\mathcal{O}(h)$. Thus, when we observe the transmitted pulse in a randomly corrected time frame, we see a deterministic pulse in the small h limit. This is what we refer to as *stabilization*. If $H = 1/2$, then β_H is standard Brownian motion that has independent increments, corresponding to independent medium fluctuations. In this case the spreading of the pulse happens on the diffusion scale h, which is a measure of the correlation length of the medium fluctuations. Spreading on this scale corresponds to that discussed by O'Doherty and Anstey in [19]. If $H < 1/2$, the increments of β_H are *negatively* correlated and the medium fluctuations are rougher than in the standard Brownian case. In this case the pulse shaping is stronger and happens on the anomalous diffusion scale $h^{H+1/2}$.

Consider next the case of $H > 1/2$; now the increments of β_H are *positively* correlated and the medium fluctuations are smoother than in the standard Brownian case. The next result shows that in this case there is no change in pulse shape on the discretization scale in the small h limit.

LEMMA 4.2. *Let $H > 1/2$ and $p(0,t) = p_0(t/h)$ be the impinging pulse at the surface. Then for every $\varepsilon > 0$ and $M > 0$,*

$$\mathbb{P}\left(\sup_{|s|<M}|p(L,\tau(L)+\chi_h(L)+hs) - p_0(s)| > \varepsilon\right) \to 0 \quad \text{as} \quad h \to 0. \tag{14}$$

The travel time correction χ_h is characterized as in Lemma 4.1.

Assume that the source pulse is supported on the inner scale: $p_0 = p_0(t/h)$. In the Brownian case with $H = 1/2$ it follows from Lemma 4.1 that on the standard diffusion scale h we observe stabilization to a *fixed* pulse shape at the *fixed* depth L. If $H < 1/2$ with a stronger pulse shaping, this result generalizes in that we see stabilization on the scale h to a fixed pulse, but for a travel distance that *decreases* with h. Analogously, for $H > 1/2$, with a weaker pulse shaping, we observe stabilization on the scale h for a travel distance that *increases* with decreasing h. This follows from the next result.

LEMMA 4.3. *Let $1/4 < H < 3/4$ and $p(0,t) = p_0(t/h)$ be the impinging pulse at the surface. Then for every $\varepsilon > 0$ and $M > 0$,*

$$\mathbb{P}\left(\sup_{|s|<M}\left|p(Lh^{1-2H},\tau(Lh^{1-2H})+\chi_h(L)+hs) - \int p_0(s-u)\mathcal{G}(u;L)\,du\right| > \varepsilon\right) \to 0 \text{ as } h \to 0. \tag{15}$$

Thus, the coherent pulse front will be confined to an $\mathcal{O}(Lh^{1-2H})$ neighborhood of the surface. The random variable χ_h is a Gaussian random variable, now with magnitude $\mathcal{O}(h^{1+H(1-2H)})$.

5. Derivation of pulse shaping.

5.1. Dynamic equations for the pulse front. In order to derive the above results we need to characterize the evolution of the pulse front and how this relates to the fluctuations in the impedance ζ. Wave propagation in the discrete medium is determined by the interface reflection coefficients r_k^h and the transmission coefficients τ_k^h. These are defined by

$$r_k^h = \frac{\zeta_{k+1}^h - \zeta_k^h}{\zeta_{k+1}^h + \zeta_k^h}, \tag{16}$$

$$\tau_k^h = \sqrt{1 - |r_k^h|^2}. \tag{17}$$

Let $D^{\pm}$ and $U^{\pm}$ be the wave components in (5) evaluated immediately to the right and left of interface k at location $x_k = kh$. The interface corresponds to a jump in the characteristic impedance ζ, and the continuity of p and u gives the appropriate interface conditions that determine the associated jumps in D and U. These jumps correspond to some of the down-propagating energy being converted to the up-propagating mode and vice-versa. The interface conditions give (see [6])

$$\begin{bmatrix} D^+ \\ U^- \end{bmatrix} = \begin{bmatrix} \tau_k^h & -r_k^h \\ r_k^h & \tau_k^h \end{bmatrix} \begin{bmatrix} D^- \\ U^+ \end{bmatrix}. \tag{18}$$

We are interested in the impulse response of the medium, that is, how a down-propagating impulse at the surface is being transformed as it propagates. The impulse response is an analogue of the Green's function for the medium. The transmitted pulse when we probe the medium with a general *down*-propagating wave is easily found by convolution of the source wave with this impulse response. At the initial time we assume that the medium is at rest and that we probe it with a down-propagating impulse:

$$\begin{aligned} D_{t=0} &= \delta(x^+), \\ U_{t=0} &= 0. \end{aligned} \tag{19}$$

Wave reflections at the interfaces in the discrete medium lead to a set of down- and up-propagating impulses. At the time instances $t = ih$, with i integer, these impulses are located at the interfaces. The down-propagating pulses are separated by integer multiples of h in the time coordinate t, and also in the travel time coordinated, x, as are the up-propagating impulses. We find it convenient to represent these "impulse-trains" by the magnitude of the impulses indexed as D_j^i and U_j^i, with i being the time index and corresponding to times $t = ih$. The index j gives the distance from the front in the x dimension, measured in units of h. Thus, D_0^i is the magnitude of the first impulse in the down-propagating pulse-train at time $t = ih$. The initial condition (19) gives

$$\begin{aligned} D_j^0 &= \delta_0(j), \\ U_j^0 &= 0. \end{aligned} \tag{20}$$

In Figure 1 we illustrate the propagation of the impulses by a sequence of "snapshots" taken at times $t = 0, h$, and $2h$. The figure makes it clear that at time instances $t = ih$ only every second interface is associated with nonzero impulses and that the support of the pulse-trains increases with increasing time, giving pulse spreading. An important aspect of the parameterization is that a finite section of the wave front evolves autonomously and can be described independently of the tail part of the wave. We make use of this fact for the analysis of the problem and also for numerical simulation of the evolution of the wave front. Consider D_2^2 in the example given in Figure 1, that is, the magnitude of the second down-propagating impulse at time $2h$. It trails the leading impulse by two sections and is determined by a double scattering event associated with the initial impulse D_0^0. Part of the initial impulse is first reflected to an up-propagating mode and then aligned with D_2^2 through a second scattering event. The change in D_{2j}^i from one time step to the next can in general be expressed *exactly* in terms of double scattering events associated with down-propagating impulses "ahead" of it when these are evaluated at previous times. We next show how by unraveling the evolution seen in Figure 1.

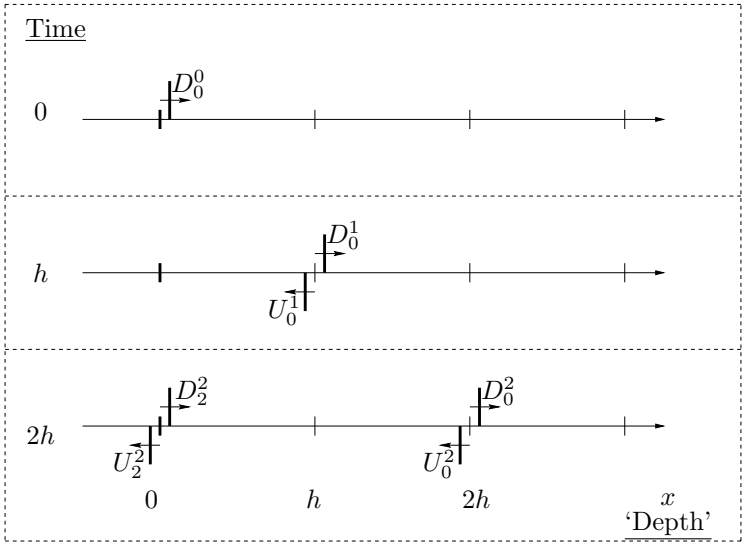


FIG. 1. *The generation of multiple reflections in the discrete medium.*

From (18) we find

$$\begin{bmatrix} D_j^{i+1} \\ U_j^{i+1} \end{bmatrix} = \begin{bmatrix} \tau_k^h & -r_k^h \\ r_k^h & \tau_k^h \end{bmatrix} \begin{bmatrix} D_j^i \\ U_{j-2}^i \end{bmatrix}, \tag{21}$$

with $k = i + 1 - j$. Define

$$d_j^i = \prod_{k=0}^{i-j} \tau_k^h \; D_j^i, \tag{22}$$

$$u_j^i = \prod_{k=0}^{i-j-1} \tau_k^h \; U_j^i,$$

with $\tau_k^h = 1$ for $k < 0$. This gives, using (18), that

$$d_j^{i+1} = d_j^i - r_k^h u_j^{i+1},$$
$$u_j^{i+1} = u_{j-2}^i + r_k^h d_j^i,$$

with $k = i + 1 - j$. Then, upon elimination of u_j^i, it follows in view of (20) that

$$\mathbf{d}^{i+1} = \mathbf{d}^i - \mathbf{A}_i^h \mathbf{d}^i, \tag{23}$$

with the vector $\mathbf{d}^i$ corresponding to the front part of the wave:

$$\mathbf{d}^i = [d_0^i, d_2^{i+1}, d_4^{i+2}, \ldots]'.$$

The matrix $\mathbf{A}_i^h = \{a_{k,l}^i\}$ is *lower triangular* with

$$a_{k,l}^i = r_{i-k+2}^h \; r_{i-l+2}^h \tag{24}$$

for $k \geq l$. In the next section we use (23) to obtain a characterization of the transmitted pulse. Note that (23) articulates how the change in a down-propagating impulse at a given time can be expressed exactly in terms of double scattering events associated with impulses ahead of it when these are evaluated at previous times. Thus, the statistics of products of reflection coefficients, corresponding to these double scattering events, will determine the evolution of the pulse shape.

5.2. Stabilization from averaging. We state the conditions and the result that describe the fascinating stabilization property of the down-propagating pulse when this is observed in a travel-time frame. With stabilization we mean that the transmitted pulse becomes essentially deterministic in the small h limit due to averaging in (23). Averaging in (23) means that we can replace $\mathbf{A}_i^h$ by its mean value, which is a lower triangular Toeplitz matrix with the entries on the ith subdiagonal being $E[r_m^h r_{m+i}^h]$, assuming here that the interface reflection coefficients form a stationary sequence. The following theorem generalizes and makes this precise.

Let $[\cdot]$ denote rounding to integer value, and define

$$\frac{d\mathcal{D}}{ds}(s,h) = -\bar{\mathbf{A}}(s,h)\; \mathcal{D}(s,h),$$
$$\mathcal{D}(0,h) = \mathbf{e}_1, \tag{25}$$

with $\bar{\mathbf{A}}(s,h)$ being a lower triangular Toeplitz matrix whose first column is

$$[a(0, s/g(h), h)/2, a(1, s/g(h), h), \ldots, a(K, s/g(h), h)]'$$

for some function a, and $\mathbf{e}_1$ a vector with one in the first entry and zero else; moreover,

$$\mathbf{D}(x,h) = [D_0^{[x/h]}, D_2^{[x/h]}, \ldots, D_{2K}^{[x/h]}]',$$
$$\mathbf{U}(x,h) = [U_0^{[x/h]}, U_2^{[x/h]}, \ldots, U_{2K}^{[x/h]}]', \tag{26}$$

with D_j^i and U_j^i as defined above. Then we have the following.

THEOREM 5.1. *If for all $\epsilon > 0$ and $\Delta \in \{0, 1, 2, \ldots\}$*

$$\lim_{h \to 0} \mathbb{P}\left[\sup_{0<s<L} \left| \sum_{m=1}^{[s/(g(h)h)]} r_m^h r_{m+\Delta}^h - \int_0^s a(\Delta, v/g(h), h)\; dv \right| > \epsilon \right] = 0, \tag{27}$$

where $0 < g(h)h = o(1)$ *and* $|a| < c$ *for some constant* c*, then for all* $\epsilon > 0$

$$\lim_{h\to 0} \mathbb{P}\left[\sup_{0<s<L} ||\mathbf{D}(s/g(h), h) - \mathcal{D}(s, h)|| > \epsilon\right] = 0, \tag{28}$$

$$\lim_{h\to 0} \mathbb{P}\left[\sup_{0<s<L} ||\mathbf{U}(s/g(h), h)|| > \epsilon\right] = 0. \tag{29}$$

The proof of this result is given in Appendix C. The formulation (25) follows from replacing $\mathbf{A}_i^h$ in (23) and the factor $\prod_{k=1}^m \tau_k^h$ in (22) by their corresponding averaged values. We apply the above result to fractal media in section 6.

The following lemma shows that the condition (27) entails that the interface reflection coefficients are small. Note, however, that this does not mean that the medium fluctuations themselves are relatively small.

LEMMA 5.2. *If for all* $\epsilon > 0$ *and* $\Delta \in \{0, 1, 2, \ldots\}$

$$\lim_{h\to 0} \mathbb{P}\left[\sup_{0<s<L} \left| \sum_{m=1}^{[s/(g(h)h)]} r_m^h r_{m+\Delta}^h - \int_0^s a(\Delta, v/g(h), h)\, dv \right| > \epsilon\right] = 0, \tag{30}$$

where $0 < g(h)h = o(1)$ *and* $|a| < c$ *for some constant* c*, then for all* $\epsilon > 0$

$$\lim_{h\to 0} \mathbb{P}\left[\sup_{1\le i\le [L/(g(h)h)]} |r_i^h| > \epsilon\right] = 0. \tag{31}$$

We prove this lemma in Appendix A.

For a given random medium model the following lemma gives a convenient way to check that the condition (27) is satisfied. Define

$$S^h(s, \Delta) = \sum_{m=1}^{[s/(g(h)h)]} r_m^h r_{m+\Delta}^h - \int_0^s a(\Delta, v/g(h), h)\, dv; \tag{32}$$

then we have the following.

LEMMA 5.3. *With* a *and* g *defined as in* (27)*, if there is an* $\alpha > 0$ *and a* $C > 0$ *such that for* $h < h_0$

$$\sup_{0<s<t<L} E[|S^h(t, \Delta) - S^h(s, \Delta)|^\alpha] \le g(h)h|t - s|C, \tag{33}$$

then the condition (27) *is satisfied.*

The proof of this lemma can be found in Appendix B.

Theorem 5.1 shows how the shape of the transmitted pulse is affected by the medium fluctuations. In the next section we give an interpretation of this modification and show that for an important class of random media models the spreading of the pulse in the random medium can be described as a convolution with a deterministic Gaussian pulse shape.

5.3. Pulse shape from the central limit theorem. Assume first that the interface reflection coefficients are stationary and that

$$E[r_i^h\ r_{i+\Delta}^h] = ha(\Delta).$$

It then follows from Theorem 5.1 that in probability $\lim_{h\to 0}\mathbf{D}(L,h)=\mathcal{D}(L)$ with

$$\mathcal{D}(L)=\exp(-L\bar{\mathbf{A}})\ \mathbf{e}_1=\exp(-La(0)/2)\exp(La(0)\mathcal{Q}/2)\ \mathbf{e}_1,$$

where $\bar{\mathbf{A}}$ and $\mathcal{Q}$ are lower triangular Toeplitz matrices whose first columns are

$$\mathbf{a}=[a(0)/2,\ a(1),\ a(2),\ldots]',$$
$$\mathbf{q}=-[0,\ 2a(1)/a(0),\ 2a(2)/a(0),\ldots]',$$

respectively. Note that multiplication with $\mathcal{Q}$ corresponds to a discrete convolution with its first column. Therefore

$$\mathcal{D}(L)=\sum_{n=0}^{\infty}p_n\mathbf{q}^{n*}\qquad\text{as}\ \ h\downarrow 0, \tag{34}$$

where $\mathbf{q}^{n*}$ denote n-fold convolution, $\mathbf{q}^0=e_1$, and where

$$p_n=\exp(-La(0)/2)(La(0)/2)^n/n!$$

is a discrete Poisson distribution. For typical media models, for instance, when ζ_k^h form a Markov process, the first column of $\mathcal{Q}$ defines a discrete probability distribution. Then $\mathcal{D}$ is the distribution of a *random sum*. A central limit theorem argument then gives that $\mathcal{D}$ is approximately a Gaussian pulse shape with standard deviation $\mathcal{O}(\sqrt{L})$ for L large. We show this in Appendix F, where we consider media with slowly varying media statistics.

6. Application to a fractal environment. In this section we consider the multiscale medium models introduced in sections 3.2 and 3.3. We show how Theorem 5.1 applies to these media and give the medium statistics that define the deterministic transformation in the shape of the propagating pulse. The results presented in section 4 follow via a transformation from the travel time coordinate to physical depth.

6.1. Fractional Brownian noise medium. We consider the medium model (9). A calculation involving the algebra of the moments of Gaussian random variables gives Lemma 6.1 below.

LEMMA 6.1. *Let ζ_k^h be defined by* (9) *and $1/4<H<3/4$; then there exists $h_0>0$ such that*

$$\lim_{h\to 0}E\left[\sum_{m=1}^{[s/g(h)h]}r_m^h r_{m+\Delta}^h\right]=sa(\Delta), \tag{35}$$

$$Var\left[\sum_{m=1}^{[s/g(h)h]}r_m^h r_{m+\Delta}^h\right]\le g(h)hs\sigma^4C(H)\quad \textit{for}\ \ h\le h_0, \tag{36}$$

with $g(h)=h^{2H-1}$ and for $\Delta\ge 1$

$$\begin{aligned}a(\Delta)&=-\Delta^{2H-4}\ (\sigma^2/8)\ \delta^4_{1/\Delta}[x^{2H}]_{x=1}\\ &\sim-\Delta^{2H-4}\ \sigma^2\ H(H-1/2)(H-1)(2H-3)\ \ \textit{as}\ \ \Delta\to\infty.\end{aligned} \tag{37}$$

In (37) we used the fourth order discrete central difference operator δ^4_Δ defined by

$$\delta_\epsilon[f(x)]=\frac{f(x+\epsilon/2)-f(x-\epsilon/2)}{\epsilon}.$$

Thus, the coefficients $a(\cdot)$ can be expressed in terms of a fourth order difference operator of the power law of the underlying fBm. This is related to the fact that these coefficients are means of products of interface reflection coefficients that themselves are obtained essentially by discrete differentiation of the impedance sequence. Lemmas 5.3 and 6.1 entail that the condition (27) in Theorem 5.1 is satisfied for the model (9) with $g(h) = h^{2H-1}$. Thus, in probability, the transmitted impulse response when evaluated at depth $\tilde{L} = Lh^{1-2H}$ satisfies

$$\lim_{h\to 0} \mathbf{D}(\tilde{L}/h, h) = \exp(-L\bar{\mathbf{A}})\ \mathbf{e}_1, \tag{38}$$
$$\lim_{h\to 0} ||\mathbf{U}(\tilde{L}/h, h)|| = 0,$$

where $\bar{\mathbf{A}}$ is a lower triangular Toeplitz matrix whose first column is

$$[a(0)/2,\ a(1),\ a(2), \ldots]$$

and $\mathbf{U}$ and $\mathbf{D}$ are defined in (26). Thus, when we probe the medium with a unit downgoing impulse at the surface, we observe the pulse shape defined by (38) at depth $\tilde{L}$. By a transformation of the independent variable to physical depth this entails that Lemmas 4.2 and 4.3 in section 4 are valid when the fractal medium is defined by (9). That Lemma 4.1 holds follows from Lemmas 5.3 and 6.1 and from Theorem 5.1 upon a transformation of the travel-time argument.

We next illustrate these results regarding the model (9) with numerical simulations. In the numerical simulations we use the initial condition (19) and propagate the pulse essentially according to (23). In practice we reformulate (23) to obtain an orthogonal propagation operator. In the figures we plot the down-propagating pulse $\mathbf{D}$ at the considered depth. Note that the origin in the plotted coordinate system corresponds to the front of the pulse, that is D_0^N, with N the total number of sections in the discrete medium. Thus, in the absence of random medium fluctuations we will see a unit impulse only at the origin. The random medium variations cause a spreading of the impulse.

First, we illustrate Lemma 4.3 using the medium model (9). In Figure 2 we use $\sigma = 5$, and the solid, dashed, and dotted lines correspond to $h_0 = 2^{-12}$, $h_1 = 2^{-14}$, and $h_2 = 2^{-16}$, respectively. The pulses are plotted at the scaled depth

$$\bar{x}(h_i) = \left(\frac{h_i}{h_0}\right)^{1-2H}.$$

The crosses give the stabilized pulse shape defined by (34). In the top plot we use $H = 0.4$, whereas in the bottom plot we use $H = 0.6$. As expected, we see stabilization to the theoretical pulse in both cases. Note that the horizontal axis is scaled by h. The pulse shaping is stronger in the antipersistent case with a rougher medium (top plot). The limiting pulse shape is close to the Gaussian pulse shape. This can be explained by the representation (34) of the impulse response. Recall that if the vector $\mathbf{q}$, the first column of $\mathcal{Q}$, is nonnegative, then (34) can be interpreted as the distribution of a random sum, and the impulse response will be close to the Gaussian pulse shape. The vector $\mathbf{q}$ is nonnegative for $H \geq 1/2$, giving in the small h limit a Gaussian pulse, as in the bottom plot. For $H < 1/2$ the sequence $\mathbf{q}$ is partly negative. In the top plot we used a value for the Hurst exponent that is slightly smaller than the Brownian case, $H = 0.4$, and the pulse shape is close to the Gaussian shape. Note that, since the pulses are plotted at depth $\propto h^{1-2H}$, we have to go shallower and shallower in the

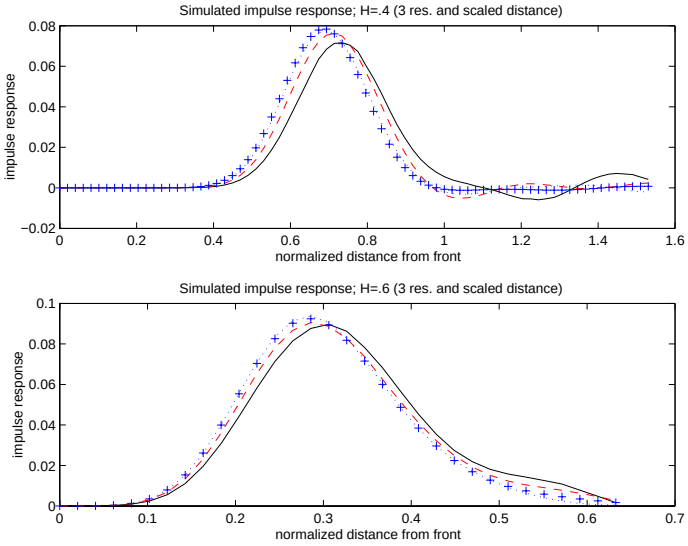


FIG. 2. *The impulse response of the fractal medium plotted at a depth that* scales *with the inner scale* h *as* h^{1-2H}. *The horizontal axis is scaled by* h. *In this frame we see stabilization to a fixed pulse in the small* h *limit. This limit is approximately a Gaussian pulse shape. In the top plot* $H = .4$, *and in the bottom* $H = .6$. *The medium model is the one defined in* (9). *The solid, dashed, and dotted lines correspond to* $h_0 = 2^{-12}$, $h_1 = 2^{-14}$, *and* $h_2 = 2^{-16}$, *respectively. The crosses give the theoretical pulse shapes.*

case $H < 1/2$ to see the stabilized pulse. This corresponds to an antipersistent fBm and to rough medium variations. In the persistent case with $H > 1/2$ and a smoother medium we have to go deeper and deeper into the medium to see the stabilized pulse. Only in the pure Brownian case with $H = 1/2$ do we observe the pulse stabilization at a *fixed* depth.

Next, we illustrate Lemmas 4.1 and 4.2 using the medium model (9). In Figure 3 we plot the same impulse responses as in Figure 2, only evaluated at the fixed depth $L = 1$. As expected, in the small h limit the impulse response in the case $H = .4 < 1/2$ (top plot) approaches a stabilized Gaussian pulse shape. Note that the impulse responses are plotted relative to the scale $h^{2H} = h^{.8}$. This is the scale at which the pulse shape stabilizes in the small h limit. The crosses give the limiting pulse shape and conform closely with the numerical simulations. The bottom plot shows the transmitted impulses when $H = .6 > 1/2$. Then the impulse response becomes close to a unit impulse for small h. The figure shows that the numerical impulse responses approach, albeit slowly, the unit impulse as h is reduced.

In Figure 4 we show how the impulse response depends rather sensitively on the value of the Hurst exponent H that gives the roughness of the medium. We use $h = 2^{-16}$, $L = 1$, $\sigma = 1$, and the model defined in (9). The solid curve corresponds to the Kolmogorov scaling law with $H = 1/3$. The dotted and dashed lines correspond, respectively, to a 20% increase/decrease in the Hurst exponent, giving less (respectively, more) spreading of the pulse.

6.2. The fractal case. Next, we let ζ_k^h be defined by the medium model in (12). Observe therefore that the medium fluctuations are strong and $\mathcal{O}(1)$, in contrast to the models considered above, where they were small. However, we will find that the interaction with the medium fluctuations can be characterized in a way similar to

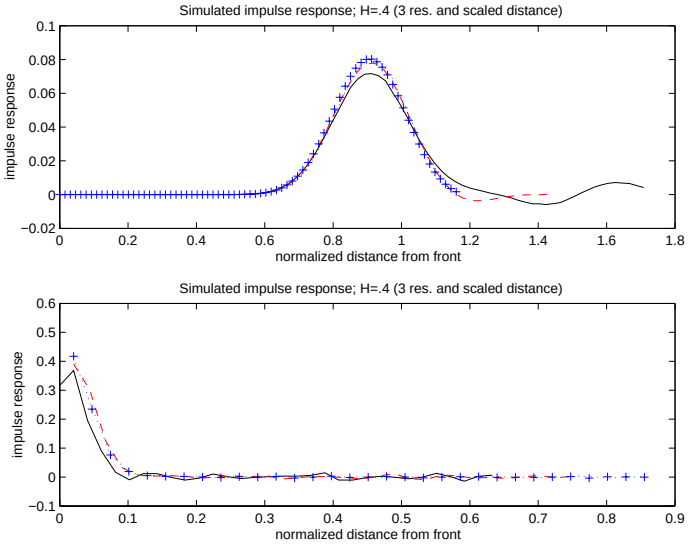


FIG. 3. *The impulse response of the fractal medium plotted at a* fixed *depth. The horizontal axis is scaled by* h^{2H}. *The top plot illustrates stabilization to a Gaussian pulse on the relative scale* $h^{2H} = h^{.8}$, *the bottom stabilization to the unit impulse. As above, in the top plot* $H = .4$, *and in the bottom* $H = .6$. *The medium model is the one defined in* (9). *The solid, dashed, and dotted lines correspond to* $h_0 = 2^{-12}$, $h_1 = 2^{-14}$, *and* $h_2 = 2^{-16}$, *respectively. The crosses give the theoretical pulse shapes corresponding to the smallest h value.*

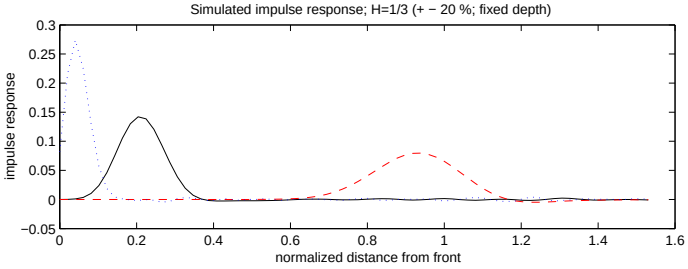


FIG. 4. *The impulse response of the fractal medium plotted at a* fixed *depth. It illustrates how the pulse shaping depends on the value of the Hurst exponent H that gives the roughness of the medium. The solid line corresponds to* $H = 1/3$, *and the dotted and dashed lines to a* 20% *increase (respectively, decrease). The medium model is the one defined in* (9).

that above. The following lemma can be shown by a generalization of the analysis that leads to Lemma 6.1.

LEMMA 6.2. *Let* ζ_k^h *be defined by* (12) *and* $1/4 < H < 3/4$; *then*

$$\lim_{h\to 0} E\left[\sum_{m=1}^{[s/g(h)h]} r_m^h r_{m+\Delta}^h\right] = sa(\Delta), \tag{39}$$

$$Var\left[\sum_{m=1}^{[s/g(h)h]} r_m^h r_{m+\Delta}^h\right] \le g(h)hs\sigma^4 C(H) \quad for \quad h \le h_0, \tag{40}$$

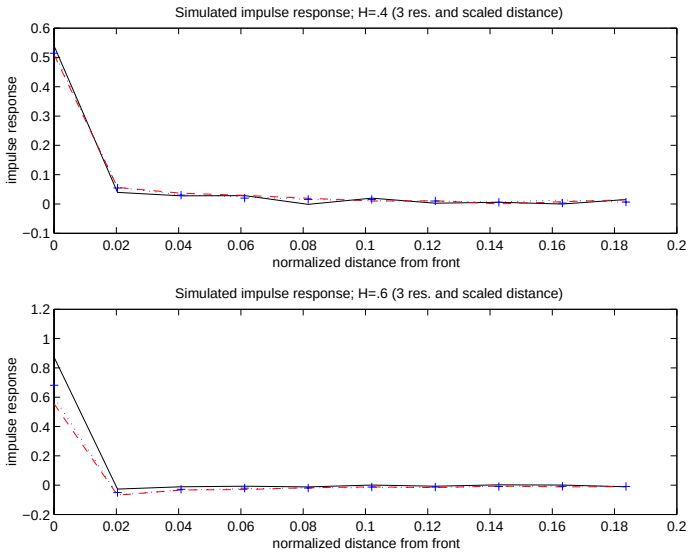


FIG. 5. *The figure corresponds to Figure* 2 *except that medium model* (12) *rather than* (9) *is used. The figure shows the impulse response of the fractal medium plotted at a depth that* scales *with the inner scale* h *as* h^{1-2H}*, and the horizontal axis is scaled by* h*. In this frame we see stabilization to a fixed pulse in the small* h *limit. In the top plot* $H = .4$*, and in the bottom* $H = .6$*. The solid, dashed, and dotted lines correspond to* $h_0 = 2^{-12}$*,* $h_1 = 2^{-14}$*, and* $h_2 = 2^{-16}$*, respectively.*

with $g(h) = h^{2H-1}$ *and for* $\Delta \geq 1$

$$a(\Delta) = \sigma^2 \Delta^{2H-2}\ \delta^2_{1/\Delta}[x^{2H}]_{x=1}/8 \sim \Delta^{2H-2}\ \sigma^2 H(H-1/2)/2 \qquad as \quad \Delta \to \infty. \tag{41}$$

Note that now the impedance is defined in terms of the fBm process itself rather than fBm noise, and that the $a(\cdot)$ coefficients thus are defined in terms of a second rather than fourth order difference operator. We show below that this has a strong effect on the impulse response. Lemma 6.2 entails that the condition (27) in Theorem 5.1 again is satisfied with $g(h) = h^{2H-1}$. Thus, in probability, the transmitted impulse response when evaluated at depth $\tilde{L} = Lh^{1-2H}$ satisfies (38), where $\bar{\mathbf{A}}$ is a lower triangular Toeplitz matrix whose first column is $\mathbf{a}' = [a(0)/2,\ a(1),\ a(2), \ldots]$, with $a(\cdot)$ now defined by (41).

Figure 5 corresponds to Figure 2 except that we used the model (12) with $\sigma = 1$. The impulse response is again plotted relative to the scale h and at depth

$$\bar{x}(h_i) = \left(\frac{h_i}{h_0}\right)^{1-2H}.$$

Again we observe stabilization in this frame. The solid, dashed, and dotted lines correspond to $h_0 = 2^{-12}$, $h_1 = 2^{-14}$, and $h_2 = 2^{-16}$, respectively. In the top plot $H = .4$, and in the bottom plot $H = .6$. The transformation of the pulse shape is weaker than above due to the smoother medium fluctuations. The crosses give the theoretical pulse shapes, and these conform closely with the numerical simulations for small h. In this case with a fractal medium the correlations decay more slowly than for the medium discussed in the previous section, as can be seen from (37) and

(41). The second moment associated with the discrete distribution $\mathbf{q}$, defined as the first column of $\mathcal{Q}$ in (34), is now unbounded, and the central limit theorem is not valid for this distribution. Thus, the pulse shape does not approach the Gaussian shape as it did above. Due to the long-range interactions in the medium fluctuations, the scattered wave energy is now spread far out and the coherent part of the pulse reduced in amplitude but not much in its shape.

Appendix A. Magnitude of medium fluctuations. We prove Lemma 5.2 stated in section 5. This lemma shows that the condition (27) entails that in the small h limit the interface reflection coefficients are small.

First, observe that the condition (27) allows us, for any given $\epsilon > 0$, to choose h_0 so small that for $h < h_0$

$$\mathbb{P}\left[\sup_{0<s<L}\left|\sum_{m=1}^{[s/(g(h)h)]}|r_m^h(\omega)|^2 - \int_0^s a(0,v/g(h),h)\ dv\right| > \epsilon^2/4\right] < \frac{\epsilon}{2}. \tag{42}$$

Now let $\{h^j\}$ be a sequence such that $\lim_{j\to\infty} h^j = 0$. Denote the associated array of interface reflection coefficients $r_i^{h^j}(\omega),\ 1 < i < N^j$, with

$$N^j = [L/(g(h^j)h^j)].$$

Assume that Lemma 5.2 is *false*. Then there is a subsequence $\{h^j\}$ of the above kind, a fixed $\epsilon > 0$, and a sequence of collections of disjoint sets $\{\mathcal{F}_i^j\}_{i=1}^{N_j}$ so that

$$\mathbb{P}\left[\bigcup_i \mathcal{F}_i^j\right] = \sum_{i=1}^{N_j}\mathbb{P}[\mathcal{F}_i^j] \ > \ \epsilon,$$

with

$$|r_i^{h^j}(\omega)| > \epsilon \ \text{ for } \ \omega \in \mathcal{F}_i^j. \tag{43}$$

We next show that this leads to a contradiction. Define

$$f(s;j) := S^{h_j}(s,0) = \sum_{m=1}^{[s/(g(h^j)h^j)]}|r_m^{h^j}(\omega)|^2 - \int_0^s a(0,v/g(h^j),h^j)\ dv.$$

Note that $f(s;j)$ has a jump discontinuity at

$$s_{(i;j)} := ig(h^j)h^j,$$

that is,

$$f(s^+_{(i;j)};j) - f(s^-_{(i;j)};j) = |r_i^{h^j}(\omega)|^2.$$

For $1 \le i \le N^j$ we therefore find

$$\sup_{0<s<L}\left|\sum_{m=1}^{[s/g(h^j)h^j]}|r_m^{h^j}(\omega)|^2 - \int_0^s a(0,v/g(h^j),h^j)\ dv\right| \ = \ \sup_{0<s<L}|f(s;j)|$$
$$\ge \ \frac{|f(s^+_{(i;j)};j) - f(s^-_{(i;j)};j)|}{2} \ = \ \frac{|r_i^{h^j}(\omega)|^2}{2},$$

and for $\omega \in \mathcal{F}_i^j$ it follows that

$$\sup_{0<s<L} \left| \sum_{m=1}^{[s/g(h^j)h^j]} |r_m^{h^j}(\omega)|^2 - \int_0^s a(0, v/g(h^j), h^j)\ dv \right| \ > \ \frac{\epsilon^2}{2}. \tag{44}$$

We can thus conclude that

$$\mathbb{P}\left[\sup_{0<s<L} \left| \sum_{m=1}^{[s/g(h^j)h^j]} |r_m^{h^j}|^2 - \int_0^s a(0, v/g(h^j), h^j)\ dv \right| > \frac{\epsilon^2}{4} \right] \geq \sum_{i=1}^{N_j} \mathbb{P}[\mathcal{F}_i^j] \ > \ \epsilon,$$

contradicting (42).

Appendix B. A stabilization criterion. In this appendix we prove Lemma 5.3 stated in section 5. To prove Lemma 5.3 we need to show that (33) implies that for $\epsilon > 0$ and $\delta > 0$ there is an h_0 such that for $h < h_0$

$$\mathbb{P}\left[\sup_{0<s<L} |S^h(s, \Delta)| \geq \epsilon \right] \leq \delta, \tag{45}$$

with S^h defined by (32):

$$S^h(s, \Delta) = \sum_{m=1}^{[s/(g(h)h)]} r_m^h r_{m+\Delta}^h - \int_0^s a(\Delta, v/g(h), h)\ dv. \tag{46}$$

Observe first that we can choose $\bar{h}$ so small that for $h < \bar{h}$

$$g(h)h \sup_{(\Delta, v, h)} a(\Delta, v, h) < \frac{\epsilon}{2} \tag{47}$$

since $|a|$ is bounded and $g(h)h = o(1)$. From (46) it follows that for $h < \bar{h}$ and i integer

$$\mathbb{P}\left[\sup_{0<s<L} |S^h(s, \Delta)| \geq \epsilon \right] \leq \mathbb{P}\left[\sup_{0 \leq i \leq [L/(g(h)h)]} |S^h(ig(h)h, \Delta)| \geq \frac{\epsilon}{2} \right].$$

Therefore, to show (45) we need to show that for h small enough:

$$\mathbb{P}\left[\sup_{0 \leq i \leq [L/(g(h)h)]} |S^h(ig(h)h, \Delta)| \geq \frac{\epsilon}{2} \right] \leq \delta. \tag{48}$$

In the rest of this section we suppress the dependence on Δ.

The result (48) follows from two bounds that we will derive below from (33). For i and j integers we have the following two bounds. First,

$$\frac{2L}{\bar{\Delta}} \sup_{0 \leq i \leq [L/\bar{\Delta}]} \mathbb{P}\left[\sup_{0 \leq j \leq [\bar{\Delta}/(g(h)h)]} |S^h(i\bar{\Delta} + jg(h)h) - S^h(i\bar{\Delta})| \geq \frac{\epsilon}{4} \right] \leq \frac{\delta}{2} \tag{49}$$

for

$$\bar{\Delta} = \min\left[\frac{(\delta/2)(\epsilon/4)^\alpha}{2LC}, L \right], \tag{50}$$

with the quantities involved being defined as in Lemma 5.3. Second,

$$\mathbb{P}\left[\sup_{0\le i\le[L/\bar{\Delta}]}|S^h(i\bar{\Delta})|\ge\frac{\epsilon}{4}\right]\le\frac{\delta}{2} \tag{51}$$

for

$$h\le\mathcal{H}\left[\frac{\bar{\Delta}(\delta/2)(\epsilon/4)^\alpha}{L(L+\bar{\Delta})C}\right], \tag{52}$$

with

$$\mathcal{H}(v)=\begin{cases}\infty & \text{if } \sup_h(g(h)h) < v,\\ \inf_h \mid [g(h)h>v] & \text{else.}\end{cases}$$

From (49) and (51) we can conclude that (48) is indeed satisfied for ϵ small if

$$h\le\mathcal{H}\left[\frac{(\delta/2)^2(\epsilon/4)^{2\alpha}}{4L^3C^2}\right],$$

and $\bar{\Delta}$ is chosen as in (50) because then

$$\begin{aligned}&\mathbb{P}\left[\sup_{0\le i\le[L/(g(h)h)]}|S^h(ig(h)h)|\ge\frac{\epsilon}{2}\right]\\ &\le\sum_{i=0}^{[L/\bar{\Delta}]}\mathbb{P}\left[\sup_{0\le j\le[\bar{\Delta}/(g(h)h)]}|S^h(i\bar{\Delta}+jg(h)h)-S^h(i\bar{\Delta})|\ge\frac{\epsilon}{4}\right]\\ &\quad+\mathbb{P}\left[\sup_{0\le i\le[L/\bar{\Delta}]}|S^h(i\bar{\Delta})|\ge\frac{\epsilon}{4}\right]\\ &\le\frac{2L}{\bar{\Delta}}\sup_{0\le i\le[L/\bar{\Delta}]}\mathbb{P}\left[\sup_{0\le j\le[\bar{\Delta}/(g(h)h)]}|S^h(i\bar{\Delta}+jg(h)h)-S^h(i\bar{\Delta})|\ge\frac{\epsilon}{4}\right]+\frac{\delta}{2}\ \le\ \delta.\end{aligned}$$

We now show (49). Define first the event

$$A(j;i)=\left[|S^h(i\bar{\Delta}+jg(h)h)-S^h(i\bar{\Delta})|\ge\frac{\epsilon}{4}\right].$$

Observe that then

$$\begin{aligned}&\mathbb{P}\left[\sup_{0\le j\le[\bar{\Delta}/(g(h)h)]}|S^h(i\bar{\Delta}+jg(h)h)-S^h(i\bar{\Delta})|\ge\frac{\epsilon}{4}\right]\\ &=\mathbb{P}\left[\bigcup_{j=1}^{[\bar{\Delta}/(g(h)h)]}A(j;i)\right]\ \le\ \sum_{j=1}^{[\bar{\Delta}/(g(h)h)]}\mathbb{P}[A(j;i)].\end{aligned}$$

Using Chebyshev's inequality and (33), we find

$$\mathbb{P}[A(j;i)]\le\frac{E[|S^h(i\bar{\Delta}+jg(h)h)-S^h(i\bar{\Delta})|^\alpha]}{(\epsilon/4)^\alpha}\ \le\ \frac{Cj(g(h)h)^2}{(\epsilon/4)^\alpha}.$$

Therefore, we can conclude

$$\mathbb{P}\left[\sup_{0\le j\le[\bar\Delta/(g(h)h)]}|S^h(i\bar\Delta+jg(h)h)-S^h(i\bar\Delta)|\ge\frac{\epsilon}{4}\right]$$
$$\le\frac{\bar\Delta}{g(h)h}\sup_{1\le j\le[\bar\Delta/(g(h)h)]}\mathbb{P}[A(j;i)]\ \le\ \frac{\bar\Delta}{g(h)h}\,\frac{C\bar\Delta g(h)h}{(\epsilon/4)^\alpha}\ =\ \frac{C\bar\Delta^2}{(\epsilon/4)^\alpha}.$$

Thus, (49) is satisfied for $\bar\Delta$ given as in (50).

Consider next showing (51). Note that

$$\mathbb{P}\left[\sup_{0\le i\le[L/\bar\Delta]}|S^h(i\bar\Delta)|\ge\frac{\epsilon}{4}\right]\le\left[\frac{L}{\bar\Delta}+1\right]\sup_{0\le i\le[L/\bar\Delta]}\frac{E[|S^h(i\bar\Delta)|^\alpha]}{(\epsilon/4)^\alpha}\le\left[\frac{L}{\bar\Delta}+1\right]\frac{CLg(h)h}{(\epsilon/4)^\alpha}.$$

Thus, (51) is satisfied if

$$\frac{(L+\bar\Delta)LCg(h)h}{\bar\Delta(\epsilon/4)^\alpha}\le\frac{\delta}{2}$$

or

$$g(h)h\le\left[\frac{\bar\Delta(\delta/2)(\epsilon/4)^\alpha}{L(L+\bar\Delta)C}\right],$$

which gives (52).

Appendix C. Stabilization. We prove Theorem 5.1 given in section 5. Let $\mathbf{X}_i^h$ satisfy

$$\begin{aligned}\mathbf{X}_{i+1}^h&=(\mathbf{I}-\mathbf{A}_i^h)\ \mathbf{X}_i^h,\\ \mathbf{X}_0^h&=\mathbf{I},\end{aligned}\tag{53}$$

with $\mathbf{A}_i^h$ defined in (23) and $\mathbf{X}\ \in\ \mathbb{R}^{[K]\times[K]}$. We show for all $\epsilon>0$

$$\lim_{h\to0}\mathbb{P}\left[\sup_{0<s<L}||\mathbf{X}^h_{[s/(g(h)h)]}-\mathbf{X}(s)||>\epsilon\right]=0,\tag{54}$$

where

$$\begin{aligned}\frac{d\mathbf{X}(s)}{ds}&=-\mathcal{A}(s,h)\ \mathbf{X}(s),\\ \mathbf{X}(0)&=\mathbf{I},\end{aligned}\tag{55}$$

and $\mathcal{A}$ is a lower triangular Toeplitz matrix whose first column is

$$[a(0,s/g(h),h),a(1,s/g(h),h),\dots,a(K,s/g(h),h)]'$$

and with the function a being defined as in (25).

In order to show (54) we introduce a continuous version of $\mathbf{X}_i^h$. Let $\mathbf{X}^h(s)$ satisfy

$$\begin{aligned}\frac{d\mathbf{X}^h(s)}{ds}&=-\mathcal{A}^h(s)\ \mathbf{X}^h(s),\\ \mathbf{X}^h(0)&=\mathbf{I},\end{aligned}$$

with

$$\mathcal{A}^h(s) = -\frac{1}{g(h)h}\ln(\mathbf{I} - \mathbf{A}_i^h) \quad \text{for} \quad (i-1)g(h)h \ \le \ s \ \le \ ig(h)h;$$

then $\mathbf{X}^h(ig(h)h) = \mathbf{X}_i^h$. (Note that, in view of Lemma 5.2, we can truncate the elements of $\mathbf{A}$.)

Next, define the residual

$$\tilde{\mathbf{X}}^h(s) = \mathbf{X}^h(s) - \mathbf{X}(s).$$

Making use of an integrating factor and that

$$\begin{aligned} \frac{d\mathbf{X}^{-1}(s)}{ds} &= \mathbf{X}^{-1}(s)\mathcal{A}(s,h), \\ \mathbf{X}^{-1}(0) &= \mathbf{I}, \end{aligned} \tag{56}$$

we find

$$\begin{aligned} \tilde{\mathbf{X}}^h(s) = &-\int_0^s \tilde{\mathcal{A}}^h(v)\ dv\ \mathbf{X}^h(s) - \int_0^s \mathbf{X}(s)\ \mathbf{X}^{-1}(v) \int_0^v \tilde{\mathcal{A}}^h(v')\ dv'\ \mathcal{A}^h(v)\ \mathbf{X}^h(v)\ dv \\ &+ \int_0^s \mathbf{X}(s)\ \mathbf{X}^{-1}(v)\ \mathcal{A}(v,h)\ \int_0^v \tilde{\mathcal{A}}^h(v')\ dv'\ \mathbf{X}^h(v)\ dv \end{aligned} \tag{57}$$

with

$$\tilde{\mathcal{A}}^h(v) = \mathcal{A}^h(v) - \mathcal{A}(v,h).$$

From (31) it follows that for all $\epsilon > 0$

$$\lim_{h\to 0} \mathbb{P}\left[\sup_{1\le i\le [L/(g(h)h)]} ||\mathbf{A}_i^h|| > \epsilon\right] = 0. \tag{58}$$

Moreover, from (27) it follows that for all $\epsilon > 0$

$$\lim_{h\to 0} \mathbb{P}\left[\sum_{i=1}^{[s/(g(h)h)]} |\mathbf{A}_i^h(k,l)| - 2\int_0^s a(0, s/g(h), h)ds > \epsilon\right] = 0, \tag{59}$$

with $\mathbf{A}_i^h(k,l)$ being the elements of the matrix $\mathbf{A}_i^h$. Given a $c_1 > 0$, we find using (58) that there exists $c_2 > 0$ such that for all $\delta > 0$ there exists $h_0 > 0$ so that

$$\mathbb{P}\left[\int_0^s ||\mathcal{A}^h(v)||dv > c_1\right] \le \mathbb{P}\left[\sum_{i=1}^{[s/(g(h)h)]} ||\mathbf{A}_i^h|| + c_2 \sup_j ||\mathbf{A}_j^h||||\mathbf{A}_i^h|| > c_1\right] + \delta$$

for $h \le h_0$. Therefore, using (58), we find that there exists $c_1 > 0$ such that

$$\lim_{h\to 0} \mathbb{P}\left[\int_0^s ||\mathcal{A}^h(v)||dv > c_1\right] = 0. \tag{60}$$

Note also that for some $c_3 > 0$

$$\int_0^s ||\mathcal{A}(v,h)||dv < c_3 \tag{61}$$

since the coefficients a are bounded. In view of (55), (56), and (61), we find $c_4 > 0$ such that

$$\max\{||\mathbf{X}^{-1}||, ||\mathbf{X}||\} < c_4, \tag{62}$$

and from (60) that there exists $c_5 > 0$ so that

$$\lim_{h\to 0} \mathbb{P}\left[\sup_{0\le s\le t} ||\mathbf{X}^h(s)|| > c_5\right] = 0. \tag{63}$$

We find using (58) that there exists $c_6 > 0$ such that for all $\delta > 0$ there exists $h_2 > 0$ so that

$$\begin{aligned}&\mathbb{P}\left[\sup_{0\le s\le L}\left\|\int_0^s \mathcal{A}^h(v) - \mathcal{A}(v,h)dv\right\| > \epsilon\right]\\ &\le \mathbb{P}\left[\sup_{0\le s\le L}\left\|\sum_{i=1}^{[s/(g(h)h)]} \mathbf{A}_i^h + \mathbf{B}\sup_{1\le i\le [L/(g(h)h)]}||\mathbf{A}_i^h|| - \int_0^s \mathcal{A}(v,h)dv\right\| > \epsilon\right] + \delta\end{aligned}$$

for $h \le h_2$ with $||\mathbf{B}|| < c_6$. The bound in (27) then gives

$$\lim_{h\to 0} \mathbb{P}\left[\sup_{0\le s\le L}\left\|\int_0^s \mathcal{A}^h(v) - \mathcal{A}(v,h)dv\right\| > \epsilon\right] = 0. \tag{64}$$

From (57) and the above it then follows that

$$\lim_{h\to 0} \mathbb{P}\left[\sup_{0\le s\le L} ||\tilde{\mathbf{X}}^h(s)|| > \epsilon\right] = 0,$$

and we have shown (54). Recall that

$$d_k^{N_s+k} = D_k^{N_s+k}\ \Pi_s^h$$

with $N_s = [s/(g(h)h)]$ and

$$\Pi_s^h = \prod_{m=1}^{[s/(g(h)h)]} \tau_m^h.$$

Finally, by using Lemma E.1, which gives the magnitude of Π_s^h, and the result of section D, which bounds the relative magnitude of the reflected mode, we obtain Theorem 5.1.

Appendix D. The reflected mode. In (5) we decompose the wavefield in terms of up- and a down-propagating wave components. Note that if the wave enters a homogeneous section with $\zeta(x) = \bar{\zeta}$ for $x > L$, then trivially the reflected wave component U vanishes for $x > L$, and the wave field is given in terms of the down-propagating wave component only. As we now show, the reflected wave component will be small in general.

Note first that from (21) it follows

$$U_{2j}^{i+1} = \sum_{n=0}^{j} r_{i+1-2(j-n)}^h \prod_{k=1}^{n} \tau_{i-1-2(j-k)}^h\ D_{2(j-n)}^{i-n}, \tag{65}$$

where

$$\prod_{k=1}^{0} := 1.$$

We parameterize the vector of wave components at the front by

$$\tilde{\mathbf{D}}_j^i = [D_0^{i-j}, D_2^{i-j+1}, \ldots, D_{2j}^i]'.$$

Observe that

$$\prod_{k=1}^{n} \tau_{i-1-2(j-k)}^h \leq 1.$$

Let $i \in \{0, \ldots, [L/(g(h)h)]\}$; then we find

$$U_{2j}^{i+1} \leq \left\{ \sup_{i+1-2j \leq k \leq i+1} |r_k^h| \right\} \, ||\tilde{\mathbf{D}}_j^i||.$$

From (19) and (21) we get the uniform bound

$$|D_j^i| \leq 1. \tag{66}$$

Thus, in view of (31), we find that for all $\epsilon > 0$ and $j \in \{0, \ldots, K\}$

$$\lim_{h \to 0} \mathbb{P} \left[\sup_{1 \leq i \leq [L/(g(h)h)]} |U_{2j}^i| > \epsilon \right] = 0$$

and that (29) is satisfied.

Appendix E. A bound on transmission. We show how the magnitude of the solution of (23) can be bounded. Recall that

$$d_k^{N_s+k} = D_k^{N_s+k} \, \Pi_s^h,$$

with $N_s = [s/(g(h)h)]$. Thus, in view of (66), we need to characterize the magnitude of Π_s^h.

LEMMA E.1. *The condition* (27) *implies that for all* $\epsilon > 0$

$$\lim_{h \to 0} \mathbb{P} \left[\sup_{0 < s < L} \left| \ln(\Pi_s^h) + \int_0^s a(0, v/g(h), h) \, dv/2 \right| > \epsilon \right] = 0, \tag{67}$$

with

$$\Pi_s^h = \prod_{m=1}^{[s/(g(h)h)]} \tau_m^h = \prod_{m=1}^{[s/(g(h)h)]} \sqrt{1 - |r_m^h|^2}. \tag{68}$$

Proof. Note first that we can write

$$\ln \sqrt{1 - |r_m^h|^2} = -\frac{|r_m^h|^2 (1 + v_m^h)}{2},$$

with

$$|v_m^h| \le |r_m^h|^2 \tag{69}$$

if $r_m^h < 1/2$. Observe next

$$\begin{aligned}
&\mathbb{P}\left[\sup_{0<s<L}\left|\ln(\Pi_s^h) + \int_0^s a(0, v/g(h), h)\ dv/2\right| > \epsilon\right] \\
&\le \mathbb{P}\left[\sup_{0<s<L}\frac{1}{2}\left|\sum_{m=1}^{[s/(g(h)h)]} |r_m^h|^2 - \int_0^s a(0, v/g(h), h)\ dv\right| + \frac{1}{2}\sup_m |v_m^h| \sum_{m=1}^{[s/(g(h)h)]} |r_m^h|^2 > \epsilon\right].
\end{aligned}$$

Let $\delta > 0$ be given; then from (27), (31), and (69) it follows that we can choose $h_0 > 0$ such that for $h \le h_0$

$$\mathbb{P}\left[\sup_{0<s<L}\frac{1}{2}\left\{\sup_m |v_m^h| \sum_{m=1}^{[s/(g(h)h)]} |r_m^h|^2\right\} > \frac{\epsilon}{2}\right] < \frac{\delta}{2}.$$

Therefore, for $h \le h_0$

$$\begin{aligned}
&\mathbb{P}\left[\sup_{0<s<L}\left|\ln(\Pi_s^h) + \int_0^s a(0, v/g(h), h)\ dv/2\right| > \epsilon\right] \\
&\le \frac{\delta}{2} + \mathbb{P}\left[\sup_{0<s<L}\frac{1}{2}\left|\sum_{m=1}^{[s/(g(h)h)]} |r_m^h|^2 - \int_0^s a(0, v/g(h), h)\ dv\right| > \frac{\epsilon}{2}\right].
\end{aligned}$$

The result (67) then follows from (27). □

Appendix F. Limiting pulse shape and the central limit theorem. Recall that in the small h limit the impulse response of the random medium is characterized by the solution of (25). The matrix $\bar{\mathbf{A}}$ in (25) is lower triangular and Toeplitz, and we find

$$\begin{aligned}
\mathcal{D}(T, h) &= \exp\left(-\int_0^T \bar{\mathbf{A}}(s, h)\ ds\right)\ \mathbf{e}_1 \\
&= \exp(-\lambda(T, h)) \exp(\lambda(T, h)\mathcal{Q}(T, h))\ \mathbf{e}_1
\end{aligned}$$

using a parameterization analogous to the one in section 5.3. Note that $\mathcal{Q}$ is a strictly lower triangular Toeplitz matrix and $\lambda(T, h) = \int_0^T \bar{\mathbf{A}}(s, h)_{(1,1)}\ ds$, where $\bar{\mathbf{A}}(s, h)_{(1,1)}$ is the main diagonal entry of the matrix $\bar{\mathbf{A}}(s, h)$. For some important random media models the first column of the matrix $\mathcal{Q}$ (denote it $\mathbf{q}$) has nonnegative entries and defines a discrete density supported on the nonnegative integers. This is the case if, for instance, the random medium is Markovian. We want to characterize $\mathcal{D}$:

$$\mathcal{D} = \mathcal{D}(\lambda) \ = \ \exp(-\lambda)\exp(\lambda\mathcal{Q}(\lambda))\ \mathbf{e}_1 \ = \ \sum_{k=0}^{\infty} \exp(-\lambda)\ \frac{\lambda^k}{k!}\ \mathbf{q}_\lambda^{k*} \tag{70}$$

in the limit of large $\lambda = \lambda(T, h) = \int_0^T \mathcal{A}(s, h)_{(1,1)}\ ds$, corresponding to large travel time depths. Note that in (70) we made use of the fact that multiplication with $\mathcal{Q}$ corresponds to a discrete convolution. This formulation shows that if indeed $\mathbf{q}_\lambda$

defines a discrete distribution, then we can regard $\mathcal{D}$ as the distribution of a random sum supported on the nonnegative integers. It follows that the pulse has constant area as it travels. By the formulas for the moments of a random sum (see [11]), we find that then the mean of $\mathcal{D}$ is $\lambda m(\lambda)$ and the variance $\lambda(v(\lambda)+m(\lambda)^2)$ when $m(\lambda)$ and $v(\lambda)$ are respectively the mean and variance associated with $\mathbf{q}_\lambda$. The next theorem shows that the normalized random sum converges in distribution to the standard normal; hence, the wave pulse attains the Gaussian shape as it penetrates deep into the medium.

LEMMA F.1. *Let $\mathbf{q}_\lambda$ define a discrete distribution with mean $m(\lambda) \le \overline{m}$ and variance $0 < \underline{v} \le v(\lambda) \le \overline{v}$, and let S_λ be distributed as a random sum according to*

$$\sum_{k=0}^{\infty} p_k^\lambda \mathbf{q}_\lambda^{k*},$$

with $p_k^\lambda = \exp(-\lambda)\lambda^k/k!$. Then, in the large λ limit,

$$X_\lambda = \frac{S_\lambda - \lambda m(\lambda)}{\sqrt{\lambda(v(\lambda)+m(\lambda)^2)}} \tag{71}$$

converges in distribution to a standardized zero mean normal random variable.

Proof. Let

$$\exp(im(\lambda)t)\phi_\lambda(t)$$

be the characteristic function of $\mathbf{q}_\lambda$. The characteristic function of $S_\lambda/\sqrt{\lambda}$ is then

$$\begin{aligned}&\exp\{-\lambda+\lambda[\exp(im(\lambda)t/\sqrt{\lambda})\ \phi_\lambda(t/\sqrt{\lambda})]\}\\ &= \exp\left\{-\lambda+\lambda\left[1+i\frac{m(\lambda)t}{\sqrt{\lambda}}-\frac{v(\lambda)t^2}{2\lambda}-\frac{m(\lambda)^2t^2}{2\lambda}\right]+o(1)\right\}\end{aligned}$$

for λ large. Thus, the characteristic function associated with X_λ defined in (71) is

$$\exp\left(\frac{-t^2}{2}+o(1)\right)$$

in the λ large limit. Hence, Lemma F.1 follows in view of Theorem 26.3 in [2]. ∎

Note also that the distribution of $S_\lambda/\sqrt{\lambda}$ can be approximated by the distribution of a random sum of Gaussian random variables. The following lemma gives this characterization of S_λ.

LEMMA F.2. *Let p_k^λ, $\mathbf{q}_\lambda$, and S_λ be defined as in Lemma* F.1. *Let E_X denote expectation with respect to the distribution of $X_\lambda = S_\lambda/\sqrt{\lambda}$, and E_g denote expectation with respect to the distribution defined by*

$$g_\lambda(x) = \sum_{k=0}^{\infty} p_k^\lambda\ \frac{\eta((x-\mu_k)/\sigma_k)}{\sigma_k}, \tag{72}$$

where $\eta(\cdot)$ is the standard normal distribution, $\mu_k = km(\lambda)/\sqrt{\lambda}$, $\sigma_k^2 = kv(\lambda)/\lambda$, and $m(\lambda)$ and $v(\lambda)$ are defined as in Theorem F.1. *Then*

$$\lim_{\lambda\to\infty}(E_g[u]-E_X[u]) = 0 \tag{73}$$

for every bounded continuous function u.

REFERENCES

[1] M. Asch, W. Kohler, G. Papanicolaou, M. Postel, and B. White, *Frequency content of randomly scattered signals*, SIAM Rev., 33 (1991), pp. 519–625.

[2] P. Billingsley, *Probability and measure*, John Wiley & Sons, New York, 1986.

[3] K.P. Bube and R. Burridge, *The one-dimensional inverse problem of reflection seismology*, SIAM Rev., 25 (1983), pp. 497–559.

[4] S.A. Bulgakov, V.V. Konotop, and L. Vazquez, *Wave interaction with a random fat fractal: Dimension of the reflection coefficient*, Waves Random Media, 5 (1995), pp. 9–18.

[5] R. Burridge, P. Lewicki, and G.C. Papanicolaou, *Pulse stabilization in a strongly heterogeneous layered medium*, Wave Motion, 20 (1994), pp. 177–195.

[6] R. Burridge, G.C. Papanicolaou, and B.S. White, *One dimensional wave propagation in a highly discontinuous medium*, Wave Motion, 10 (1988), pp. 19–44.

[7] J. Chillan and J.P. Fouque, *Pressure fields generated by acoustical pulses propagating in randomly layered media*, SIAM J. Appl. Math., 58 (1998), pp. 1532–1546.

[8] J.F. Clouet and J.P. Fouque, *Spreading of a pulse traveling in random media*, Ann. Appl. Probab., 4 (1994), pp. 1083–1097.

[9] D.J. Crossley and O.G. Jensen, *Fractal velocity models in refraction seismology*, Pure Appl. Geophys., 131 (1989), pp. 61–76.

[10] J. Feder, *Fractals*, Plenum Press, New York, 1988.

[11] W. Feller, *An Introduction to Probability Theory and Its Applications*, John Wiley & Sons, New York, 1971.

[12] J.P. Fouque and K. Sølna, *Time-reversal aperture enhancement*, SIAM J. Appl. Math, (2003), to appear.

[13] F. Herrmann, *A Scaling Medium Representation, A Discussion on Well-Logs, Fractals and Waves*, Ph.D. thesis, Delft University of Technology, Delft, The Netherlands, 1997.

[14] T.A. Hewett, *Modeling reservoir heterogeneities with fractals*, Proceedings of the 4th International Geostatistics Congress, J. European Union of Geosciences, Terra Abstracts, Suppl. 3, (1992).

[15] W. Kohler, G. Papanicolaou, and B. White, *Reflection and transmission of acoustic waves by a locally layered slab*, in Diffuse Waves in Complex Media, J.-P. Fouque, ed., Math and Physical Sciences Series 531, Kluwer, Dordrecht, The Netherlands, 1999, pp. 347–382.

[16] P. Lewicki, R. Burridge, and M.V. de Hoop, *Beyond effective medium theory: Pulse stabilization for multimode wave propagation in high-contrast layered media.*, SIAM J. Appl. Math., 56 (1996), pp. 256–276.

[17] V.V. Konotop, Z. Fei, and L. Vazquez, *Wave interaction with a fractal layer*, Phys. Rev. E, 48 (1993), pp. 4044–4048.

[18] B.B. Mandelbrot and J.W. Van Ness, *Fractional Brownian motions, fractional noises and applications*, SIAM Rev., 10 (1968), pp. 422–437.

[19] R.F. O'Doherty and N.A. Anstey, *Reflections on amplitudes*, Geophys. Prospecting, 19 (1971), pp. 430–458.

[20] G. Papanicolaou and K. Sølna, *Wavelet based estimation of Kolmogorov turbulence*, in Long-Range Dependence: Theory and Applications, P. Doukhan, G. Oppenheim, and M.S. Taqqu, eds., Birkhäuser-Boston, Cambridge, MA, 2002, pp. 473–505.

[21] M. Pilkington and J.P. Todoeschuck, *Stochastic inversion for scaling geology*, Geophys. J. Int., 102 (1990), pp. 205–217.

[22] P.G. Richards and W. Menke, *The apparent attenuation of a scattering medium*, Bull. Seism. Soc. Amer., 73 (1983), pp. 1005–1021.

[23] L.C.G. Rogers, *Arbitrage with fractional Brownian motion*, Math. Finance, 7 (1997), pp. 95–105.

[24] K. Sølna, *Focusing of time-reversed reflections*, Waves Random Media, 12 (2002), pp. 365–385.

[25] K. Sølna and G. Papanicolaou, *Ray theory for a locally layered medium*, Waves Random Media, 10 (2000), pp. 155–202.

[26] X. Sun and D.L. Jaggard, *Wave interaction with a generalized Cantor bar fractal multilayer*, J. Appl. Phys., 70 (1991).

SIAM J. APPL. MATH.
Vol. 63, No. 5, pp. 1789–1830

ORTHONORMAL POLYNOMIAL WAVELETS ON THE INTERVAL AND APPLICATIONS TO THE ANALYSIS OF TURBULENT FLOW FIELDS*

J. FRÖHLICH† AND M. UHLMANN‡

Abstract. The paper presents an orthogonal wavelet basis for the interval using a linear combination of Legendre polynomials. Expansion coefficients are taken as appropriate roots of the Chebyshev polynomials of the second kind. The new transform is first applied to analytical data, and appropriate definitions of a scalogram are presented. The transform is then extended to the multidimensional case, finding the tensor-product construction more appropriate than the multiresolution. The new method offers the possibility to determine meaningful spectra for signals on bounded domains which are constructed in a global and in a local fashion. Analyses of one- and two-dimensional data from a direct numerical simulation of turbulent channel flow are presented and demonstrate the potential of the method.

Key words. wavelets, Legendre polynomials, bounded domains, turbulence, turbulent channel flow

AMS subject classifications. 42C40, 42C20, 76F40

DOI. 10.1137/S0036139902404116

1. Introduction.

Since their advent in the 1980s, wavelets have been put to use in many different scientific areas. Applications to turbulent flows have been amongst the first and are comprehensively reviewed in [10]. Recent activity in this area includes analysis of the flow field in space by means of either continuous or discrete transforms [3, 26], data compression using orthogonal schemes [28], and the discretization of the governing Navier–Stokes equations in terms of wavelet functions [13]. Here, we are concerned with the first of these tasks: data analysis. This is motivated by the need for an improved understanding of local nonlinear transfer of turbulent kinetic energy in scale space.

While wavelet analysis has proven to be a valuable tool for investigating spatially homogeneous configurations, i.e., flow in periodic domains [3, 26, 7, 22], there are only very few publications dealing with bounded flows. One such example is reference [9], where, however, only planes parallel to the nonhomogeneous spatial direction are investigated. We believe that this situation is at least partially due to the lack of wavelet functions specifically designed for use on bounded intervals which are fully appropriate for the analysis. A proposal for such a construction is made in the present paper.

Fischer and Prestin [11] have developed a general method for constructing wavelet bases on the interval, starting from a set of orthogonal polynomials and making use

*Received by the editors March 15, 2002; accepted for publication (in revised form) January 20, 2003; published electronically July 26, 2003.
http://www.siam.org/journals/siap/63-5/40411.html

†Institute for Hydromechanics, University of Karlsruhe, 76128 Karlsruhe, Germany. Current address: Institute for Chemical Technology, University of Karlsruhe, 76128 Karlsruhe, Germany (froehlich@ict.uni-karlsruhe.de). This author's research was funded through the joint DFG-CNRS programme "Numerical Flow Simulation."

‡Potsdam Institute for Climate Impact Research, 14412 Potsdam, Germany. Current address: Departamento de Combustibles Fósiles, CIEMAT, Avda. Complutense 22, 28040 Madrid, Spain (markus.uhlmann@ciemat.es). This author's research was supported through the DFG-funded project KL 611-10.

of their reproducing kernel property. However, these authors present only one specific basis—constructed from Chebyshev polynomials of the second kind—which is orthogonal. These wavelets are related to a scalar product weighted by the function $w(x) = (1-x^2)^{1/2}$, and orthogonality is obtained with respect to this weighted product. For the purpose of data analysis, any weight other than unity is undesirable since the interpretation of coefficient values with respect to their energy contribution turns out to be nonintuitive. Prestin (in a private communication) proposed a modification to the original construction in which a "hybrid" basis would instead be built from a combination of Chebyshev and Legendre polynomials, thereby carrying over the orthogonality of the original pure Chebyshev basis to a weight function of unity.

In the present paper we first discuss the specific requirements a wavelet basis should fulfill in order to be useful for the purpose of analysis of turbulent flows. This is followed by a concise overview of related constructions developed in the literature. The new construction is presented in section 3 and applied to different types of analytical signals. This leads to important issues, such as the visual presentation of the coefficients (scalogram) and the study of the local power spectral density, for which we propose appropriate definitions in section 4. In a further step, the construction is extended to the multidimensional case (section 5). Applications of the method to data from turbulent plane channel flow in section 6 give an impression of its potential for the analysis of nonperiodic turbulent data.

2. Requirements and previous constructions.

2.1. Requirements. The term wavelet is often used in a very broad sense and can designate functions used in quite different multiscale methods. Features of such schemes are (i) a certain number of vanishing moments reflected by an oscillatory nature of the functions; (ii) localization in space; (iii) translational invariance; (iv) localization in frequency; (v) a rescaling mechanism. In practice, some of these properties are often watered down due to practical constraints. Since compromises need to be made it is important to fix desired properties a priori according to the needs of a target application. This issue is briefly discussed in the following, considering for notational ease the one-dimensional case. The term *frequency* will be used when referring to a Fourier basis while the term *scale* is employed in a more general sense.

Two- and three-dimensional data sets from turbulent flows tend to be extremely large. A redundant representation of these data can increase their size considerably and pose problems of computation time and storage requirements. A discrete, nonredundant representation therefore seems indispensable [26], and the continuous transform will not be considered in the present context.

The long-range goal of our research is the investigation of energy transfer mechanisms in turbulent flows. One approach to perform this is to represent a turbulent signal $f(x) \in L_2(\Omega)$, where Ω is the spatial domain, through an orthonormal set of basis functions $\beta_\lambda(x) \in L_2(\Omega)$, viz.

$$f(x) = \sum_\lambda a_\lambda \, \beta_\lambda(x). \tag{1}$$

The orthonormality property then allows us to decompose the energy E into contributions related to each basis function as

$$E = \frac{1}{|\,\Omega\,|} \int_\Omega f^2(x)\, dx \; = \; \sum_\lambda a_\lambda^2. \tag{2}$$

In mathematical terms, the L^2 norm of f is represented by the expansion coefficients with respect to the basis β_λ, according to the Plancherel identity. The scalar product generating this norm is denoted $\langle \cdot, \cdot \rangle$. In the next step, the representation (1) can be inserted into the governing Navier–Stokes equations to obtain equations for the evolution of the coefficients a_λ and hence equations for the energy transfer [26].

Turbulent flows exhibit motions over a wide range of scales. If the basis functions β_λ are not very smooth, they contain high-frequency contributions and are not well localized in the upper frequency range. On the other end of the spectrum, the number of vanishing moments determines the localization in the low-frequency range. If the number of these vanishing moments is insufficient, this can yield a decay of the wavelet spectrum which differs significantly from that of the analogous Fourier spectrum, which is an undesirable feature [30].

For the physical interpretation of the transform (1) and quantities derived from it, a pronounced asymmetry of the basis functions also is undesirable. Daubechies' wavelets, e.g., which have compact support, exhibit such an asymmetry. It can be alleviated to some extent, but compact support and symmetry exclude each other if orthogonality is required [5], except for the Haar basis, which is not smooth.

To sum up, we require vanishing moments, orthonormality, smoothness, and some sort of translational invariance and symmetry for the wavelets to be employed. The last two notions have to be relaxed for a basis on the interval as discussed below. Multiresolution algorithms often yield fast numerical schemes due to recursions over the refinement level. This last issue will be disregarded in the present paper and postponed to later work.

Finally, it should be stressed that a nonperiodic transform of turbulent data is important for signals in space rather than in time. The latter can always be made very long—provided the flow is statistically stationary—such that end effects are removed by an appropriate windowing. Signals in space, on the other hand, are often limited by the geometry of the flow. Moreover, physically interesting features frequently develop in direct proximity of the boundaries. A prototype case is the turbulent plane channel considered below.

2.2. Real line and periodic case. For later reference we recall some constructions for the real line. Here, it is convenient to work in Fourier space defined by the Fourier transform

$$\widehat{f}(\omega) = \int f(x)\, e^{-2\pi i x \omega} dx. \tag{3}$$

The trigonometric functions are not localized in space and are maximally localized in frequency. In order to design basis functions with localization in space, as well as frequency, localization in frequency is sacrificed to some extent by lumping together basis functions with neighboring frequencies. In the following we denote a wavelet by ψ and a scaling function by φ, using j and i as scale and shift indices, respectively. In the classical case this leads to a dyadic set of wavelet functions with

$$\psi_{ji} = 2^{j/2}\, \psi(2^j x - i), \tag{4}$$

and analogously for φ_{ji}. The Shannon wavelet, also termed the Littlewood–Paley basis, [5] is obtained by selecting only frequencies in the band $\omega \in [1/2, 1]$:

$$\widehat{\psi}^S(\omega) = \begin{cases} 1 & \text{if} \quad \mid \omega \mid \in [\frac{1}{2}, 1], \\ 0 & \text{else.} \end{cases} \tag{5}$$

Fourier theory immediately yields the asymptotic decay rate of $\psi \propto 1/x$ in space. Better decay is possible only with higher smoothness in frequency space. Meyer wavelets are constructed by an ingenious incorporation of neighboring decades using a blending function

$$\widehat{\psi}^M(\omega) = \begin{cases} e^{-i\omega x}\, S(3 \mid \omega \mid -1) & \text{if} \quad \mid \omega \mid \in [\frac{1}{3}, \frac{2}{3}], \\ e^{-i\omega x}\, S(2 - \frac{3}{2} \mid \omega \mid) & \text{if} \quad \mid \omega \mid \in [\frac{2}{3}, \frac{4}{3}], \\ 0 & \text{else}, \end{cases} \tag{6}$$

where S is a cosine-based smooth function with $S(0) = 0, S(1) = 1$, and $S^2(t)+S^2(1-t) = 1$ for $t \in [0;1]$ [5]. The smoother the blending function, the faster the asymptotic decay of the wavelets in space. It is polynomial for this family if S is not in C^∞. Other constructions such as spline wavelets have noncompact support in frequency space and decay exponentially in physical space. Wavelets on different scales are obtained by multiplying ω by a power of two, introducing a logarithmic decomposition of the frequency axis. But whether it is decomposed into logarithmic segments or some other intervals, the asymptotic decay rate in physical space is unaltered if the regularity in Fourier space is maintained.

A periodic basis, i.e., a basis on the circle $\mathbb{T} = [0, 1]$, can be obtained from a wavelet basis on the real line by restricting the frequency to integer values $\widehat{f}_k = \widehat{f}(\omega = k), k \in \mathbb{Z}$. This introduces a coarsest scale represented by $\widehat{f}_0$, i.e., the constant function. When considering the periodic case, asymptotic decay in space refers to the behavior in the limit $j \to \infty$.

2.3. Orthogonal wavelets on the interval. By functions on the interval (here $I = [-1, 1]$ for later convenience) we understand that in contrast to the circle $\mathbb{T}$ no periodicity is imposed.

Embedding the interval in a larger periodic domain or the real line by padding with zero and using standard transforms on the larger domains usually creates artifacts at the boundaries of the interval due to the generation of a strong singularity. This is very inconvenient for the analysis.

On the interval, translational invariance therefore has to be relaxed in some way. In addition to the length scale introduced by the scale index j of the wavelet, the distance to the nearest boundary unavoidably appears as a second length scale contradicting complete shift invariance. This fact has to be reflected to some extent by the construction, and translational invariance can therefore be realized only in some relaxed sense. For the Daubechies wavelets adapted to the interval [4] this is performed through modification of those translates touching the boundaries. Due to the orthogonalization procedure, the modified functions tend to take a quite irregular shape. Furthermore, a minimal refinement level is required to separate the regions of modification close to both boundaries. According to the above requirements we therefore do not use this construction for the present application.

2.4. Using the Chebyshev transform. A very elegant method to turn a periodic wavelet basis into a basis on the interval I is the mapping

$$x = \cos\theta, \qquad \theta \in [0, \pi]. \tag{7}$$

This is the mapping relating cosines and Chebyshev polynomials through $T_n(x) = cos(n\theta)$. In fact, if $\beta_\lambda^{\mathbb{T}}$ is a basis of periodic functions on $\mathbb{T}$, orthogonal with respect

to the scalar product

$$\langle \cdot, \cdot \rangle_{\mathbb{T}} = \int_0^1 \cdot \; \cdot \, dx, \tag{8}$$

the functions

$$\beta_\lambda^I(\theta) = \beta_\lambda^{\mathbb{T}}(\theta) + \beta_\lambda^{\mathbb{T}}(1-\theta) \tag{9}$$

constitute an orthogonal basis for the weighted scalar product

$$\langle \cdot, \cdot \rangle_w = \int_{-1}^1 \cdot \; \cdot \, w(x)\, dx \tag{10}$$

with the Chebyshev weight $w(x) = 1/\sqrt{1-x^2}$ [19, 31]. In fact, this was already announced and used in [24], where unfortunately the second entry of the sum in (9) was overlooked. The transform (7), (9) allows us to map an arbitrary periodic basis onto a basis for the interval I. Computations can then be done with the classical algorithms (employing fast convolution by FFT for long filters) requiring only a rescaling of the abscissa. The price, however, is the introduction of the Chebyshev weight.

For the reasons mentioned above we require a basis which is orthonormal with respect to the unweighted scalar product. A naive way of achieving this is to incorporate the weight into the basis by defining

$$\tilde{\beta}_\lambda^I = \sqrt{w(x)}\beta_\lambda^I \tag{11}$$

and using $\tilde{\beta}_\lambda^I$ instead of $\beta_\lambda^{\mathbb{T}}$ in (1). Approximations of square-integrable functions on I using this basis with finer and finer scale converge in an integral sense. Pointwise convergence at the interval boundaries, however, is destroyed by the singularities of the weight at ± 1 which are introduced into the basis. This is illustrated in Figure 1, where a very smooth function is approximated by a truncated series of such wavelets based on periodic spline wavelets. Hence, this approach is not useful for the present purpose.

2.5. Wavelets based on Jacobi polynomials. In [11] the authors have constructed wavelets based on orthogonal polynomials P_n by means of the reproducing kernel polynomials. The construction is possible for general orthogonal polynomials, but in the present context we take P_n to be the Jacobi polynomial of degree n defined on the interval $I = [-1, 1]$ and orthogonal with respect to the weighted scalar product (10) with $w(x) = w^{\alpha,\beta}(x) = (1-x)^\alpha(1+x)^\beta$. (The indices α, β are dropped from now on for convenience.) The reproducing kernel polynomial is

$$K_n(x, y) = \sum_{k=0}^n P_k(x)\, P_k(y). \tag{12}$$

Scaling functions are defined as

$$\varphi_{n,i} = K_n(x, y_i^{(n+1)}), \qquad i = 0, \ldots, n, \tag{13}$$

with a suitable set of points $y_i^{(n+1)}$, e.g., the zeros of the polynomial P_{n+1}. Wavelets are constructed as

$$\psi_{n,i} = K_{2n}(x, y_i^{(n)}) - K_n(x, y_i^{(n)}), \qquad i = 0, \ldots, n-1. \tag{14}$$

Based on this approach a multiresolution analysis (MRA) [25] of nested subspaces $V_j \subset V_{j+1}$ with difference spaces W_j can be generated. Orthogonality of all transla-

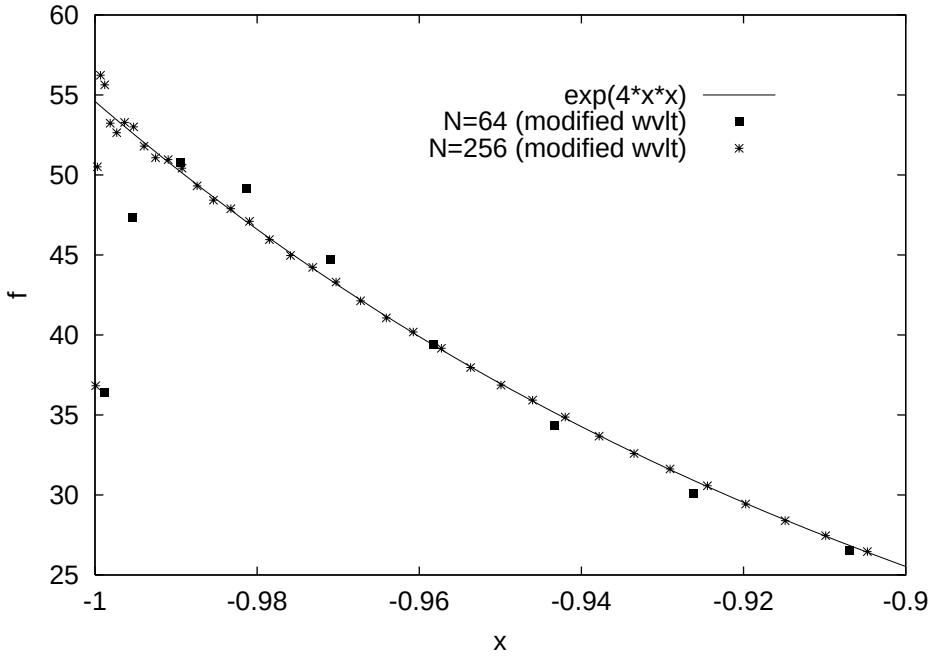


FIG. 1. *The reconstructed signal when approximating the function* $u(x) = \exp(-4x^2)$ *with the modified wavelets* $\tilde{\beta}$ *defined in* (11). *Two different resolutions are used. The first employs a grid with* $N = 64$ *points, using* 3 *of the computed scales for reconstruction; the second employs* $N = 256$ *points, using* 6 *scales for reconstruction.*

tions with respect to the related scalar product (including the Jacobi weight) is not necessarily obtained but can be investigated as described in [11]. In this reference one example of an orthogonal basis is presented using the Chebyshev polynomials of the second kind U_n. The same construction, when applied directly to Legendre polynomials L_n, however, does not yield an orthogonal basis.

3. New construction.

3.1. The general setting. In [11] linear combinations of orthogonal polynomials $P_k(x)$ are used for constructing scaling functions and wavelets by regrouping low-order and high-order polynomials via

$$\varphi_{ji}(x) = \sum_{k=0}^{2^j} a_{jik}\, P_k(x)\,, \tag{15a}$$

$$\psi_{ji}(x) = \sum_{k=2^j+1}^{2^{j+1}} b_{jik}\, P_k(x)\,. \tag{15b}$$

For wavelets and scaling functions spanning an orthonormal basis of an MRA, these need to fulfill the following orthogonality conditions:

$$\langle \varphi_{ji}, \varphi_{jl} \rangle_w = \delta_{il}\,, \tag{16a}$$

$$\langle \psi_{ji}, \psi_{ml} \rangle_w = \delta_{il}\, \delta_{jm}\,, \tag{16b}$$

$$\langle \varphi_{ji}, \psi_{ml} \rangle_w = 0 \qquad (m \geq j)\,. \tag{16c}$$

Introducing the ansatz (15), we obtain

$$\langle \varphi_{ji}, \varphi_{jl} \rangle_w = \sum_{k=0}^{2^j} \sum_{n=0}^{2^j} a_{jik} a_{mln} \langle P_k, P_n \rangle_w \,, \tag{17a}$$

$$\langle \psi_{ji}, \psi_{ml} \rangle_w = \sum_{k=2^j+1}^{2^{j+1}} \sum_{n=2^m+1}^{2^{m+1}} b_{jik} b_{mln} \langle P_k, P_n \rangle_w \,, \tag{17b}$$

$$\langle \varphi_{ji}, \psi_{ml} \rangle_w = \sum_{k=0}^{2^j} \sum_{n=2^m+1}^{2^{m+1}} a_{jik} b_{mln} \langle P_k, P_n \rangle_w \,. \tag{17c}$$

It is clear that due to the orthogonality of the polynomials (i.e., $\langle P_k, P_n \rangle_w = \delta_{kn}$) the choice of the coefficients a, b alone determines the orthogonality properties of the basis. Therefore, we find

$$\langle \varphi_{ji}, \varphi_{jl} \rangle_w = \sum_{k=0}^{2^j} a_{jik}\, a_{jlk} \,, \tag{18a}$$

$$\langle \psi_{ji}, \psi_{ml} \rangle_w = \delta_{jm} \sum_{k=2^j+1}^{2^{j+1}} b_{jik}\, b_{jlk} \,, \tag{18b}$$

$$\langle \varphi_{ji}, \psi_{ml} \rangle_w = 0 \qquad\qquad (m \geq j) \,. \tag{18c}$$

(The factor δ_{jm} in relation (18b) follows from the fact that the bounds of the two sums in (17b) need to be equal if the scalar product is to be nonzero.) As a consequence it is possible to interchange freely the particular type of polynomial—amongst the class of orthogonal ones—without changing the above properties. Therefore, we can go about and modify a given basis whose a's and b's are such that (16) is verified and replace the functions $P_k(x)$ and the weight w with any other orthogonal function and associated weight, in particular with Legendre polynomials related to the weight $w(x) = 1$. The construction will now be described in detail.

3.2. Definition of the basis functions. We define wavelets and scaling functions based upon Legendre polynomials $L_k(x)$ and coefficients related to the Chebyshev polynomials of the second kind $U_k(x)$ as follows:

$$\varphi_{ji}(x) = C_{ji}^{\varphi} \sum_{k=0}^{2^j} U_k(y_i^{(2^j+1)}) \sqrt{k+1/2}\, L_k(x) \,, \qquad j = 0, 1, \ldots, i = 0 \ldots 2^j \,, \tag{19a}$$

$$\psi_{ji}(x) = C_{ji}^{\psi} \sum_{k=2^j+1}^{2^{j+1}} U_k(y_i^{(2^j)}) \sqrt{k+1/2}\, L_k(x) \,, \qquad j = 0, 1, \ldots, i = 0 \ldots 2^j - 1 \,. \tag{19b}$$

With the above index bounds these functions span the scale spaces V_j and the difference spaces W_j of an MRA. This MRA fulfills the classical requirements of [25]. The

polynomials $L_k(x)$ and $U_k(x)$ on the interval $x \in [-1,1]$ can be defined by [1]:

$$L_k(x) = \frac{1}{2^k} \sum_{l=0}^{\mathrm{int}(k/2)} (-1)^l \begin{pmatrix} k \\ l \end{pmatrix} \begin{pmatrix} 2k-2l \\ k \end{pmatrix} x^{k-2l}\,, \tag{20a}$$

$$U_k(x) = \frac{\sin((k+1)\,\arccos(x))}{\sin(\arccos(x))}\,, \tag{20b}$$

(where $\mathrm{int}(r)$ is the largest integer less or equal r). For an efficient computation their three-term recursion formulae

$$L_{k+1}(x) = \frac{2k+1}{k+1} x\, L_k(x) - \frac{k}{k+1}\, L_{k-1}(x)\,, \qquad L_0(x) = 1\,, \quad L_1(x) = x\,, \tag{21a}$$

$$U_{k+1}(x) = 2x\, U_k(x) - U_{k-1}(x)\,, \qquad U_0(x) = 1\,, \quad U_1(x) = 2x \tag{21b}$$

are used. The parameters $y_i^{(n)}$ in equations (19) are the zeros of the nth-order Chebyshev polynomial of the second kind, i.e.,

$$y_i^{(n)} = -\cos\left(\frac{(i+1)\,\pi}{n+1}\right)\,, \qquad i = 0 \ldots n-1. \tag{22}$$

For convenience, the present numbering is different from the standard numbering in that we have $y_i^{(n)} < y_{i+1}^{(n)}$. Equations (19) define the coefficients a_{ijk} and b_{ijk} in (15). In [11] the orthogonality of the resulting basis with $P_k = U_k$ in (15) is proved. As discussed above, this property carries over to the functions defined by (19). The factors

$$C_{ji}^{\varphi} = \sqrt{\frac{2}{2^j+2}}\, \sin\left(\pi \frac{i+1}{2^j+2}\right)\,, \tag{23a}$$

$$C_{ji}^{\psi} = \sqrt{\frac{2}{2^j+1}}\, \sin\left(\pi \frac{i+1}{2^j+1}\right) \tag{23b}$$

have been introduced here for the purpose of normalization in order to fulfill equations (16) without further constants. The derivation makes use of a trigonometric identity given in [15, p. 14].

From the presentation it is obvious that the linear approximation properties of the wavelet functions ψ_{ji} are those of the spaces V_j spanned by the Legendre polynomials up to degree 2^j. In particular, these polynomials constitute an unconditional basis of $L_2([-1,1])$ so that any square-integrable function $u(x)$, $x \in [-1,1]$, can be decomposed as

$$u(x) = c_{00}\, \varphi_{00}(x) + c_{01}\, \varphi_{01}(x) + \sum_{j=0}^{\infty} \sum_{i=0}^{2^j-1} d_{ji} \psi_{ji}(x), \tag{24}$$

where, by orthonormality, the coefficients are obtained from

$$d_{ji} = \int_{-1}^{1} u(x) \psi_{ji}(x)\, \mathrm{d}x\,, \tag{25a}$$

$$c_{ji} = \int_{-1}^{1} u(x) \varphi_{ji}(x)\, \mathrm{d}x. \tag{25b}$$

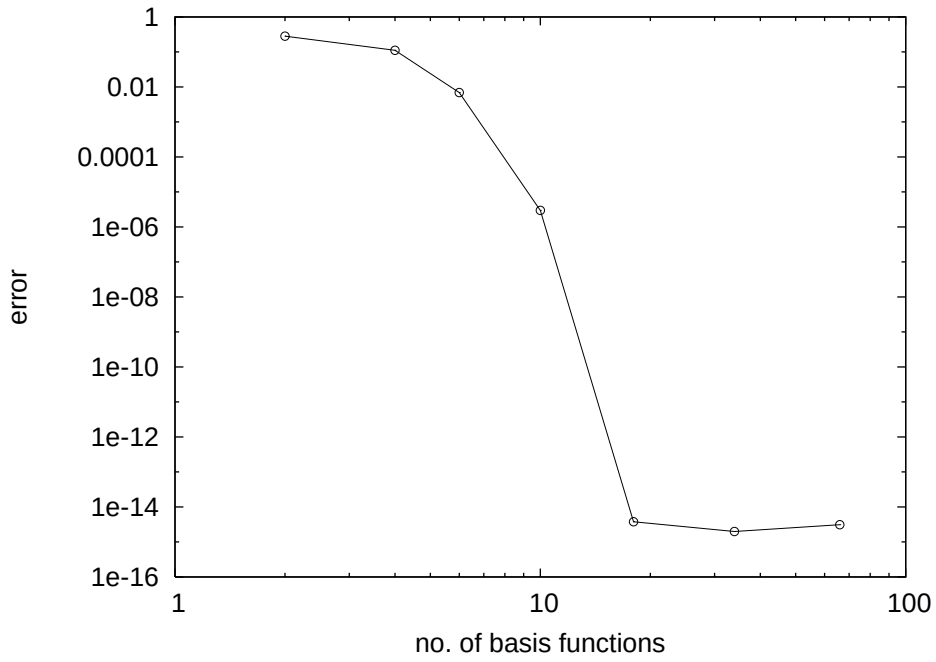


FIG. 2. *Convergence of the approximation when analyzing the function* $u(x) = \exp(-4x^2)$ *for* $x \in [-1; 1]$. *The curve shows the maximum relative error as a function of the number of basis functions used in the reconstruction. The r.m.s. error (not shown) has a similar behavior.*

Again due to orthonormality the decomposition (24) yields a corresponding decomposition of the "energy" of the signal in terms of the coefficients

$$\int_{-1}^{1} u(x)^2 \mathrm{d}x = c_{00}^2 + c_{01}^2 + \sum_{j,i} d_{ji}^2. \tag{26}$$

4. Properties and definition of secondary quantities.

4.1. Implementation and convergence of the approximation. We illustrate the global convergence of the approximation of a function by its wavelet expansion through numerical tests with analytical signals. For this purpose, a partial reconstruction according to (24) with $j \leq J < \infty$ is performed. As a representative example, Figure 2 shows the variation of the maximum error for the signal $u(x) = \exp(-4x^2)$ when the truncation index J is increased. Spectral convergence is observed as expected. The equivalent to Figure 1 is not shown here since at these resolutions no difference between the exact and the approximated data can be discerned. We also note that each function ψ_{ji} has at least 2^j vanishing moments, a fact resulting from the bounds of the sum in (19) and the orthogonality of the Legendre polynomials.

At present, the scalar products (25) are evaluated by a Gauss–Lobatto quadrature, i.e., first performing a Legendre transform of the data—sampled on a Gauss–Lobatto grid—and then computing the linear combination of Legendre coefficients, which leads to the respective wavelet coefficients. If data are given in terms of coefficients of orthogonal polynomials of a different type, like Chebyshev polynomials of the first kind

as, e.g., used in a spectral simulation, explicit conversion formulas [2] might be used. Another means for such a conversion is spectral interpolation onto the Legendre grid as applied in section 4.5 below. The construction of a fast recursive implementation of the present algorithm using classical relations for orthogonal polynomials is left as a future extension.

4.2. Localization properties. Figure 3 shows sample wavelet functions of scale $j = 5$. It can be observed that they are almost translationally invariant near the center of the interval, while they visibly increase their amplitude and frequency near the boundary. This effect of varying shape is more vividly illustrated in Figure 4, where the envelope of the square of several wavelet functions is shown. Particularly, the existence of a second local maximum of the amplitude at the nearest boundary can be observed. From the semilogarithmic plots in Figure 5 the spatial decay of the functions around their center location can be judged. The envelopes approximately decay like $\mathcal{O}(x^{-2})$ (cf. Figure 6) which means that the wavelets themselves decay at a rate of $1/x$. This can be understood by referring to the Littlewood–Paley basis recalled above in section 2.2. In fact, for increasing degree, the zeros of the orthogonal polynomials become more and more uniformly spaced in the center of the interval so that in this region the analysis locally resembles a Fourier analysis. For functions φ_{ji} and ψ_{ji} defined by (15) this amounts to approaching Shannon wavelets due to the employed summation bounds. We can therefore conjecture that with nonoverlapping summation bounds in (15) it will not be possible to improve the decay rate of these functions. Overlapping bounds, however, would add an additional level of complexity to the construction as this destroys the automatic interscale orthogonality which is readily obtained with (15) due to the orthogonality of the underlying polynomials.

The decay, however, is only local. Close to the boundaries the wavelet functions have a tendency to increase and to exhibit the "tails" mentioned above. Table 1 gives the contribution of these tails to the energy of the wavelet, i.e., the integral $\int_{x_{tail}}^{+1} \psi_{ji}^2 \mathrm{d}x$ with x_{tail} being the location where the slope of the envelope reverses. This quantity is below one percent for centrally located wavelets. For comparison, Figure 7 shows the corresponding decay of the wavelets of Fischer and Prestin [11], which are based upon Chebyshev polynomials of the second kind (U_k instead of L_k in (19)). In the latter case the tails are similar and even more pronounced. This is also reflected in the values of Table 1.

We recall that both families of wavelets are related through the basic equation (15) inasmuch as they have common coefficients a_{ijk}, b_{ijk} and differ only in the definition of the associated polynomial function $P_k(x)$. These coefficients a_{ijk} are plotted in Figure 8, where the same indices i, j as in the previous graphs have been chosen. At the same time, the coefficients represent the Legendre spectrum of the present wavelets. It is evident from the graphs that the exact spectral distribution of the basis functions varies with the position index. Furthermore, the low-pass nature of the scaling functions and the band-pass property of the wavelets is obvious.

One question which arises naturally with respect to the usefulness of the current basis is its ability to pick up existing features of a given signal without creating artifacts due to the particular shape of the wavelet functions near the boundary. We will address this point in section 4.4.

4.3. Definition of a scale number. From the definition of the wavelets and from the plots in Figure 3 it is obvious that the period of oscillation of wavelets with

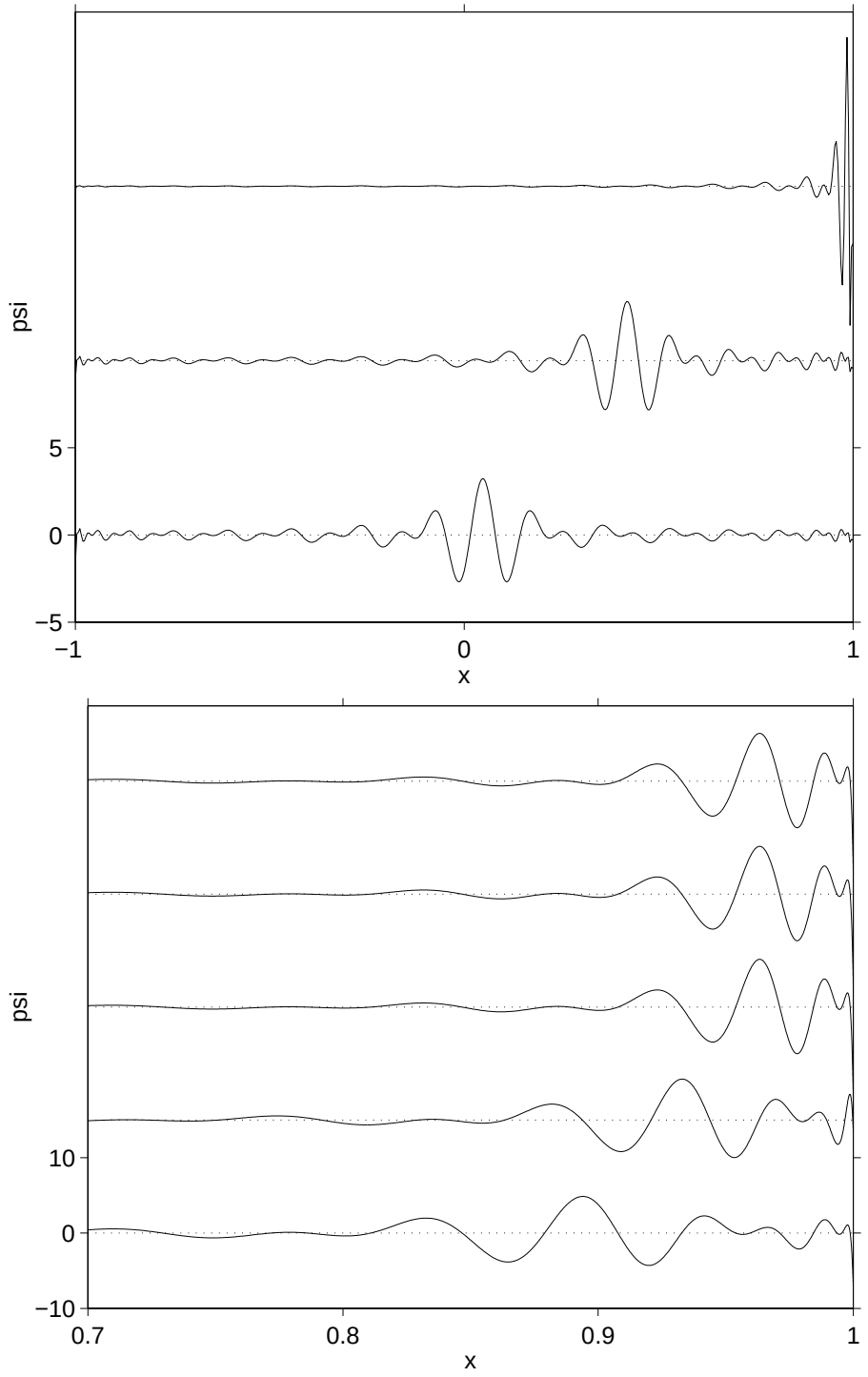


FIG. 3. *Wavelet functions of scale index $j = 5$ computed on a grid with $N = 4096$ points. Position indices are $i = 16, 19, 29$ (top) and $i = 26, 27, 28, 29, 30$ (bottom). Observe that the abscissa of the lower plot is zoomed.*

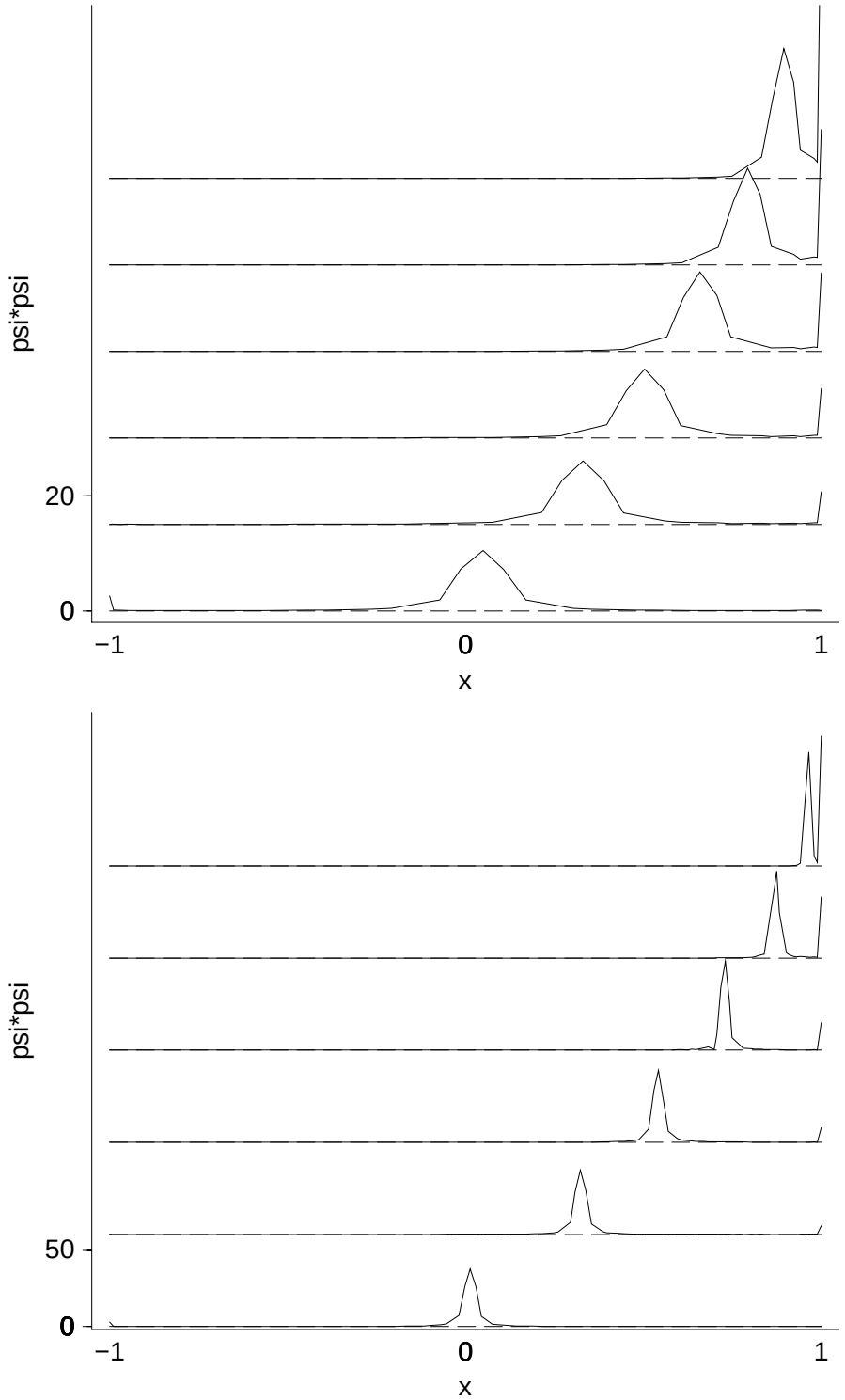


FIG. 4. *Envelope of the square of wavelet functions with scale index $j = 5$ (top) and $j = 7$ (bottom) for different center locations, i.e., computed on a grid with $N = 512$ points.*

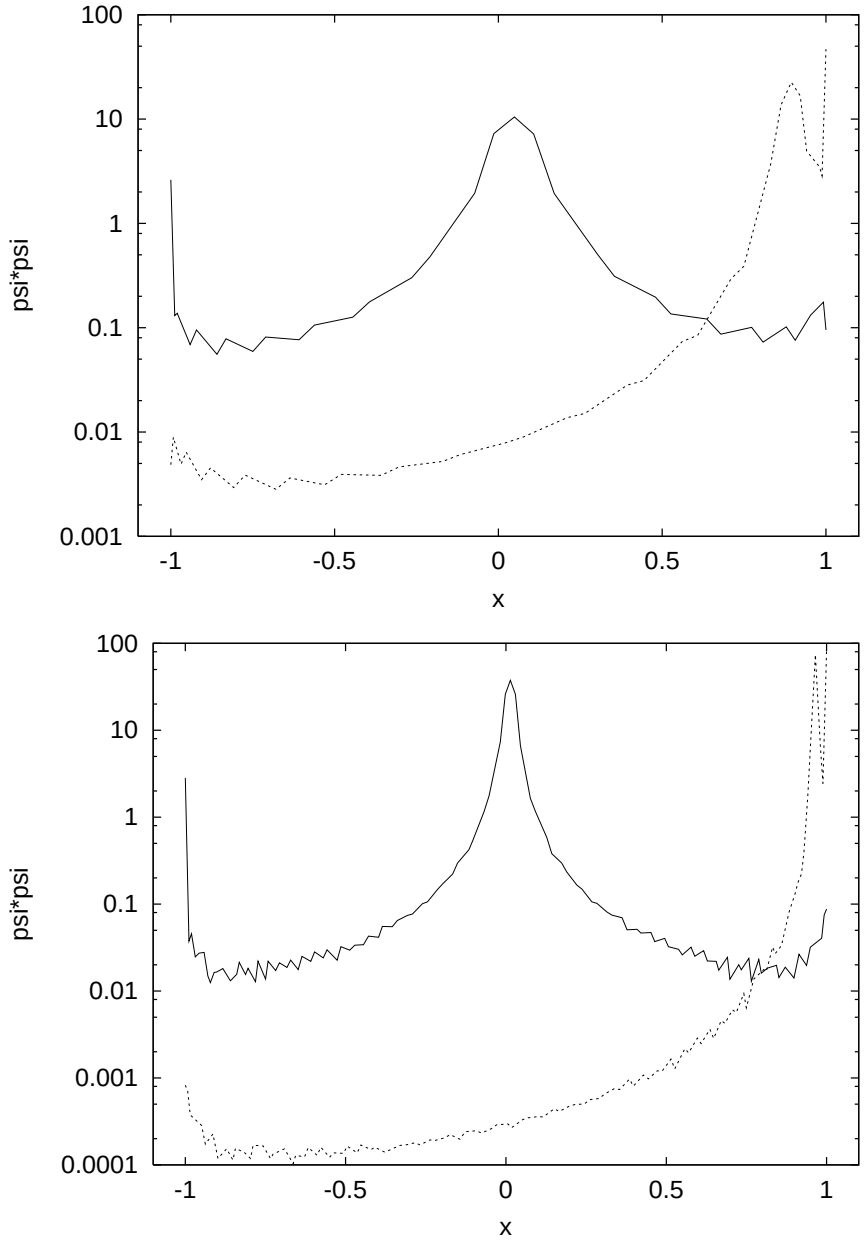


FIG. 5. *Semilogarithmic plot of the envelope of the square of wavelet functions with scale index $j = 5$ (top) and $j = 7$ (bottom) in the center of the interval and close to the boundary, i.e., $i = 15 = 6, 27$ (top) and $i = 64, 117$ (bottom) computed on a grid with $N = 512$ points.*

the same scale index j varies over the interval. Hence, it is important to distinguish between the scale index j and the "physical scale." We therefore attribute a scale number s_{ji} to each wavelet which, at constant j, changes with the position index i. Defining a "scale" and drawing a scalogram hence becomes a nontrivial issue. Here we use the centers of the wavelet functions for this purpose as described in the following.

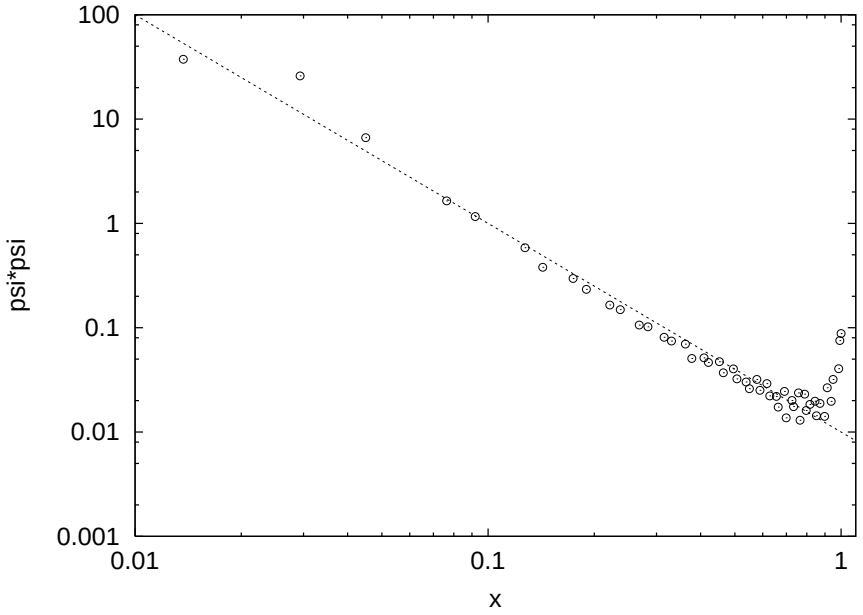


FIG. 6. *Double-logarithmic plot of the decay of the square of the wavelet functions with scale index $j = 7$ and with center location in the middle of the interval. The straight line has a slope of -2.*

TABLE 1

Energy contained in the tails of the wavelet functions at two different levels j and various positions i (cf. Figures 4 and 5). The "tail-location" x_{tail} corresponds to the local minimum of the envelope and has been determined visually using a grid with $N = 1024$ points. The integral has been evaluated by a low-order quadrature. For comparison, the last two columns show the corresponding quantities computed for the wavelets of Fischer and Prestin [11] based upon Chebyshev polynomials of the second kind.

		Legendre wavelets		Chebyshev wavelets	
j	i	x_{tail}	$\int_{x_{tail}}^{+1} \psi_{ji}^2 \mathrm{d}x$	x_{tail}	$\int_{x_{tail}}^{+1} \psi_{ji}^2 \mathrm{d}x$
5	15	0.879	6.17e-3	0.933	4.40e-2
	12	0.850	1.46e-2	0.933	1.17e-1
	10	0.922	1.67e-2	0.933	1.59e-1
	8	0.922	2.80e-2	0.962	2.03e-1
	6	0.922	5.54e-2	0.996	2.15e-1
	4	0.957	7.87e-2	0.996	3.31e-1
7	63	0.844	1.79e-3	0.953	1.12e-2
	50	0.932	2.15e-3	0.992	2.29e-2
	40	0.932	3.79e-3	0.992	3.47e-2
	30	0.932	7.64e-3	0.992	5.29e-2
	20	0.975	1.04e-2	0.992	8.66e-1

Recall that the zeros of the Legendre polynomials are not available in closed form. As a consequence, the locations of the "centers" of the wavelets and scaling functions defined here are not available in closed form but need to be determined numerically. To be specific, by "centers" z_{ji}^{φ}, z_{ji}^{ψ} we mean the position of the largest positive local maximum values, excluding the boundaries of the interval, which can be obtained, e.g.,

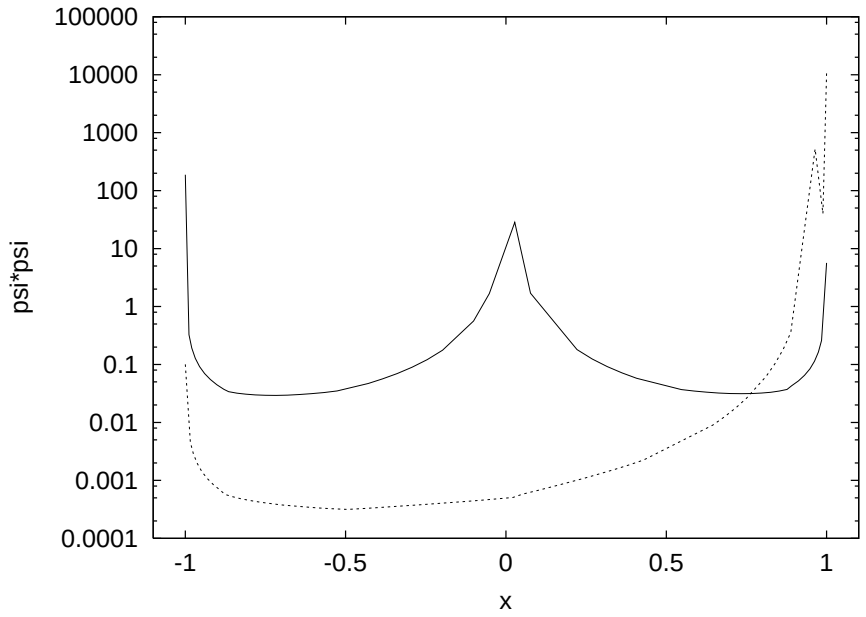


FIG. 7. *As the second plot of Figure* 5*, but for the wavelet functions of* [11]*, being based upon Chebyshev polynomials of the second kind instead of the present Legendre polynomials.*

by a fixed point iteration of the first derivative of (19). In practice, this procedure is, however, very cumbersome and an analytic expression, even approximate, would be preferable, particularly in view of the way of presenting information with respect to scale as discussed below. Therefore, we propose—solely for the definition of the "scale" of a wavelet function—to work with the roots of the Chebyshev polynomials of the second kind. Instead of the centers z_{ji}^{φ}, z_{ji}^{ψ} defined above we therefore use the approximations

$$\hat{z}_{ji}^{\varphi} = y_i^{(2^j+1)}\,, \tag{27a}$$

$$\hat{z}_{ji}^{\psi} = y_i^{(2^j)}\,, \tag{27b}$$

with $y_i^{(n)}$ given in (22). Figure 9 shows the relative difference between the two definitions z_{ji}^{ψ} and $\hat{z}_{ji}^{\psi}$. It exhibits a minimum in the center of the interval and near the boundaries.

For the purpose of data analysis, we associate a "physical scale" with each wavelet function. In the classical MRA, where wavelets are translationally invariant, the scale is simply $s_j = 2^{-j} L_x$, with L_x being the size of the domain. Here we define the scale number s_{ji} for the nonperiodic wavelets as follows:

$$s_{ji} = \frac{L_x}{2} \begin{cases} \dfrac{z_{ji+1}^{\psi} + z_{ji}^{\psi}}{2} + 1 & \text{if} \qquad i = 0\,, \\ 1 - \dfrac{z_{ji}^{\psi} + z_{ji-1}^{\psi}}{2} & \text{if} \qquad i = 2^j - 1\,, \\ \dfrac{z_{ji+1}^{\psi} - z_{ji-1}^{\psi}}{2} & \text{otherwise}\,. \end{cases} \tag{28}$$

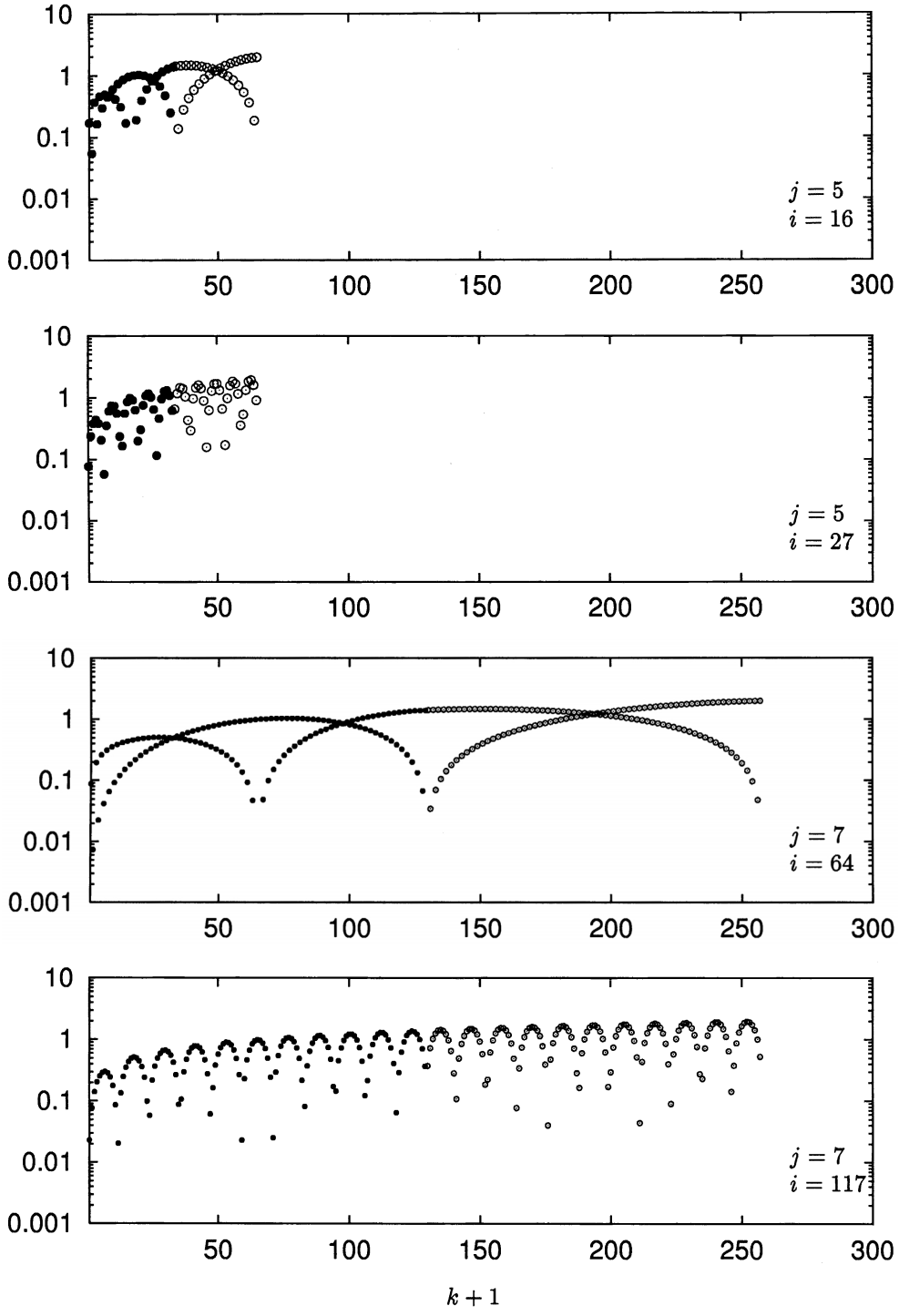


FIG. 8. *Semilogarithmic plot of the absolute value of the Legendre spectral coefficients (as a function of the wavenumber k) of wavelet functions (open symbols) and scaling functions (full symbols). Plots are displayed for scale index* $j = 5$ *and* $i = 4$ *(top),* $i = 15$ *(second), as well as for* $j = 7$ *with* $i = 64$ *(third) and* $i = 117$ *(bottom), respectively.*

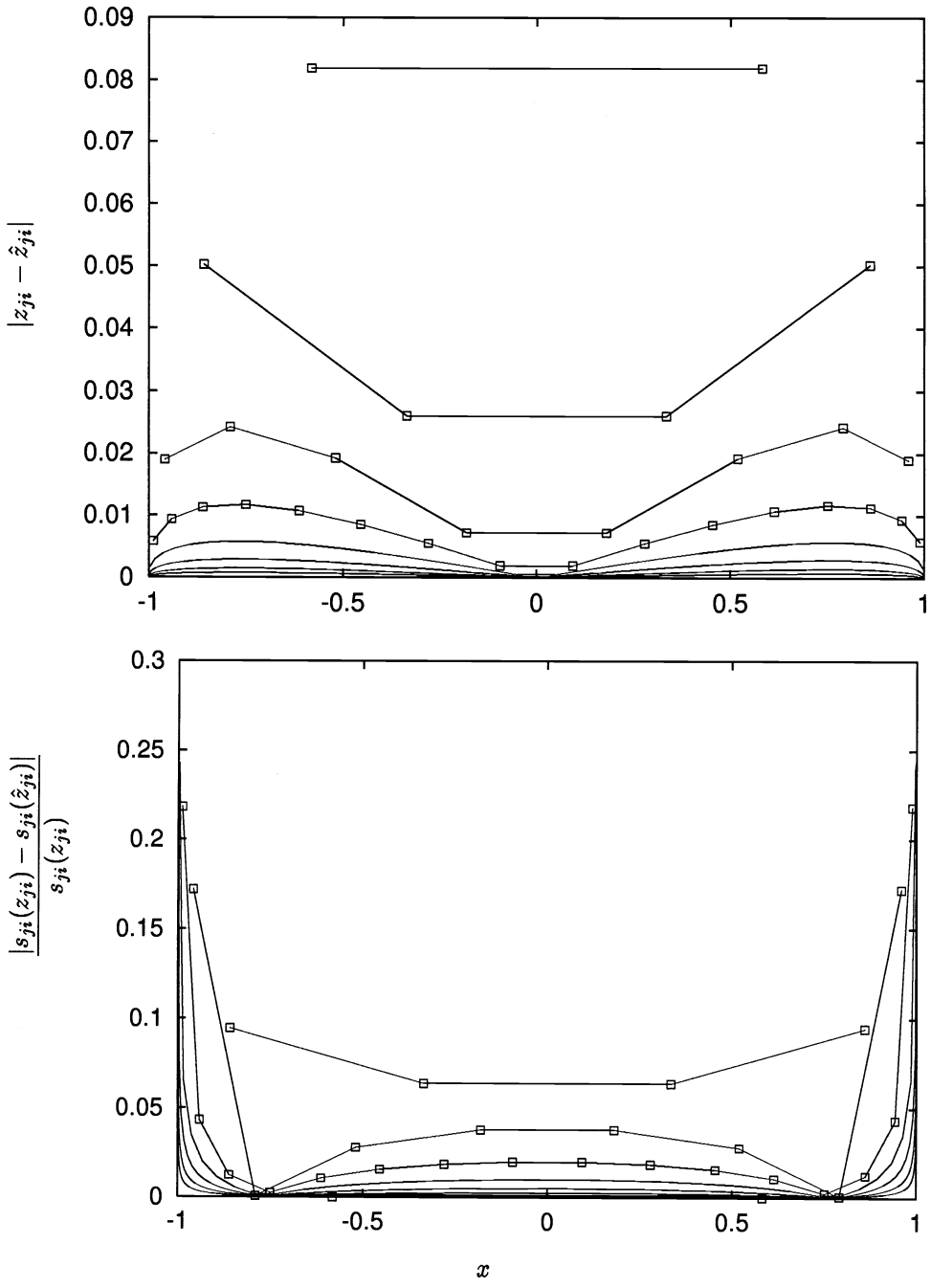


FIG. 9. *The difference in the definition of the center locations (top) and the scale number (bottom) corresponding to each wavelet ψ_{ji} of the present Legendre basis: $s_{ij}(\hat{z}_{ji})$ is the scale defined by the spatial variation of the Chebyshev roots $\hat{z}_{ji}$ given in* (28)*; $s_{ji}(z_{ji})$ is the value obtained by numerically computing the distance between neighboring centers z_{ji} of wavelets.*

This results from partitioning the interval into subintervals bounded by the midpoints between neighboring center locations. Consequently, it implies that $\sum_i s_{ji} = L_x$ and in particular that $s_{00} = L_x$. As discussed above, the quantity s_{ji} cannot be determined analytically if the exact centers z^ψ_{ji} are used. As a remedy we propose using, instead of z^ψ_{ji}, the roots of the Chebyshev polynomials $\hat{z}^\psi_{ji}$ in (28) which upon substitution of (22) yield

(29)
$$s_{ji}(\hat{z}^\psi_{ji}) = \frac{L_x}{2} \begin{cases} \cos\left(\dfrac{(2^j - 1/2)\pi}{2^j + 1}\right) \cos\left(\dfrac{\pi/2}{2^j + 1}\right) + 1 & \text{if} \quad i = 0\,, \\ 1 - \cos\left(\dfrac{3/2\pi}{2^j + 1}\right) \cos\left(\dfrac{\pi/2}{2^j + 1}\right) & \text{if} \quad i = 2^j - 1\,, \\ \sin\left(\dfrac{(i+1)\pi}{2^j + 1}\right) \sin\left(\dfrac{\pi}{2^j + 1}\right) & \text{otherwise}\,. \end{cases}$$

Figure 9 shows the resulting relative difference in scale between the definition of s_{ji} with the exact and with the approximated centers. These differences are only appreciable near the boundary, where they reach a very localized maximum of 25 percent. In the present situation the definition of a "scale" associated with a wavelet necessarily has to be somewhat arbitrary, in particular close to the boundaries. We therefore feel that the definition (29) suits the purpose of data analysis and visualization. This is backed by the examples below.

Based on the two definitions of the scale parameter s_{ji} two ways of presenting the coefficients of the present transform can be constructed. They will be detailed and illustrated by means of analytical signals in the following paragraph. In what follows, we will then adopt definition (29).

4.4. Transform of analytical signals and coefficient scheme. Before application to real-life signals it is instructive to study the transform itself by means of analytical signals. We consider as a first case the transform of a periodic signal, $u(x) = \sin(2\pi\, x\, a)$, with various frequencies a. This is a particular case of a signal which can be analyzed with the present new algorithm as well as with a standard method for periodic signals.

Let us now discuss the representation of the wavelet coefficients. In Figures 10 and 11 we compare two different ways of drawing scalograms for the same coefficient values. In the first method, used in Figure 10, the "exact"—i.e., iteratively determined—center locations z^ψ_{ji} and resulting scales $s_{ji}(z_{ji})$ are used to define rectangular cells Ω_{ji} in the following way:

(a1) The center of Ω_{ji} is defined by the coordinate pair $(z^\psi_{ji}, -\log_2(s_{ji}))$.
(a2) The width of Ω_{ji} is equal to the scale s_{ji}.
(a3) The height of Ω_{ji} is set to an arbitrary constant value.

The plot is then obtained by drawing each cell colored according to the absolute value of the corresponding wavelet coefficient. Note that due to some overlapping of the rectangles they seem to have the shape of more irregular polygons. In Figure 10 as well as the subsequent scalograms below, 8 wavelet levels $j = 0, \ldots, 7$ have been computed and plotted. The coefficients of the finest scales are often not visible since their amplitude is below the threshold for the grayscale.

The second type of visual presentation is displayed in Figure 11. It is based on the approximate center points and scales via the roots of the Chebyshev polynomials of the second kind. Introducing the parameter $\theta_{ji} = \pi(i+1)/(2^j + 1)$, we can rewrite

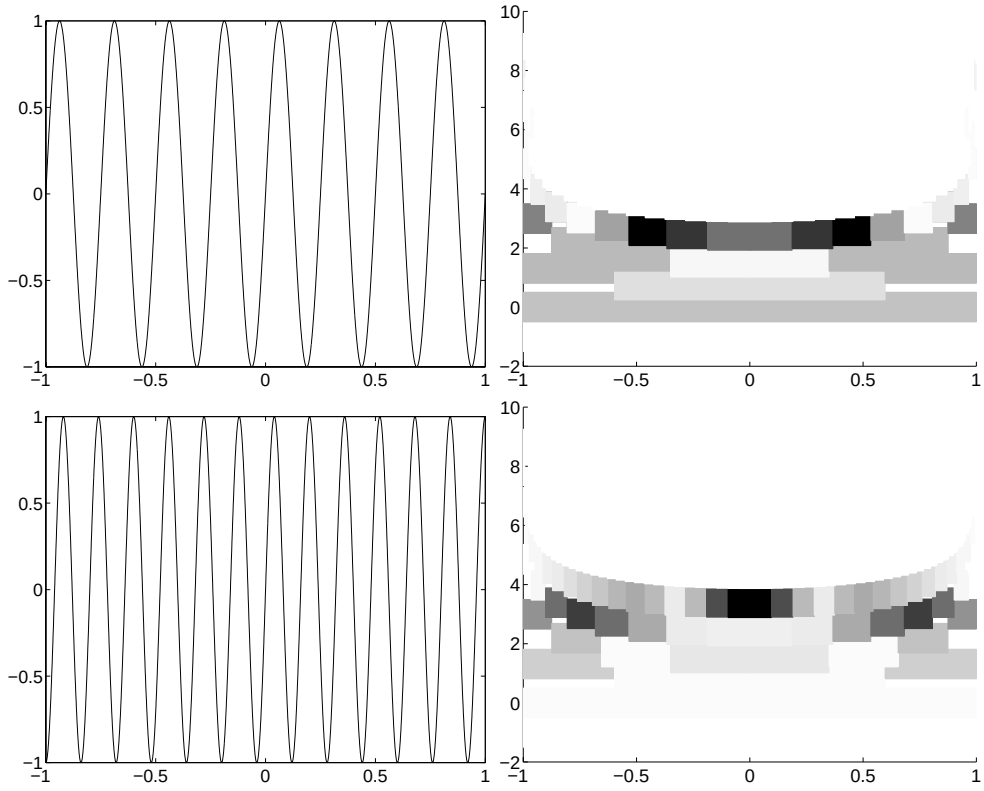


FIG. 10. *Wavelet transform of a periodic signal:* $u(x) = \sin(2\pi\, x\, a)$ *with* $a = 4$ *(top) and* $a = 6.25$ *(bottom). Note that the second signal is not periodic on the present interval* $[-1, 1]$. *In the scalogram on the right the abscissa relates to the position in the domain* $x \in [-1, 1]$ *while the ordinate gives the inverse of the scale number in logarithmic scale, i.e.,* $-\log_2(s_{ji})$. *Darker shading indicates higher (absolute) coefficient values on a linear scale. The total number of modes is* $N + 1 = 257$. *The scale is defined by means of the "exact," numerically determined centers* z^{ψ}_{ji} *of the wavelets. Therefore, the visualization features "cells" which have a seemingly polygonal shape through overlapping and are filled with the grayscale value according to the coefficients.*

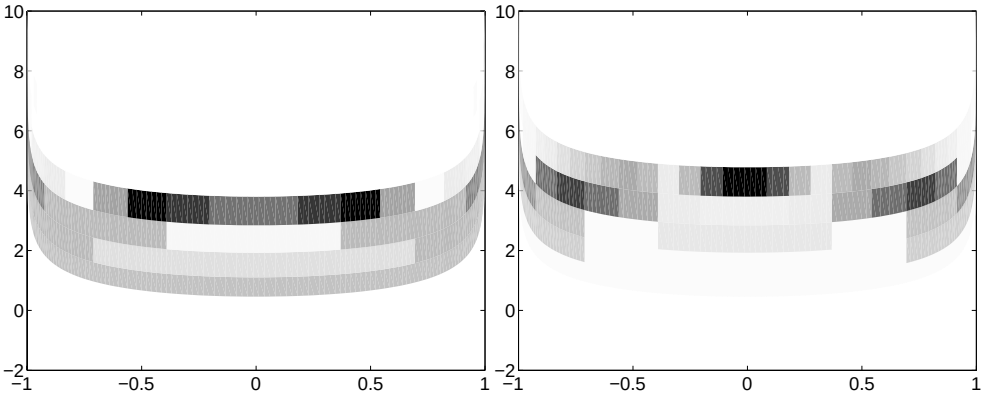


FIG. 11. *Coefficient scalograms as in Figure* 10, *but with the scale defined by means of the approximate centers* $\hat{z}^{\psi}_{ji}$ *of the wavelets and with the related transform used to determine the shape of the cells representing the coefficients.*

the definitions (27), (29) as

$$\hat{z}_{ji}^{\psi} = \cos(\theta_{ji})\,, \tag{30a}$$

$$s_{ji}(\hat{z}_{ji}^{\psi}) = \frac{L_x}{2}\,\sin(\theta_{ji})\sin\left(\frac{\pi}{2^j+1}\right), \qquad 0 < i < 2^j - 1\,. \tag{30b}$$

These relations represent a discrete mapping from dyadic θ_{ji} to $\hat{z}_{ji}$ and s_{ji}, respectively, which can be extended to the continuous case by replacing θ_{ji} with $\theta \in]0, \pi[$. In this fashion, a scalogram is constructed with cell boundaries progressively deformed such as to indicate a spatial change of scale at fixed scale index j. In practice, we proceed as follows:

(b1) We define a classical scalogram with rectangular cells centered at $(\theta_{ji}, -\log_2(s_{ji}(\hat{z}_{j,i=2^j/2}))$, i.e., using the scale of the centrally located wavelets given in (29), and we separate the cells at the midpoints between neighbors.

(b2) We transform the coordinate locations of the cell boundaries by the maps $x = \cos(\theta)$, $y = -\log_2(\sin(\theta))$.

The result is a pattern with strips of coefficients of common scale index j which are bent upwards near the boundaries. By this method, the coefficient values corresponding to small-scale indices j appear at different physical scales along their horizontal extent, which in a way is a visual representation of the fact that a single wavelet undergoes a similar variation in frequency along the interval. As Figures 11 and 10 demonstrate, both methods of visualization are of fairly similar quality with respect to the readability of frequency content and position of the signal. For its smoothness and because it allows for a natural partitioning of the ordinate without gaps we will henceforth retain the second type of scalogram.

It can be seen in Figure 11 that the present basis correctly shows a response at approximately constant scale across the interval. Recall that the use of a real-valued wavelet always tends to yield small-scale oscillations of the coefficients due to cancellations between the signal and the wavelet itself [8]. Therefore, a pure sine wave does not show up as a solid line in the scalogram but rather as a horizontal band with alternating values. It is particularly noteworthy that in the coefficient plots no artifacts are generated at the ends of the interval. Although the signal is periodic in the upper plot, the analysis is entirely independent of this fact. No relation between both boundaries is imposed or assumed. In the lower graph of Figures 10 and the right-hand graph of Figure 11 the period length of the sine was chosen to be incommensurate with the length of the interval—a configuration which is not compatible with a periodic analysis. In the present case, however, the plots of the coefficients remain similar, even if the values $u(x = 1)$ and $u(x = -1)$ are different and not zero.

Next, let us turn to the transform of a Gaussian bump, $u(x) = \exp(-(x - x_c)/(2\sigma^2))$, with different standard deviations σ and center locations x_c. Here, the question is whether the position of the peak can be correctly determined from the scalogram and if information on the characteristic scale can be extracted in this fashion. Figures 12 to 13 show that this indeed is the case. The maximum amplitudes of the coefficients are not exactly pyramid-shaped when x_c is off-center—an effect common to all wavelet transforms with downsampling between different levels—but they do point to the x_c-locations. Furthermore, the scale of the cusp (i.e., the smallest scale where a large amplitude is recorded) corresponds to the scale of the signal. The observed difference between the cusps of the scalograms in Figures 12 and 13 is of

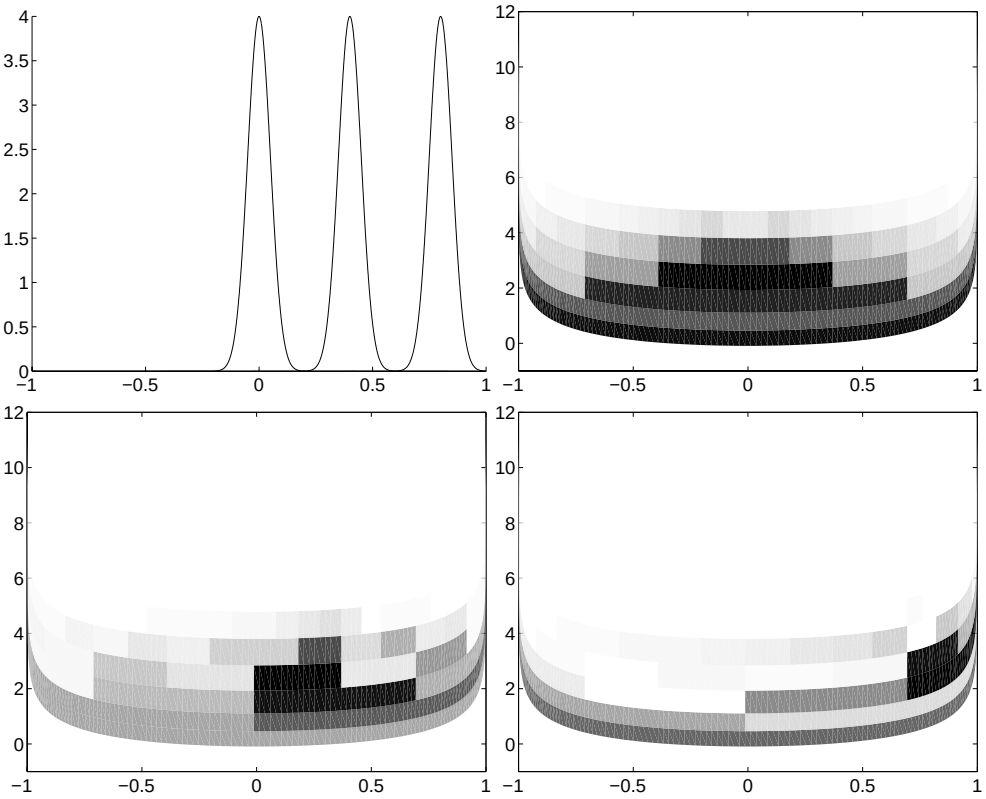


FIG. 12. *Wavelet transform of a single Gaussian bump:* $u(x) = \exp(-((x - x_c)/(2\sigma))^2)$ *with* $\sigma = 0.05$ *and the locations* $x_c = 0.0, 0.4, 0.8$ *from top to bottom and from left to right. Darker shading indicates higher (absolute) coefficient values on a linear scale. The abscissa relates to the position in the domain* $x \in [-1, 1]$ *while the ordinate gives the inverse of the scale number in logarithmic scale, i.e.,* $-\log_2(s_{ji})$*. The total number of modes is* $N + 1 = 257$*.*

roughly two octaves, i.e., a factor of 4, while the standard deviation of the data varies fivefold.

4.5. Definition of wavelet spectra. Let us now introduce a pseudowavenumber as the inverse of the scale parameter, $k_{ji} = 1/s_{ji}$. As a *global* power-spectral density per unit wavenumber we then define

$$E(k_m) = \frac{1}{\Delta k_m} \sum_{j,i \,/\, k_L^m \le k_{ji} \le k_R^m} d_{ji}^2. \tag{31}$$

The index pairs (j, i) in the sum are selected such that the pseudowavenumber of the corresponding wavelet falls into the interval $[k_L^m, k_R^m]$, where the whole wavenumber range considered is partitioned into M such bins, $1 \le m \le M$. In the present work, the bins are spaced logarithmically. The function (31) is normalized by the wavenumber increment Δk_m so that the following identity holds for the total energy:

$$E_{tot} = \int_{-1}^{+1} u(x)^2 \, \mathrm{d}x = \sum_{i=0}^{1} c_{0i}^2 + \sum_{j=0}^{J} \sum_{i=0}^{2^j-1} d_{ji}^2 = \sum_{i=0}^{1} c_{0i}^2 + \sum_{m=1}^{M} E(k_m) \, \Delta k_m. \tag{32}$$

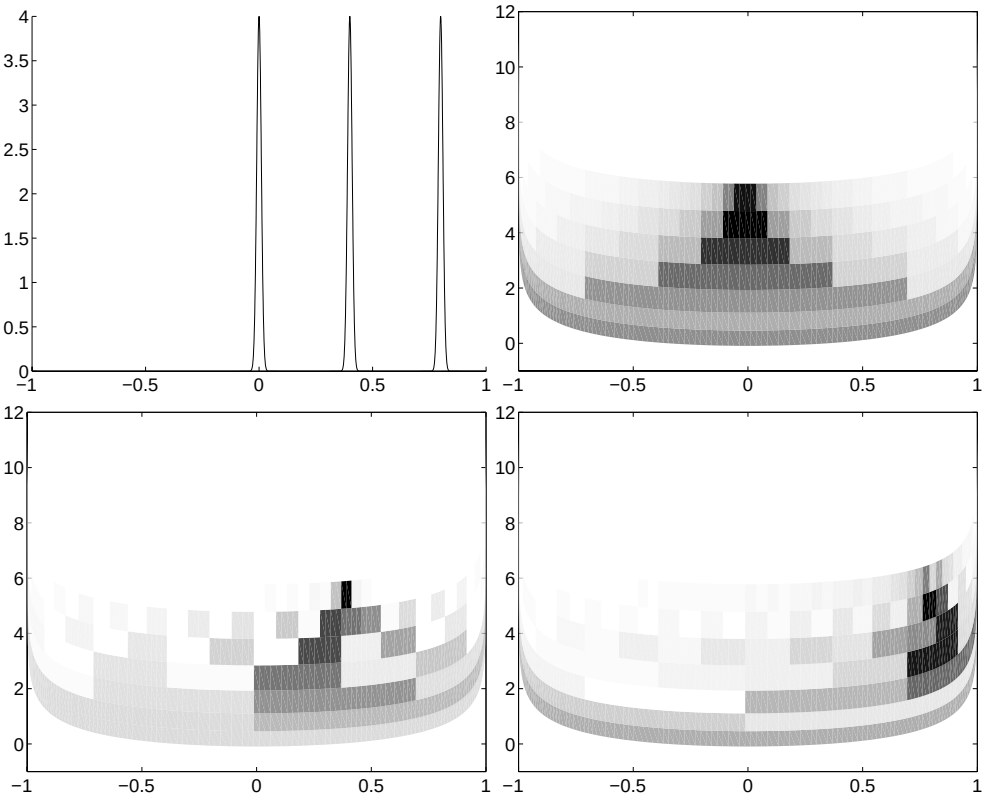


FIG. 13. *As Figure* 12*, but the signal is a narrower bump with* $\sigma = 0.01$.

Next, we define a *local* power-spectral density per unit wavenumber

$$E(k_j, x) = 2^j \frac{d_{ji^*}^2}{\Delta k_j}, \tag{33}$$

where the position index $i^* = i^*(x)$ corresponds to the wavelet at scale index j whose center $\hat{z}_{ji}^{\psi}$ lies closest to the location x. In other words, the function $E(k_j, x)$ represents a cut through the scalogram at the abscissa x. Therefore, there are exactly $J + 1$ such spectral values at each location and the largest scales are obviously redundantly reproduced in spectra evaluated at small distances from each other. In particular, the $j = 0$-coefficient will enter all local spectra. In contrast to the fixed wavenumber increment Δk_m in (31), the increment Δk_j is based on $k_{j,i^*} = 1/s_{j,i^*}$, i.e., on the scale parameters of the coefficients actually selected. The factor 2^j in (33) is introduced for compatibility with the global spectrum (cf. [7]). Applications of spectral analysis using the present wavelet basis are given in section 6.

5. Multidimensional basis. The construction of a wavelet basis in more than one dimension typically proceeds along either one of the following lines [5, p. 313]:

A Performing a tensor product of one-dimensional bases in each coordinate direction. With this procedure two separate scale indices j_x, j_y are introduced and the mechanism of rescaling is not directionally invariant. Therefore, this method is sometimes called *rectangular transform* [16].

B Performing a tensor product of one-dimensional MRAs with a "global" scale index j (*square transform*) and different wavelets for picking up the various directional features.

C Design of genuinely multidimensional wavelet/scaling functions with the desired orthonormality and angular selectivity properties.

Due to its additional complexity, a construction of type **C** is beyond the scope of the present paper. On the other hand, both constructions **A** and **B** are straightforward once a suitable one-dimensional basis has been found. Regarding their use as a tool for data analysis, these two options differ in various respects. We have found method **A** more useful in the present context and will retain it for subsequent applications. The reasons for this choice will become obvious during the following presentation of both methods. The discussion is performed for two space dimensions; for higher dimensions the situation is analogous.

5.1. Method A: Tensor product of one-dimensional wavelet functions. The two-dimensional basis according to procedure **A** consists of the following functions:

$$\text{(34a)}\quad \varphi^{0,j_y}_{i_x,i_y}(x,y) = \varphi_{0,i_x}(x)\,\psi_{j_y,i_y}(y), \qquad i_x = 0\dots 1, \qquad i_y = 0\dots 2^{j_y}-1,$$

$$\text{(34b)}\quad \varphi^{j_x,0}_{i_x,i_y}(x,y) = \psi_{j_x,i_x}(x)\,\varphi_{0,i_y}(y), \qquad i_x = 0\dots 2^{j_x}-1, \quad i_y = 0\dots 1,$$

$$\text{(34c)}\quad \psi^{j_x,j_y}_{i_x,i_y}(x,y) = \psi_{j_x,i_x}(x)\,\psi_{j_y,i_y}(y), \qquad i_x = 0\dots 2^{j_x}-1, \quad i_y = 0\dots 2^{j_y}-1,$$

where $j_x, j_y = 0, 1, \dots$ and the one-dimensional functions $\psi_{j,i}$ and $\phi_{j,i}$ are defined in (19).

Below we wish to analyze two-dimensional data from a turbulent plane channel flow computation which possesses one periodic (x) and one bounded (y) coordinate direction, i.e., we consider the space $L^2(\mathbb{R}/\mathbb{Z}\times[-1,1])$. For this task we propose a hybrid construction composed of a periodic wavelet basis and the Legendre wavelet basis of section 3.2. The two-dimensional scaling functions and the wavelet functions are then defined as follows:

$$\text{(35a)}\quad \varphi^{0,j_y}_{i_x,i_y}(x,y) = \widetilde{\varphi}_{0,i_x}(x)\,\psi_{j_y,i_y}(y), \qquad i_x = 0, \qquad i_y = 0\dots 2^{j_y}-1,$$

$$\text{(35b)}\quad \varphi^{j_x,0}_{i_x,i_y}(x,y) = \widetilde{\psi}_{j_x,i_x}(x)\,\varphi_{0,i_y}(y), \qquad i_x = 0\dots 2^{j_x}-1, \quad i_y = 0\dots 1,$$

$$\text{(35c)}\quad \psi^{j_x,j_y}_{i_x,i_y}(x,y) = \widetilde{\psi}_{j_x,i_x}(x)\,\psi_{j_y,i_y}(y), \qquad i_x = 0\dots 2^{j_x}-1, \quad i_y = 0\dots 2^{j_y}-1.$$

The functions $\varphi(y)$, $\psi(y)$ are defined in (19). The periodic functions $\widetilde{\varphi}(x)$, $\widetilde{\psi}(x)$ employed here are spline wavelets of order 4 and can be found in detail in references [29, 12]. With the basis (35), the approximation of a two-dimensional function up to a scale J reads as follows:

$$\begin{aligned} u(x,y) = &\sum_{j_y=0}^{J}\sum_{i_y=0}^{2^{j_y}-1} c^{0,j_y}_{0,i_y}\,\varphi^{0,j_y}_{0,i_y}(x,y) + \sum_{j_x=0}^{J}\sum_{i_x=0}^{2^{j_y}-1}\sum_{i_y=0}^{1} c^{j_x,0}_{i_x,i_y}\,\varphi^{j_x,0}_{i_x,i_y}(x,y) \\ &+ \sum_{j_x=0}^{J}\sum_{i_x=0}^{2^{j_x}-1}\sum_{j_y=0}^{J}\sum_{i_y=0}^{2^{j_y}-1} d^{j_x,j_y}_{i_x,i_y}\,\psi^{j_x,j_y}_{i_x,i_y}, \end{aligned} \tag{36}$$

which leads to a total number of $N(N+1)$ coefficients, where $N = 2^{J+1}$. Due to

orthogonality, the coefficients are obtained from the following scalar products:

$$d_{i_x,i_y}^{j_x,j_y} = \int_x \int_y u(x,y)\, \psi_{i_x,i_y}^{j_x,j_y}(x,y)\, \mathrm{d}y\mathrm{d}x\,, \tag{37a}$$

$$c_{i_x,i_y}^{j_x,j_y} = \int_x \int_y u(x,y)\, \varphi_{i_x,i_y}^{j_x,j_y}(x,y)\, \mathrm{d}y\mathrm{d}x\,. \tag{37b}$$

These integrals can be factorized during the computation due to the tensorial nature of the wavelets and the scaling functions. Therefore, we can first apply the standard Mallat algorithm to each "row" of data at constant y and then proceed column-wise by computing the remaining integration in the y-direction by the new scheme of section 4.1.

Figure 14 shows the shape of the wavelets $\psi_{i_x,i_y}^{j_x,j_y}(x,y)$, for the scale indices being combinations of 2 and 5 and at two locations, in the center of the domain and close to the boundary $y = 1$. The localization properties are quite different in the two coordinate directions. Spline wavelets have an exponential decay while the Legendre wavelets decay roughly as x^{-1} and exhibit the characteristic tails near the boundaries as discussed above.

For visual presentation, the coefficients can be arranged in matrix fashion, i.e., collocated blockwise according to the values of the index pair (j_x, j_y) (cf. Figure 15). Each coefficient within a block is represented by a rectangle colored according to the coefficient's absolute value. The size of this rectangle is determined according to the wavelet centers, i.e., uniform in the x-direction and using the approximate center locations of the Legendre wavelets (27) in the y-direction. This leads to flattened cells near the boundaries $y = \pm 1$, reflecting the two length scales by their aspect ratio s_x/s_y. Examples are provided below.

5.2. Method B: Multidimensional MRA. Following Mallat [25] we define the following two-dimensional scaling functions and a set of three different types of wavelet functions:

$$\varphi_{i_x,i_y}^{j}(x,y) = \widetilde{\varphi}_{j,i_x}(x)\, \varphi_{j,i_y}(y), \qquad i_x = 0 \ldots 2^j - 1, \qquad i_y = 0 \ldots 2^j, \tag{38a}$$

$$\psi_{i_x,i_y}^{j,1}(x,y) = \widetilde{\varphi}_{j,i_x}(x)\, \psi_{j,i_y}(y), \qquad i_x = 0 \ldots 2^j - 1, \qquad i_y = 0 \ldots 2^j - 1, \tag{38b}$$

$$\psi_{i_x,i_y}^{j,2}(x,y) = \widetilde{\psi}_{j,i_x}(x)\, \varphi_{j,i_y}(y), \qquad i_x = 0 \ldots 2^j - 1, \qquad i_y = 0 \ldots 2^j, \tag{38c}$$

$$\psi_{i_x,i_y}^{j,3}(x,y) = \widetilde{\psi}_{j,i_x}(x)\, \psi_{j,i_y}(y), \qquad i_x = 0 \ldots 2^j - 1, \qquad i_y = 0 \ldots 2^j - 1, \tag{38d}$$

where $j = 0, 1, \ldots$. Similar to the one-dimensional case this defines a multidimensional MRA of $L^2(\mathbb{R}/\mathbb{Z}\times[-1,1])$. Removing the tilde and adjusting the indices in x-direction provides the analogous basis for the fully nonperiodic case. Observe that the two-dimensional scaling function φ_{i_x,i_y}^{j} represents the smooth content of the signal in both directions at scale index j while the wavelets $\psi_{i_x,i_y}^{j,1}$, $\psi_{i_x,i_y}^{j,2}$, $\psi_{i_x,i_y}^{j,3}$ pick up the detailed information with respect to vertical (y), horizontal (x), and diagonal variations of the signal, respectively. Figure 16 shows the shape of the three types of wavelets $\psi_{i_x,i_y}^{j,q}(x,y)$, $q = 1, 2, 3$, for the scale index $j = 5$ at two locations: in the center of the domain and close to the boundary $y = 1$.

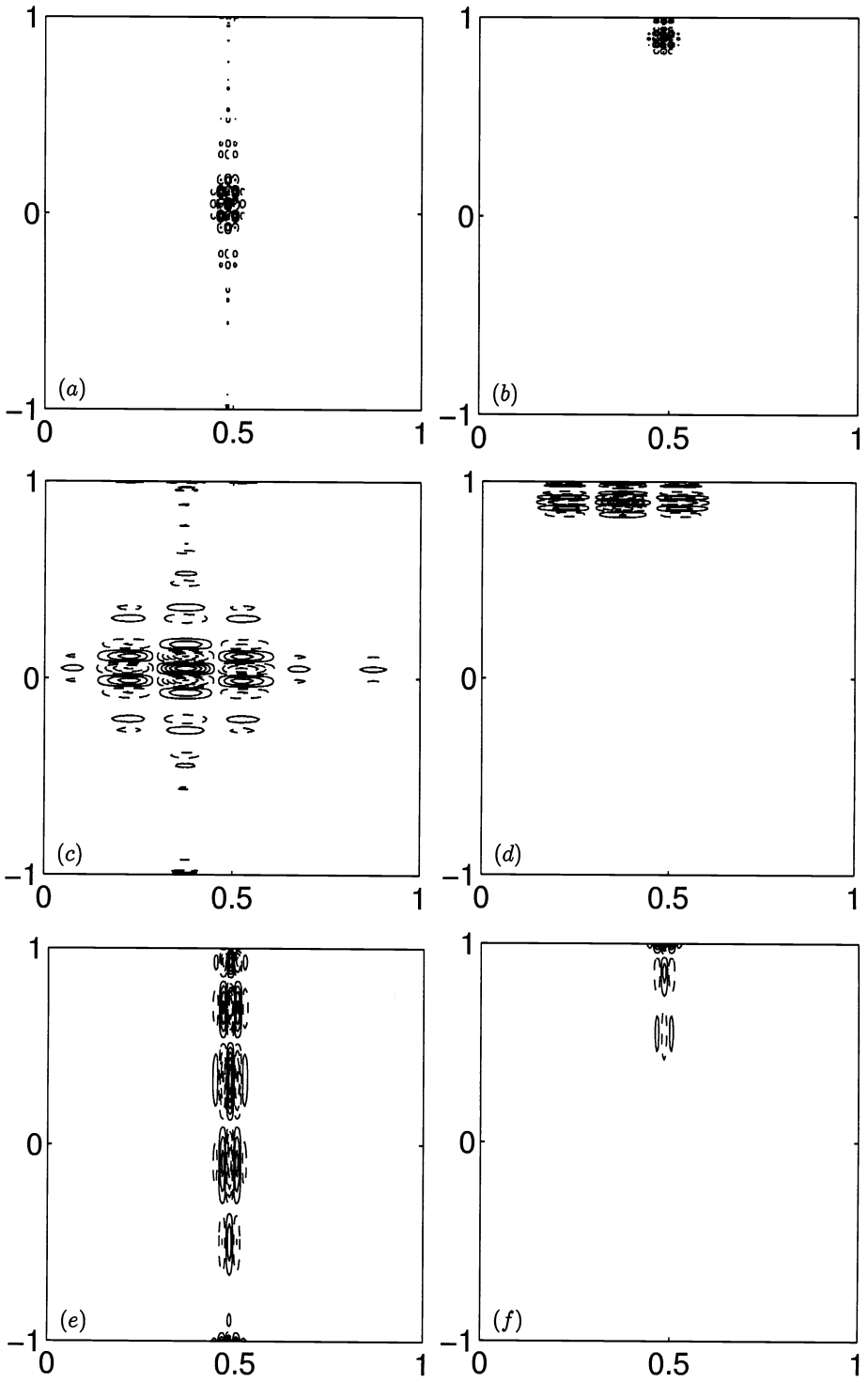


FIG. 14. *Illustration of the basis functions of the tensor-product construction according to method* **A** *sampled on a grid with* 256^2 *points. The values for the quadruplet of indices* j_x, j_y, i_x, i_y *are* (a) $5, 5, 16, 16$*;* (b) $5, 5, 16, 27$*;* (c) $2, 5, 2, 16$*;* (d) $2, 5, 2, 27$*;* (e) $5, 2, 16, 2$*;* (f) $5, 2, 16, 3$.

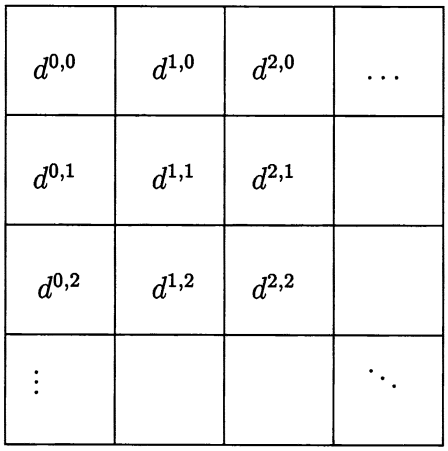


FIG. 15. *The graphical representation of the wavelet coefficients* $d^{j_x,j_y}_{i_x,i_y}$ *of the two-dimensional tensor-product basis (method* **A***) defined by* (35).

With (38), the approximation of a two-dimensional function up to a scale J reads

$$u(x,y) = c^0_{0,0}\,\varphi^0_{0,0}(x,y) + c^0_{0,1}\,\varphi^0_{0,1}(x,y) + \sum_{j=0}^{J}\sum_{i_x=0}^{2^j-1}\left[\sum_{i_y=0}^{2^j} d^{j,2}_{i_x,i_y}\,\psi^{j,2}_{i_x,i_y}(x,y) + \sum_{q=\{1,3\}}\sum_{i_y=0}^{2^j-1} d^{j,q}_{i_x,i_y}\,\psi^{j,q}_{i_x,i_y}(x,y)\right], \tag{39}$$

which leads to a total number of $2+\sum_{j=0}^{J}\{2^j\,(2^j+1)+2\cdot 2^{2j}\} = N(N+1)$ coefficients. Splitting the inner sum results from the different number of scaling functions and wavelets in the Legendre construction, as already reflected by the index bounds in (38). When representing the coefficients obtained from a "classical" two-dimensional MRA graphically, one customarily uses a block diagram where at each level j the rectangular domain is divided into quarters. Three of them are used for representing the coefficients of level j, while the fourth is subdivided again for the following level $j-1$ and so on [25]. We locate the coefficients $d^{j,1}_{i_x,i_y}$ in the lower left quadrant, $d^{j,3}_{i_x,i_y}$ in the lower right, and $d^{j,2}_{i_x,i_y}$ in the upper right (cf. the schematic in Figure 17), an arrangement differing from the one of Mallat [25, 5]. Its advantage is that in the one-dimensional limit $u(x,y) = u_1(x)$ the coefficients in the scheme located on the uppermost horizontal line yield the coefficients of the one-dimensional analysis of $u_1(x)$, while analogously those on the leftmost vertical line reproduce the one-dimensional analysis of $u(x,y) = u_2(y)$.

5.3. Transform of a two-dimensional test function. In what follows, the domain has been mapped to $\Omega = \pi\mathbb{T} \times [-1,1]$, which corresponds to the domain of the turbulence data we wish to analyze below in section 6.3. As an analytical test we consider a two-dimensional Gaussian bump,

$$\begin{aligned} u(x,y) &= g\,(\,(y-y_c)\sin(\alpha) + (x-x_c)\cos(\alpha)\,,\\ &\qquad (y-y_c)\cos(\alpha) - (x-x_c)\sin(\alpha)\,)\,,\\ g(x,y) &= \exp\left(-\frac{1}{2}\left(\frac{x}{\sigma_x}\right)^2 - \frac{1}{2}\left(\frac{y}{\sigma_y}\right)^2\right), \end{aligned} \tag{40}$$

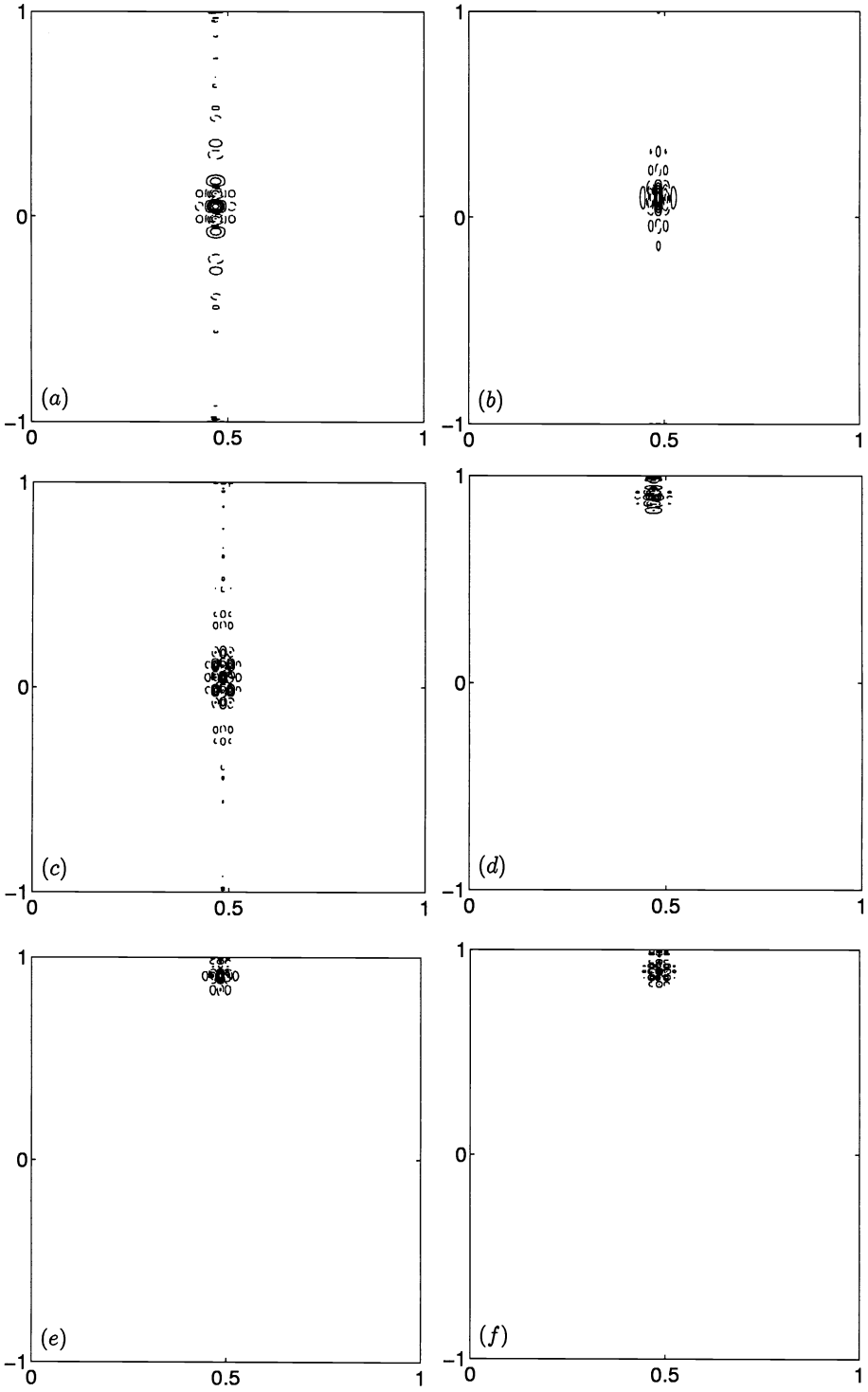


FIG. 16. *Illustration of the basis functions of the two-dimensional MRA (method* **B***) at scale number* $j = 5$*, sampled on a grid with* 256^2 *points.* (a) $q=1$, $i_y=16$, (b) $q=2$, $i_y=16$, (c) $q=3$, $i_y=16$, (d) $q=1$, $i_y=27$, (e) $q=2$, $i_y=27$, (f) $q=3$, $i_y=27$.

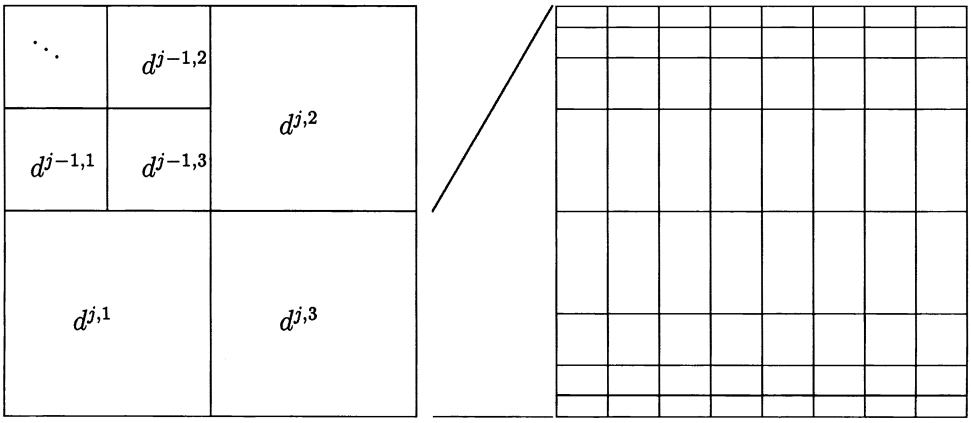


FIG. 17. *Graphical representation of the wavelet coefficients* $d^{j,q}_{i_x,i_y}$ *of the two-dimensional MRA (method* **B***) given in* (38)*. The enlargement on the right shows that the individual colored cells of each block of data (coefficients which have common* j *and* q *indices) are uniformly spaced in the horizontal direction and spaced according to the approximate definition of the Legendre wavelet centers in* (29) *in the vertical direction. Note that the position of* $d^{j,2}$ *and* $d^{j,1}$ *is interchanged with respect to* [25, 5].

centered around the position (x_c, y_c), having the two characteristic length scales σ_x and σ_y and possibly a rotation by an angle α.

The coefficient diagrams of the transforms for various values of the parameters y_c, σ_x, σ_y, α (Figures 18 to 21) demonstrate several characteristics of the present hybrid MRA which are common to methods **A** and **B**:

(i) In both cases **A** and **B**, the position of the bump can be correctly determined from the location with the largest scale index (or scale index pair) at which a significant response is obtained.

(ii) Due to the different localization properties of $\tilde{\psi}(x)$ and $\psi(y)$, the response appears more smeared out in the vertical direction than in the horizontal direction, especially when the bump is centered near the boundary of the interval.

(iii) Since the characteristic vertical scale s_y of the Legendre wavelets varies with the position index i_y, the response appears at lower values of j_x (respectively, at lower j for $q = 1$ with method **B**) when the bump is located closer to the boundary. Locally, however, both variants of our hybrid base still bear the strict hierarchical feature of the original MRA in the sense that at a fixed location the scale varies exponentially with the scale indices j_x, j_y (or j, respectively).

With method **A**, directional information is solely represented by the aspect ratio s_x/s_y of the basis functions. We define a global index I_s for the aspect ratio:

$$I_s = \frac{\sum_{j_x,j_y,i_x,i_y} (d^{j_x,j_y}_{i_x,i_y})^2 \, \dfrac{s_x(j_x)}{s_y(j_y,i_y)}}{\sum_{j_x,j_y,i_x,i_y} (d^{j_x,j_y}_{i_x,i_y})^2}. \tag{41}$$

The values obtained for various choices of the parameters in the example (40) are shown in Table 2. The fact that I_s takes a value significantly higher than unity

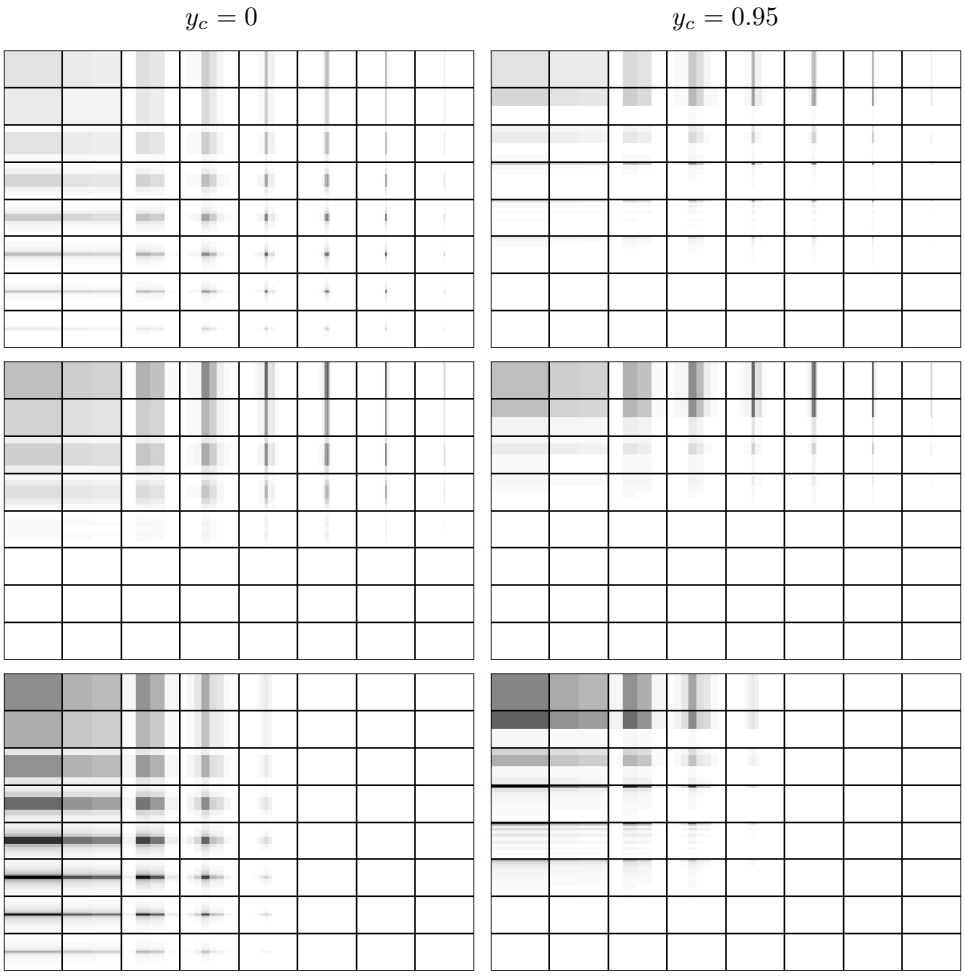


FIG. 18. *Coefficient schemes of the transform of a two-dimensional Gaussian bump using the tensor-product basis* $\mathbf{A}$ *with* $N = 256$ *modes.* $\sigma_x = \sigma_y = 0.01$ *(top row);* $\sigma_x = 0.01$, $\sigma_y = 0.1$ *(center);* $\sigma_x = 0.1$, $\sigma_y = 0.01$ *(bottom row). The grayscale representing the absolute value of the wavelet coefficients is linear.*

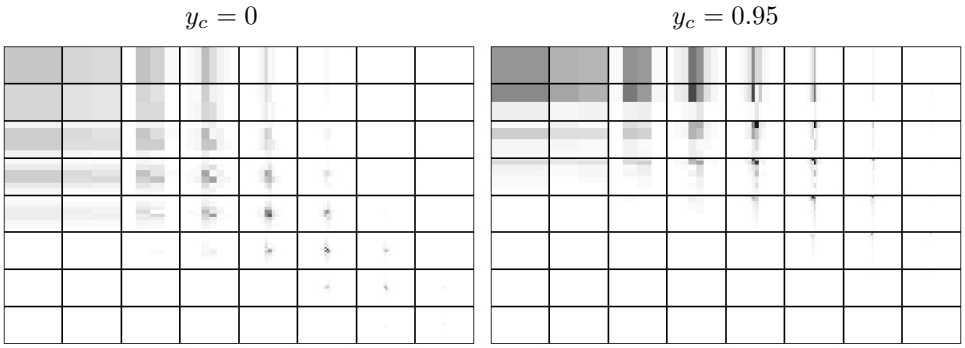


FIG. 19. *As Figure* 18*, but the Gaussian bump with* $\sigma_x = 0.1$, $\sigma_y = 0.01$ *has been rotated by* $-45°$ *with respect to the horizontal axis.*

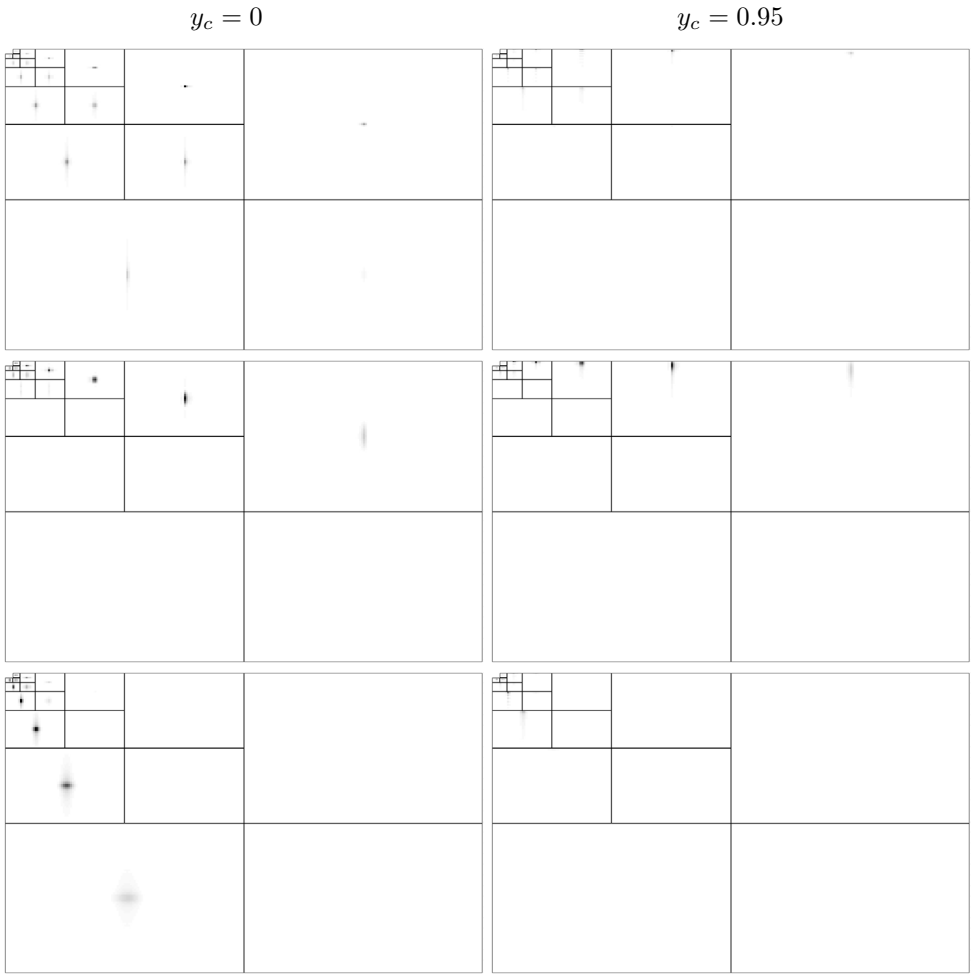


FIG. 20. *As Figure* 18*, but the transform has been performed with method* **B**.

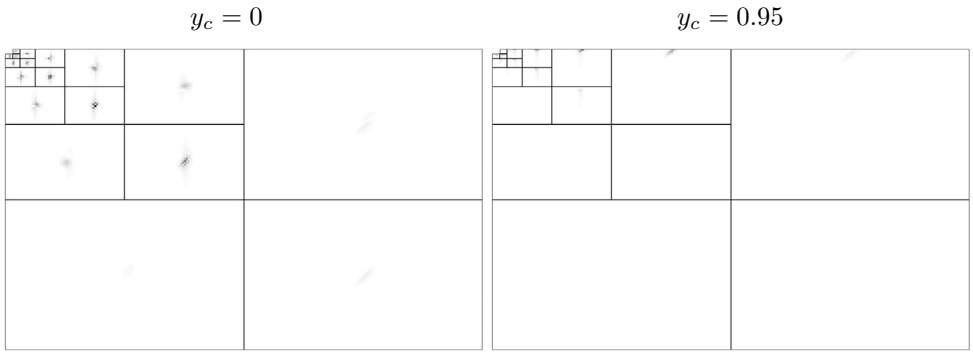


FIG. 21. *As Figure* 20*, but the Gaussian bump with* $\sigma_x = 0.1$*,* $\sigma_y = 0.01$ *has been rotated by* -45° *with respect to the horizontal axis.*

TABLE 2

Aspect ratio for different test signals according to the two-dimensional Gaussian bump (40) *determined by the index I_s for method* **A** *and index I_q for method* **B** *of constructing a two-dimensional basis. The transform was performed with $N = 256$.*

				Method **A**	Method **B**
y_c	σ_x	σ_y	α	I_s	I_q
0.0	0.01	0.01	0	3.43	1.14
0.95	0.01	0.01	0	3.25	5.4
0.0	0.01	0.1	0	0.41	21.0
0.95	0.01	0.1	0	0.23	57.5
0.0	0.1	0.01	0	19.45	0.06
0.95	0.1	0.01	0	18.45	0.23
0.0	0.1	0.01	$-45°$	1.36	1.24
0.95	0.1	0.01	$-45°$	0.88	26.0

although $\sigma_x/\sigma_y = 1$ again is a consequence of the different localization properties of the underlying one-dimensional wavelet bases used for the two coordinate directions. Important, however, is here that this index is approximately invariant with respect to a vertical shift of the center of the bump (except when $\sigma_x/\sigma_y = 1/10$ in which case a large part of the bump lies outside the upper boundary when $y_c = 0.95$). This quantity hence provides a useful characterization of the anisotropy of a signal. It can also be extended to a local coefficient or a scalewise coefficient by an appropriate restriction of the summation bounds in (41).

Turning now to method **B**, we remark that the angular selectivity (i.e., the index q) and the aspect ratio s_x/s_y are not independent but jointly represent the signal's directional properties. We therefore define the index

$$I_q = \frac{\sum_{j,i_x,i_y} (d^{j,1}_{i_x,i_y})^2}{\sum_{j,i_x,i_y} (d^{j,2}_{i_x,i_y})^2} \tag{42}$$

reflecting the ratio of the energies of coefficients with $q = 1$ and $q = 2$. Table 2 shows that, in contrast to I_s, the latter coefficient changes considerably when the signal is shifted vertically. Hence, changing y_c not only provokes a shift of the index j_y, but also causes the response to shift between the three wavelet types $q = 1, 2, 3$. Attempts to construct a robust joint directional index for method **B** failed, mainly due to the problem of attributing a meaningful physical scale to the scaling functions. Similar effects have to be expected for the doubly nonperiodic case. Because of the apparent difficulties in interpreting the coefficients from transform **B** in terms of orientation, we decided to work with method **A** from here on.

As discussed in section 4.5, interesting quantitative information can be extracted from the transformation by means of local spectra. For method **A** we therefore define the two-dimensional equivalent of the local power-spectral density (33) as

$$E(k_{j_x}, k_{j_y}, x, y) = 2^{j_x} 2^{j_y} \frac{\left(d^{j_x,j_y}_{i_{x^*},i_{y^*}}\right)^2}{\Delta k_{j_x}\, \Delta k_{j_y}}, \tag{43}$$

with, again, the indices i_{x^*}, i_{y^*} determined from the wavelet center nearest to the point (x, y).

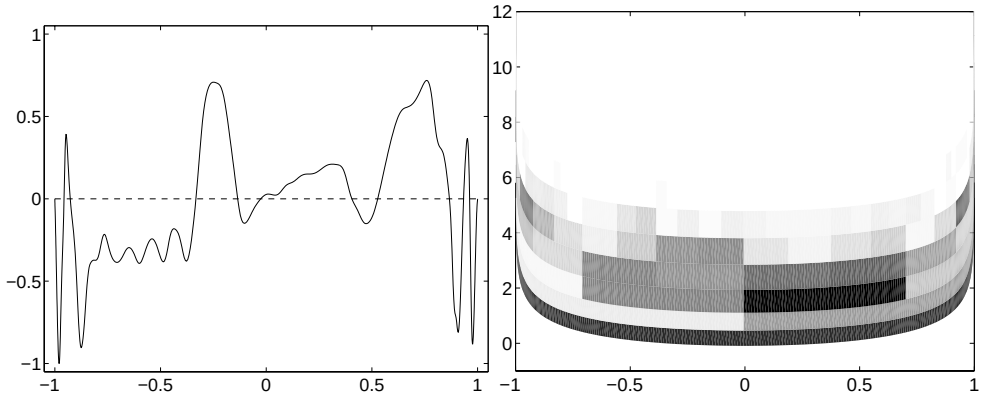


FIG. 22. *Signal and wavelet coefficient diagram of an instantaneous cut along a wall-normal line in a turbulent channel flow at friction velocity Reynolds number* $Re_\tau = 590$. *The signal is the fluctuation of the streamwise velocity component, normalized such that its maximum absolute value is unity.*

6. An application: Local scales in turbulent plane channel flow.

6.1. Flow configuration. We consider the fully developed turbulent flow in a doubly periodic box between two parallel walls which are spaced apart by $2h$. Here, x is the streamwise, y the wall-normal, and z the spanwise coordinate with u, v, w being the corresponding velocity components. Fluctuations with respect to the time-averaged signal are denoted by a prime. The characteristic Reynolds number $Re_\tau \equiv u_\tau h/\nu$ is based upon the wall-friction velocity $u_\tau = \sqrt{\nu|\overline{\partial u/\partial y}|_{y=\pm h}}$ (averaged over the channel walls and in time), the channel half-width h, and the kinematic viscosity ν. The same reference quantities are used to form a nondimensional length scale $l^+ = lu_\tau/\nu$, the so-called wall-scaling or wall units, which is the analogue to Kolmogorov scaling in homogeneous-isotropic turbulence and indicates the size of the smallest features in the flow close to the wall. The superscript + added when numerical values are given refers to quantities normalized by l^+.

We use flow data from a pseudospectral direct numerical simulation of the second author, performed at $Re_\tau = 590$ in a computational domain of size $\Omega_{DNS} = 2\pi h\mathbb{T} \times [-h, h] \times \pi h\mathbb{T}$ using $600 \times 385 \times 600$ discrete Fourier/Chebyshev modes, respectively. This case is similar to the highest Reynolds number case studied in [27], except that the spatial resolution has been increased substantially.

6.2. Local one-dimensional spectra. We extract instantaneous wall-normal profiles of velocity fluctuations and interpolate them spectrally to a $N = 256$ Legendre–Gauss–Lobatto grid before performing the wavelet transform given by (25).

In turbulent channel flow, high gradients and small structures are generated close to the solid surfaces. In the one-dimensional cuts these are hence located close to the extrema of the interval. Figure 22 displays such a snapshot, where the fluctuations of the streamwise velocity component u' show such features near $y=-1$ and $y=1$, both of which are not unlike the narrow Gaussian bumps considered in section 4.4. The wavelet coefficient scalogram again allows a localization of these peaks as well as an approximate determination of their relative scales. Several coarser undulations of u' towards the center of the interval produce responses at larger scales.

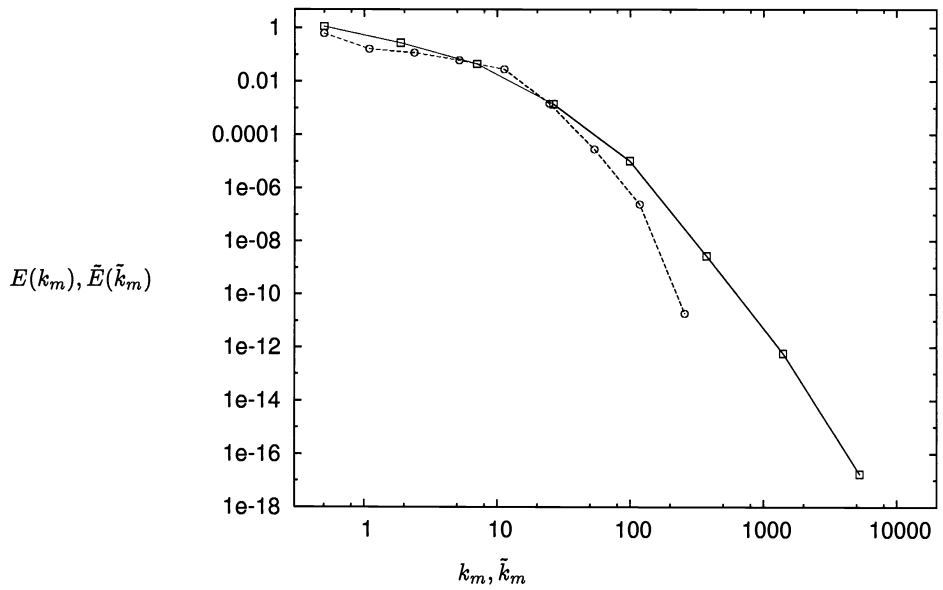


FIG. 23. *The global wavelet spectrum $E(k_m)$ as a function of the inverse of the scale number $k_m = 1/s_{ij}$ (solid line) corresponding to the transform of the data in Figure 22. The dashed line shows the Legendre coefficient spectrum $\tilde{E}(\tilde{k}_m)$. In both cases, the values of the spectrum are accumulated over logarithmically spaced bins and normalized such that their integral amounts to unity.*

Figure 23 shows the global wavelet energy spectrum for the signal reported in Figure 22. Also included is the corresponding Legendre coefficient spectrum. Both curves are normalized by the respective total energy E_{tot}. A very close comparison is delicate due to the different meaning of "scale" in both cases. The wavenumber associated with a Legendre polynomial is naturally defined by its degree n as $\tilde{k}_n = n/L_x$. The spacing of zeros and extrema, however, varies over the interval. Since these functions do not have local character, physical information from features of different size is summed up over the whole interval. The physical significance of this averaging over the nonhomogeneous flow direction is unclear and motivates the present approach for defining a global spectrum by means of wavelets. Here, it is assumed that only contributions from the same physical scale are accounted for in $E(k_m)$ (see (31)).

Figure 24 shows the local energy spectra of the streamwise velocity data from Figure 22, evaluated at different locations. Not surprisingly, close to the center the largest scales dominate the flow. About halfway towards the lower wall ($x = -0.6$) a distinct medium-scale peak is observed, while at $x = -0.94$ the maximum energy is recorded for even smaller scales. There are of course strong temporal variations and any significant statement about turbulent flow structures requires a statistical approach. This will be done in section 6.4 below.

6.3. Local two-dimensional spectra. Data in planes (containing the wall-normal and either the streamwise or the spanwise direction) extracted from the raw three-dimensional fields have been transformed to wavelet space by method **A**. To this end, we have first spectrally interpolated the data on a grid comprised of $N_1 = 512$ uniformly spaced points times $N_2 = 512$ Legendre–Gauss–Lobatto points before evaluating the scalar products in (37).

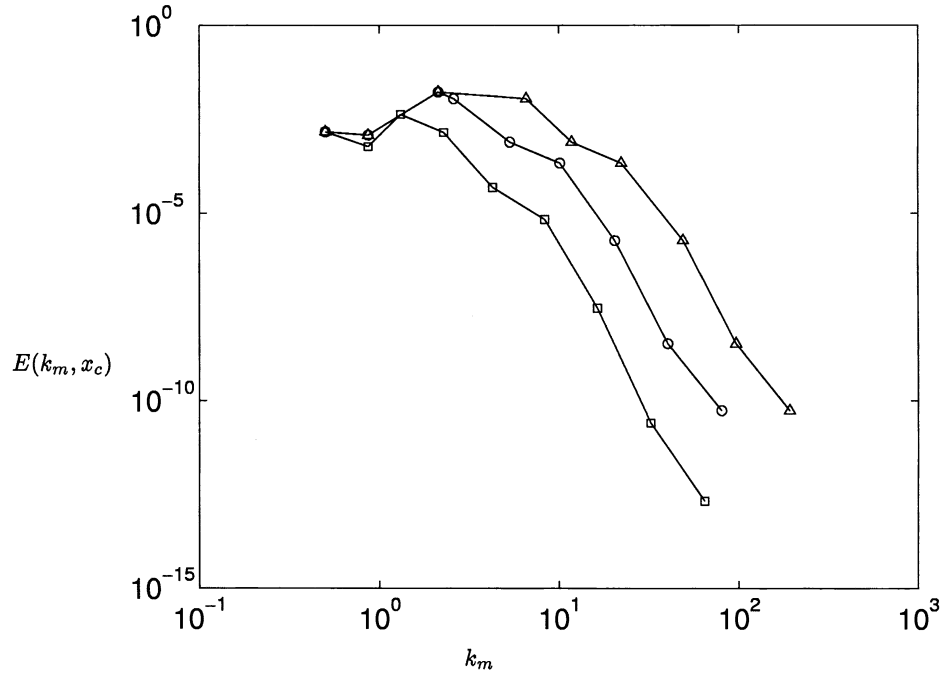


FIG. 24. *Local wavelet spectra $E(k_m, x_c)$ as a function of the inverse of the scale number $k_m = 1/s_{ij}$ corresponding to the data of Figure 22 and taken at the following positions within the interval $x \in [-1, 1]$: □, $x_c = 0.1$; ○, $x_c = -0.6$; △, $x_c = -0.94$.*

Figure 25 shows a snapshot of the streamwise velocity fluctuation in a spanwise/wall-normal plane. The signal bears a vast number of features and one clearly needs a formalism which helps to extract the desired information. Figure 26 displays the wavelet representation of this signal by means of the two-dimensional periodic/non-periodic transform defined above. Inspecting these coefficients, we see that the well-known small-scale intermittency [18] is apparent. For this purpose, the small-scale coefficients have been overexposed by multiplying all coefficient values with a factor of $2^{\max(j_x, j_y)}$ in the lower graph. This reveals that the high intensity regions become increasingly localized, i.e., less space-filling, with increasing values of $\max(j_x, j_y)$.

In the following we consider two locations in the near-wall region, i.e., at a wall-distance of $y^+ = 89$, one of which is located in what seems a lifted low-speed streak (marked "B" in Figure 25), the other (marked "A") well away from such events. The third location, marked "C," is situated on the centerline of the channel. Figures 27 and 28 show the two-dimensional power-spectral density evaluated at these locations and plotted for each value of the horizontal scale number k_{j_x} as a function of k_{j_y}. For clarity, only the higher horizontal scale numbers $k^+_{j_x} \geq 4.3 \cdot 10^{-3}$ are included.

Comparing the center to the wall region ("C" vs. "A"), it is visible that for all horizontal scale numbers k_{j_x} the decay of the signal's energy with increasing k_{j_y} is much faster in the center of the channel, especially at vertical scale numbers around $k^+_{j_y} = 10^{-2}$. This is an indication that smaller vertical scales of motion are active at point "A" while the horizontal scales are comparable to those at "C."

Comparing locations "A" and "B," more fine-scale contributions are observed at point "B" in vertical as well as horizontal direction. While at "A" a smooth and

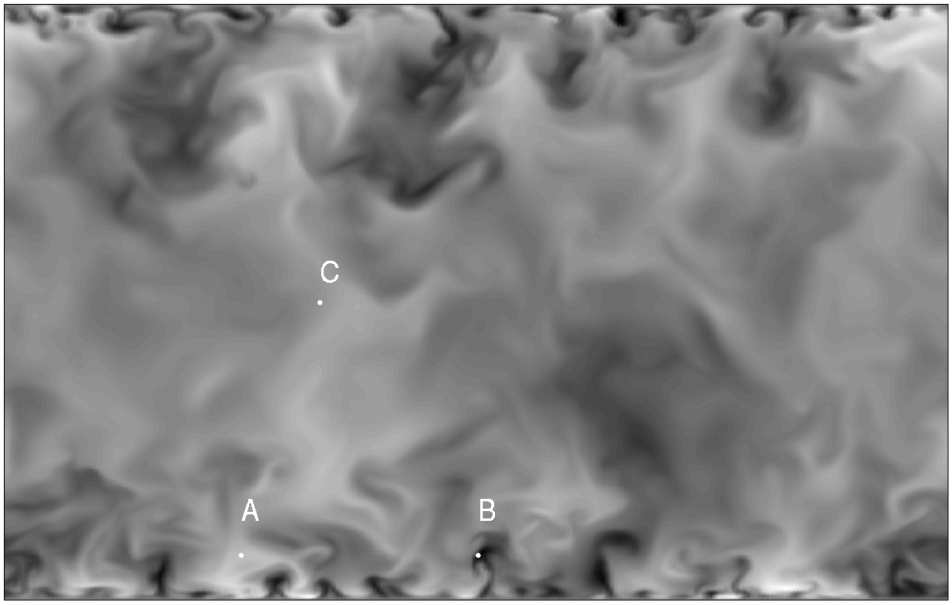


FIG. 25. *Streamwise velocity fluctuations of a snapshot from a turbulent channel flow at friction velocity Reynolds number $Re_\tau = 590$. The plane is spanwise/wall-normal, i.e., the mean flow is perpendicular to the plane. The aspect ratio reflects the physical size of the domain. The locations marked "A," "B," and "C" are used in subsequent graphs.*

continuous decay of the vertical coefficients is detected for all horizontal wavenumbers, the spectrum forms a plateau at "B," in particular for larger k_{j_x}. Furthermore, for $k^+_{j_x} = 8.6 \cdot 10^{-3} \ldots 3.5 \cdot 10^{-2}$ a local maximum is observed around $k^+_{j_y} = 1.5 \cdot 10^{-2}$. This corresponds to features in the flow having a size about the distance of this point from the wall which indeed is the case for such a low-speed streak. Towards higher wavenumbers k_{j_y} a regular decay of the energy is observed. Its rate at point "B" is only half of the one at point "A." Furthermore, the same values are observed for different k_{j_x} in the former case, which means that saturation is attained in the streamwise direction.

6.4. Statistical results for wall-normal scales. In the past, the wall-normal scales of turbulent flow have basically been characterized by two methods: inspection of two-point autocorrelations (e.g., [20]) or analysis of the most energetic modes from proper orthogonal decomposition (e.g., [23]). In neither case, a clear correspondence between an a priori defined length scale and its energy is available. The present wavelet basis for the interval provides just these two ingredients and thus allows for a quantitative description of the energy content of wall-normal scales and their position.

For the generation of statistical data we have used a total of $n_{stat} = 150$ streamwise/wall-normal planes from 5 flow fields covering a time span equivalent to one flow-through time of the computational domain. The quantities considered are the three velocity components u_α ($\alpha = 1 \ldots 3$) with their transform coefficients denoted as $d^{j_x,j_y}_{i_x,i_y}(u_\alpha)$. Furthermore, we focus on the distribution of the wall-normal scales and therefore sum over the indices j_x and i_x of the statistically homogeneous direction. This leads us to the following definition of the ensemble-averaged one-dimensional

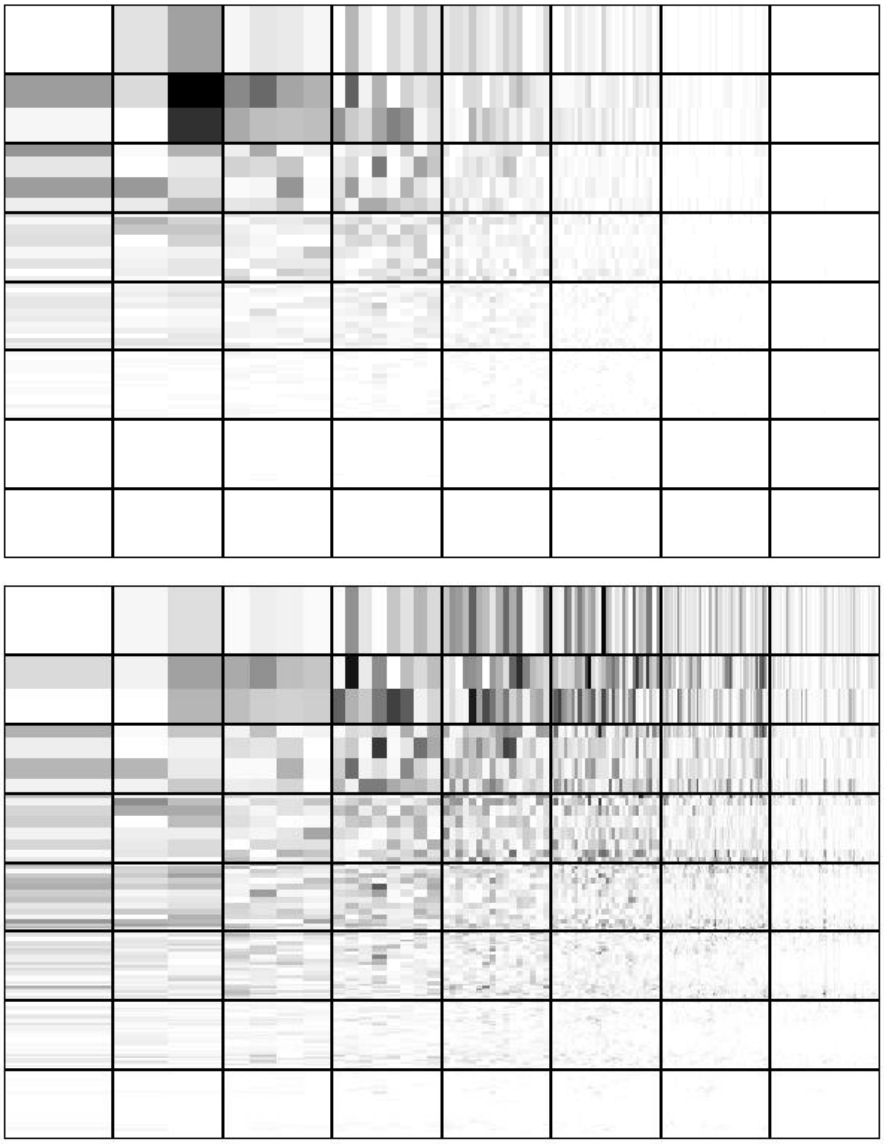


FIG. 26. *The absolute value of the wavelet coefficients of the transform of streamwise velocity fluctuations in a turbulent channel flow at friction velocity Reynolds number* $Re_\tau = 590$. *The plane is spanwise/wall-normal. The numerical grid has a dimension of* $N = 2^9$. *The grayscale coloring is chosen such that white corresponds to zero intensity and black to maximum intensity. Both graphs originate from the same data. In the lower graph, the small-scale coefficients are overexposed by a factor of* $2^{\max(j_x,j_y)}$.

power-spectral density as a function of wall-normal position y and wavenumber k_{j_y}:

$$E_{\alpha\alpha}(k_{j_y}, y) = 2^{j_y} \sum_{j_x,i_x} \frac{\left\langle \left(d^{j_x,j_y}_{i_{x_c},i_y}(u_\alpha) \right)^2 \right\rangle_{n_{stat}}}{\Delta k_{j_y}}, \tag{44}$$

where $\langle\cdot\rangle_{n_{stat}}$ denotes an average over the available samples.

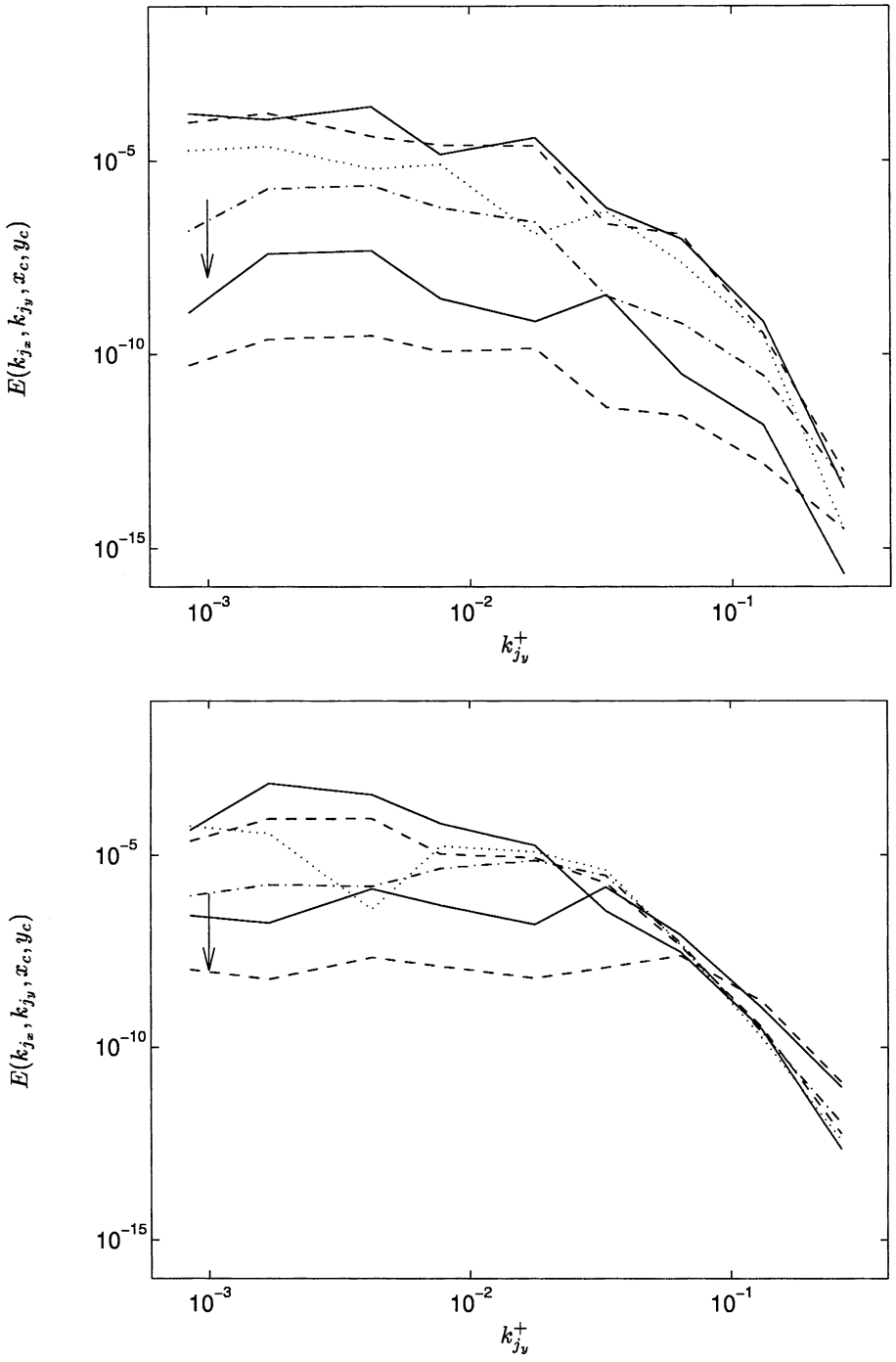


FIG. 27. *Local spectral energy density of the streamwise velocity fluctuations, determined by the two-dimensional wavelet analysis. The results presented refer locations "A" (top) and "B" (bottom), indicated in Figure* 25*, both at a wall-distance of* $y^+ = 89$*. The various lines correspond to different values of the spanwise (i.e., horizontal) scale number, stepping through* $k^+_{j_x} = 4.3 \cdot 10^{-3}$ *(solid),* $8.6 \cdot 10^{-3}$ *(dashed),* $1.7 \cdot 10^{-2}$ *(dotted),* $3.5 \cdot 10^{-2}$ *(dash-dotted),* $6.9 \cdot 10^{-2}$ *(solid),* 0.14 *(dashed) in the direction indicated by the arrow.*

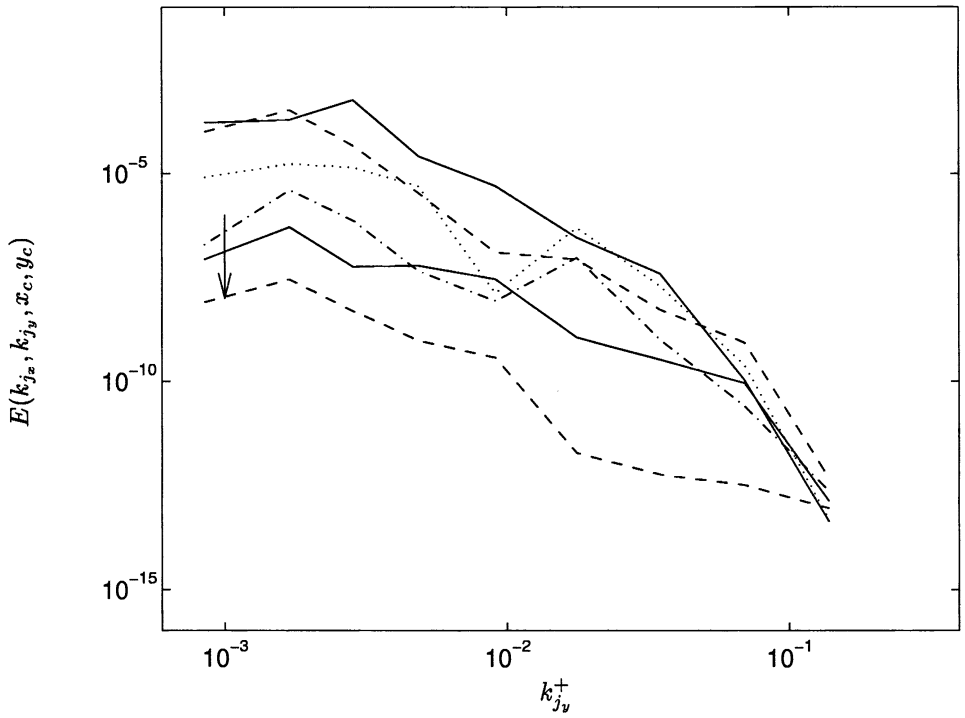


FIG. 28. *As Figure* 27*, but for location "C" on the centerline of the channel.*

Figure 29 shows the premultiplied spectra $k_{j_y} E_{\alpha\alpha}(k_{j_y}, y)$ for the streamwise ($\alpha = 1$) and the wall-normal component ($\alpha = 2$) of velocity, evaluated at different wall-distances across the channel. This representation allows for a direct comparison with results for the corresponding streamwise and spanwise spectra available through Fourier analysis. (For a compilation of experimental and numerical data cf. [17].) Here we observe a very distinct behavior of the velocity components. The wall-normal velocity has the maximum energy contained in scales which increase with the wall-distance, from a peak scale of 15 wall units at $y^+ = 5$ to the largest scale (i.e., the full channel width) at the centerline. Contrarily, the scale of the energy peak is *always* the largest scale in the case of the streamwise velocity fluctuations. These observations imply that the energetic scales of the wall-normal velocity are strongly constrained near the wall, while those of the streamwise velocity are affected only to a much lesser extent. It has long been recognized (e.g., [32, p. 150]) that wall-impermeability has a strong direct effect upon the wall-normal velocity in turbulent shear flow while not restricting the motion in planes parallel to the wall. With the present analytical tools it is finally possible to describe the above effect quantitatively. We believe that the local wall-normal spectrum could play a particularly important role during the analysis of new data from ongoing efforts to capture the dynamics of the very large scales of the logarithmic layer [6].

7. Conclusions. Starting from [11] we have constructed an orthonormal wavelet basis for the interval by an appropriate recombination of Legendre polynomials. These functions have been implemented together with routines to perform the corresponding

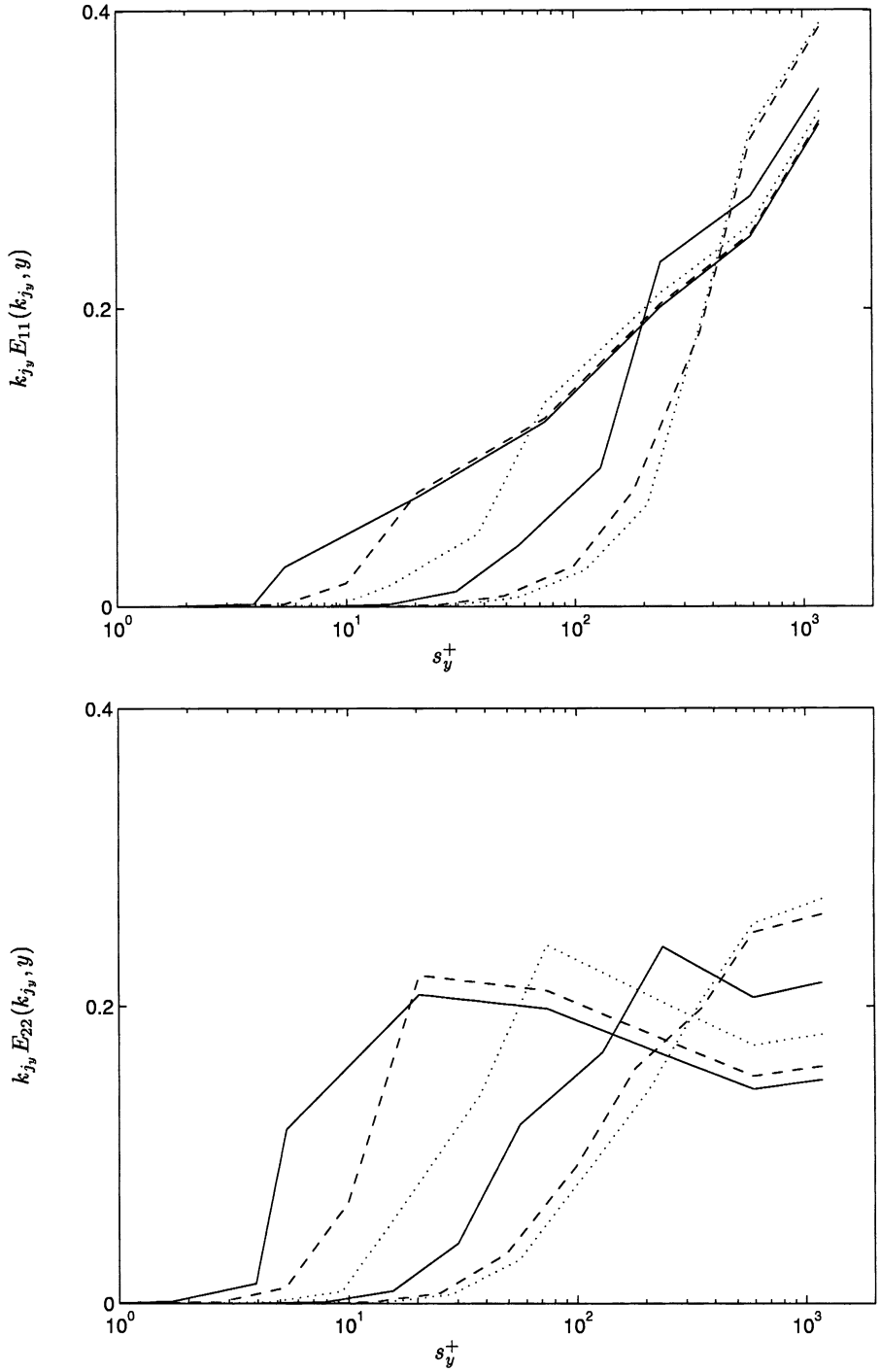


FIG. 29. *Premultiplied, ensemble-averaged, wall-normal power spectra* $k_{j_y} E_{\alpha\alpha}(k_{j_y}, y)$ *as a function of scale* s_y^+ *in plane channel flow at* $Re_\tau = 590$ *for streamwise* ($\alpha = 1$) *and wall-normal* ($\alpha = 2$) *velocity components. The different lines correspond to different wall distances* $y^+ = \{5, 10, 30, 100, 300, 500\}$*; in both graphs, increasing* y^+ *results in a shift towards larger scales while line styles rotate through solid, dashed, dotted. The spectra are normalized to unit area.*

forward and backward transform. As a consequence of the inhomogeneity, the spatial scale of the wavelets depends to some extent not only on the scale index but also on the translation index. We have devised a suitable definition of the scale number and developed a scheme for the representation and analysis of the coefficients. The usefulness of the present basis for data analysis is demonstrated by studying the transforms of analytical functions as well as data from turbulent flow simulations. We have defined local power-spectral density functions and find that they constitute an important tool for the analyst. We have argued that with the present lumping of blocks of polynomials no decay better than $\propto 1/x$ is possible. While the rate of decay cannot be changed, the actual values and the properties of the tails might be improved. Another route is to modify the lumping by using a smoother selection of polynomial coefficients as employed in [14]. This work is currently under way.[1]

In a second part of the paper two variants of a hybrid two-dimensional MRA have been proposed and implemented. Method **A** is constructed from a tensor product between periodic spline wavelets in the first direction and the present Legendre wavelets in the second direction. Method **B** uses the more widespread procedure of a tensor product of the two corresponding one-dimensional MRAs, leading to three types of wavelets with different directional properties. In both cases, the implementation of two nonperiodic directions can be accomplished analogously by using Legendre wavelets in both directions. Higher dimensions are also straightforward.

The graphical representation of a two-dimensional wavelet analysis is genuinely more difficult than in the one-dimensional case. We have discussed the implications of a spatially varying scale parameter and performed visualizations with an adaptation of the classical scheme combining coefficients with the same scale and, in the case of method **B**, direction indices in blocks. It was found that due to the interaction between directional properties and varying scale ratio s_x/s_y, method **B** is less useful for the purpose of data analysis.

Finally, we have applied the new transform (method **A**) to the analysis of data from direct numerical simulation of turbulent plane channel flow. The qualitative analysis of intermittency of a plane extracted from a snapshot showed velocity fluctuations which are more intermittent at small scales than at large scales—an observation which is consistent with previous wavelet analyses of spatially homogeneous flows [26, 3, 21]. We then performed ensemble-averaging and reduced the data to the form of local wall-normal power spectra. It was found that near the wall, the most energetic scales of the wall-normal velocity are much smaller than those of the streamwise velocity, probably due to the constraining effect of wall-impermeability.

The constructions presented in this paper offer numerous perspectives for future extensions. One direction is the construction itself in the one-dimensional and multidimensional cases, such as a variant for semi-infinite intervals, and its optimized implementation. A second direction is the application to signals of various other areas and the definition of further secondary quantities based on the wavelet coefficients.

Acknowledgments. The authors like to thank J. Prestin for pointing them to the possibility of replacing the Chebyshev polynomials by Legendre polynomials in the wavelet construction.

[1] This construction has been completed in the meantime. A copy of the manuscript can be obtained from the authors upon request.

REFERENCES

[1] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions*, Dover, New York, 1964.

[2] R. Askey, *Orthogonal Polynomials and Special Functions*, SIAM, Philadelphia, 1975.

[3] J. Brasseur and Q. Wang, *Structural evolution of intermittency and anisotropy at different scales analyzed using three-dimensional wavelet transforms*, Phys. Fluids A, 4 (1992), pp. 2538–2554.

[4] A. Cohen, I. Daubechies, and P. Vial, *Wavelets on the interval and fast wavelet transform*, Appl. Comput. Harmon. Annal., 1 (1994), pp. 54–81.

[5] I. Daubechies, *Ten Lectures on Wavelets*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 61, SIAM, Philadelphia, 1992.

[6] J. del Álamo and J. Jiménez, *Direct numerical simulation of the very large anisotropic scales in a turbulent channel*, Ann. Research Briefs, Center for Turbulence Research, Stanford University, Stanford, CA, (2001), pp. 329–341.

[7] M. Do-Khac, C. Basdevant, V. Perrier, and K. Dang-Tran, *Wavelet analysis of* $2d$ *turbulent fields*, Phys. D, 76 (1994), pp. 252–277.

[8] M. Farge, *Wavelet transforms and their applications to turbulence*, Ann. Rev. Fluid Mech., 24 (1992), pp. 395–457.

[9] M. Farge, Y. Guezennec, C. Ho, and C. Meneveau, *Continuous wavelet analysis of coherent structures*, in Proceedings of the CTR Summer Program, Center for Turbulence Research, Stanford University, Stanford, CA, 1990, pp. 331–348.

[10] M. Farge, N. Kevlahan, V. Perrier, and E. Goirand, *Wavelets and turbulence*, Proc. IEEE, 84 (1996), pp. 639–669.

[11] B. Fischer and J. Prestin, *Wavelets based on orthogonal polynomials*, Math. Comp., 66 (1997), pp. 1593–1618.

[12] J. Fröhlich and K. Schneider, *An adaptive wavelet Galerkin algorithm for one- and two-dimensional flame computations*, Eur. J. Mech. B Fluids, 13 (1994), pp. 439–471.

[13] J. Fröhlich and K. Schneider, *Computation of decaying turbulence in an adaptive wavelet basis*, Phys. D, 134 (1999), pp. 337–361.

[14] R. Girgenson and J. Prestin, *Lebesgue constants for an orthogonal polynomial Schauder basis*, Comp. Anal. Appl., 2 (2000), pp. 159–175.

[15] D. Gottlieb and S. A. Orszag, *Numerical Analysis of Spectral Methods: Theory and Applications*, SIAM, Philadelphia, 1977.

[16] B. Jawerth and W. Sweldens, *An overview of wavelet based multiresolution analyses*, SIAM Rev., 36 (1994), pp. 377–412.

[17] J. Jiménez, *The largest scales of turbulent wall flows*, CTR Res. Briefs, (1998), pp. 137–154.

[18] J. Jiménez and A. Wray, *On the characteristics of vortex filaments in isotropic turbulence*, J. Fluid Mech., 373 (1998), pp. 255–282.

[19] T. Kilgore and J. Prestin, *Polynomial wavelets on the interval*, Constr. Approx., 12 (1996), pp. 95–110.

[20] J. Kim, *On the structure of pressure fluctuations in simulated channel flow*, J. Fluid Mech., 205 (1989), pp. 421–451.

[21] K. Kishida, *Analysis of Turbulence in the Orthonormal Divergence-Free Wavelet Representation*, Ph.D. thesis, Hiroshima University, Hiroshima, Japan, 2000.

[22] K. Kishida, K. Araki, S. Kishiba, and K. Suzuki, *Local or nonlocal? Orthonormal divergence-free wavelet analysis of nonlinear interactions in turbulence*, Phys. Rev. Letters, 83 (1999), pp. 5487–5490.

[23] Z. Liu, R. Adrian, and T. Hanratty, *Large-scale modes of turbulent channel flow: Transport and structure*, J. Fluid Mech., 448 (2001), pp. 53–80.

[24] Y. Maday and J. Ravel, *Adaptativité par ondelettes: conditions aux limites et dimensions supérieures*, C.R. Acad. Sci. Paris Sér. I Math., 315 (1992), pp. 85–90.

[25] S. Mallat, *A theory for multiresolution signal decomposition: The wavelet representation*, IEEE Trans. Pattern Analysis Mach. Intell., 11 (1989), pp. 674–693.

[26] C. Meneveau, *Analysis of turbulence in the orthonormal wavelet representation*, J. Fluid Mech., 232 (1991), pp. 469–520.

[27] R. Moser, J. Kim, and N. Mansour, *Direct numerical simulation of turbulent channel flow up to* $Re_\tau = 590$, Phys. Fluids, 11 (1999), pp. 943–945.

[28] W. Nitsche, H. Thiele, M. Farge, C. Pellegrino, and K. Schneider, *Wavelet filtering of three-dimensional turbulence*, ZAMM Z. Angew. Math. Mech., 81 (2001), p. 465.

[29] V. Perrier and C. Basdevant, *Periodical wavelet analysis, a tool for inhomogeneous field investigation. Theory and algorithms*, Rech. Aérosp., 3 (1989), pp. 54–67.

[30] V. Perrier, T. Philipovich, and C. Basdevant, *Wavelet spectra compared to fourier spectra*, J. Math. Phys., 36 (1995), pp. 1506–1519.
[31] J. Prestin and K. Selig, *Interpolatory and orthonormal trigonometric wavelets*, in Signal and Image Representation in Combined Spaces, J. Zeevi and R. Coifman, eds., Academic Press, San Diego, CA, 1998, pp. 201–255.
[32] A. Townsend, *The Structure of Turbulent Shear Flow*, 2nd ed., Cambridge University Press, Cambridge, UK, 1976.

SIAM J. APPL. MATH.
Vol. 63, No. 5, pp. 1831–1848

INITIATION OF FREE-RADICAL POLYMERIZATION WAVES*

L. R. RITTER†, W. E. OLMSTEAD†, AND V. A. VOLPERT†

Abstract. Frontal polymerization is a process of converting a monomer into a polymer by means of a self-propagating thermal reaction wave. We study initiation of polymerization waves by a high temperature heat source. A five species reaction model is considered with a focus on the evolution of two of these species and the temperature of the mixture. The temperature is tracked from the inert heating to the transition stage. Through an asymptotic analysis, the first correction to the temperature in transition is found as the solution to an integral equation. Two parameters govern the qualitative behavior of the solution to the integral equation. Depending on the magnitude of these parameters, the solution exhibits either bounded or unbounded behavior indicating the onset or inhibition of propagation of a polymerization wave.

Key words. frontal polymerization, integral equation, initiation

AMS subject classifications. 35, 45, 79

DOI. 10.1137/S0036139902413246

1. Introduction.

Frontal polymerization is the process of converting a monomer into a polymer by means of a self-propagating high temperature reaction wave. The chemical process involves two species: a monomer and an initiator, which is needed to start the growth of polymer chains. In a typical experiment, the species are placed into a test tube, and the temperature at the end of the tube is increased by applying a heat source. The increase in temperature induces decomposition of the initiator, which produces active radicals, and the highly exothermic propagation process begins. The resulting heat release promotes initiator decomposition ahead of the front, and a self-sustained reaction wave travels through the mixture leaving polymer in its wake.

Experimental and theoretical studies of frontal polymerization began in the 1970s (see references 1–4 in [5]). In [5] and [6], a mathematical model for the five species reaction is presented, and traveling wave solutions are sought. In these theoretical examinations of the process, the focus has been on the propagation of the thermal front and its velocity, the spatial profiles of the species involved, the degree of conversion of monomer, and the final temperature of the mixture. Initiation of a polymerization front is presumed. From experimental work, however, it is found that initiation of the front does not always occur. It is desirable to determine the dependence of initiation on the amount of reactants at the onset of the experiment, the initial temperature, the heat control imposed at the end of the test tube, and the properties of the initiator.

The purpose of this paper is to examine the initiation process necessary for propagation. In this respect, the current study is similar to ignition considerations in solid phase combustion problems. Unlike the combustion problem with a single reactant, the frontal polymerization process involves several chemical reaction steps with different reaction rates and activation energies. However, the reaction mechanism in both types of problem is assumed to be Arrhenius, and, upon nondimensionalization of the kinetic equations governing frontal polymerization, we can obtain a system of partial

*Received by the editors August 19, 2002; accepted for publication (in revised form) February 10, 2003; published electronically August 6, 2003. This work was supported in part by the National Science Foundation under grant DMS-0103856.
http://www.siam.org/journals/siap/63-5/41324.html
†Department of Engineering Science and Applied Mathematics, Northwestern University, Evanston, IL 60208-3125 (lritter67@northwestern.edu, weo@nwu.edu, v-volpert@northwestern.edu).

differential equations of a form similar to those arising in solid phase combustion. For this reason, the techniques applied in the current analysis are similar to those employed in [1, 2, 3, 4] in the examination of ignition of a combustible half-space. Lasseigne and Olmstead [4] consider the effects of reactant consumption, and they derive an integral equation governing the temperature at the ignition site. In this paper, we show that the mechanism governing initiation of the polymerization front gives rise to a similar two parameter integral equation governing the temperature in the transitional heating stage. In fact, under certain limiting conditions, the integral equation in [4] for a first order Arrhenius reaction appears as a special case of the integral equation presented in the current work. An asymptotic analysis of this integral equation is given, and numerical results are presented.

2. The mathematical model. The typical experiment in free radical frontal polymerization involves placing a mixture of initiator and monomer into a test tube. Assuming that the cross-sectional area of the tube is small relative to its length, we can model the tube as a thin semi-infinite channel $\hat{x} \geq 0$. A boundary condition on the heat flux will be prescribed at the end ($\hat{x} = 0$). The evolution of the reactants and the temperature can then be tracked. A mathematical model for a five species reaction is derived in [5] and [6], and the following system of equations governing the kinetics at time $\hat{t}$ in dimensional coordinates is given:

$$\frac{dI}{d\hat{t}} = -k_d I, \tag{2.1}$$

$$\frac{dR}{d\hat{t}} = 2fk_d I - k_i RM - k_e RR_p, \tag{2.2}$$

$$\frac{dM}{d\hat{t}} = -k_i RM - k_p MR_p, \tag{2.3}$$

$$\frac{dR_p}{d\hat{t}} = k_i RM - k_e RR_p - 2k_t R_p^2, \tag{2.4}$$

$$\frac{dP}{d\hat{t}} = k_e RR_p + k_t R_p^2. \tag{2.5}$$

The five species are the initiator, free radicals, monomer, polymer radicals, and the final polymer, denoted by I, R, M, R_p, and P, respectively. The parameter f appearing in the second equation is the ratio of primary radicals in the polymer to the primary radicals formed by the initiator. In practice, its value is taken to be $1/2$ (see [5]). The quantities, written above as k with a subscript, are assumed to have an Arrhenius dependence on the temperature T of the system. Thus, they can be expressed as

$$k_\alpha(T) = k_\alpha^0 \exp\left(\frac{-E_\alpha}{R_g T}\right) \quad \text{for} \quad \alpha = d, i, p, e, t.$$

Here, k_α^0 is the frequency factor, E_α is the activation energy for the corresponding reaction, and R_g is the universal gas constant. The subscripts correspond to the five reaction steps—initiator decomposition d, polymer chain initiation i, chain propagation p, free radical termination e, and polymer radical termination t.

To formulate the heat balance in the system, we note that the decomposition step is slightly endothermic but that each of the four subsequent reactions is exothermic. However, the most significant heat release occurs in the propagation step [9]. Thus, only this contribution to the net energy of the system will be considered here. Letting

$T(\hat{t},\hat{x})$ denote the temperature of the mixture at time $\hat{t}$ and at the point $\hat{x}$, $\kappa > 0$ the thermal diffusivity of the mixture, and $q > 0$ the increase in temperature induced per unit reacted monomer, we can write the following reaction diffusion equation governing the temperature:

$$\frac{\partial T}{\partial \hat{t}} = \kappa \frac{\partial^2 T}{\partial \hat{x}^2} - q\frac{\partial M}{\partial \hat{t}}. \tag{2.6}$$

Equations (2.1)–(2.6), together with appropriate initial and boundary conditions, completely describe the state of the mixture. Because we are interested in initiation of a polymerization front, we will consider a reduced system obtained by imposing the quasi-steady-state assumption (QSSA) [6], reducing the number of unknowns as in [5] and [6], and considering only the evolution of the initiator, the monomer, and the temperature. The QSSA states that the level of free and polymer radicals in the mixture is nearly constant. Hence, we set $(d/d\hat{t})(R + R_p) = 0$. In addition, we make the following simplifying assumptions as justified in [6]:

$$k_i = k_p,\ k_e = k_t, \quad \text{and} \quad R_p \gg R.$$

Summing (2.2) and (2.4) and making the aforementioned assumptions yields

$$R + R_p \approx \sqrt{\frac{2fk_d}{k_t}}\sqrt{I}.$$

Then (2.3) becomes

$$\frac{dM}{d\hat{t}} = -k_p\sqrt{\frac{2fk_d}{k_t}}M\sqrt{I}.$$

Noting that the coefficient in front of M in the above equation is an Arrhenius exponential motivates the following convenient notation for the effective reaction rate:

$$k_{eff} = k_p\sqrt{\frac{2fk_d}{k_t}}, \quad k^0_{eff} = k^0_p\sqrt{\frac{2fk^0_d}{k^0_t}}, \quad \text{and} \quad E_{eff} = \frac{1}{2}(E_d - E_t) + E_p.$$

The initial amounts of monomer and initiator present are known and will be denoted by M_0 and I_0. Similarly, the initial temperature of the system is given as T_0. In the current work, we will assume that the boundary condition on the temperature at $\hat{x} = 0$ will be a Neumann condition. That is, the heat flux is prescribed as

$$\frac{\partial T}{\partial \hat{x}} = -\hat{h}(\hat{t}) \quad \text{for} \quad \hat{x} = 0,\ \hat{t} > 0.$$

Further, we assume that $\hat{h}(\hat{t}) > 0$ for all $\hat{t}$. This restriction implies an energy input at the end of the test tube. Finally, the temperature far from the end is assumed to be equal to the initial temperature. The reduced, dimensional form of the system to be studied can then be written as

$$\frac{\partial I}{\partial \hat{t}} = -k_d(T)I, \quad I(0) = I_0, \tag{2.7}$$

$$\frac{\partial M}{\partial \hat{t}} = -k_{eff}(T)M\sqrt{I}, \quad M(0) = M_0, \tag{2.8}$$

$$\frac{\partial T}{\partial \hat{t}} = \kappa\frac{\partial^2 T}{\partial \hat{x}^2} + qk_{eff}(T)M\sqrt{I}, \quad T(0,\hat{x}) = T_0, \quad \hat{x} \geq 0, \tag{2.9}$$

$$\frac{\partial T(\hat{t},0)}{\partial \hat{x}} = -\hat{h}(\hat{t}), \quad \text{and} \quad T \to T_0 \quad \text{as} \quad \hat{x} \to \infty. \tag{2.10}$$

3. Scaling and nondimensionalization. Because the activation energies are relatively large, the Arrhenius reaction terms are insignificant, provided that the temperature is relatively small. Thus, we will consider a critical value of the temperature T_c at which the reaction terms become appreciable. The value of T_c will be made more precise later. Further, the largeness of the activation energies facilitates a perturbation scheme in solving for the temperature. Hence, we introduce the small parameter

$$\epsilon = \frac{R_g T_c}{E_{eff}}$$

and define the quantities

$$r = \frac{E_d}{E_{eff}}, \qquad \tilde{k}_d^0 = k_d^0 e^{-r/\epsilon}, \qquad \tilde{k}_{eff}^0 = k_{eff}^0 e^{-1/\epsilon},$$

$$t_* = (\tilde{k}_{eff}^0 \sqrt{I_0}\,)^{-1}, \qquad x_* = \sqrt{\kappa t_*}.$$

We also introduce the nondimensional variables

$$\phi = \frac{I}{I_0}, \quad \psi = \frac{M}{M_0}, \quad \theta = \frac{T}{T_c}, \quad \theta_0 = \frac{T_0}{T_c},$$
$$h(t) = \frac{x^*}{T_c}\hat{h}(\hat{t}), \quad t = \frac{\hat{t}}{t_*}, \quad \text{and} \quad x = \frac{\hat{x}}{x_*}.$$

From (2.7)–(2.9), we obtain the corresponding nondimensional system:

$$\frac{\partial \phi}{\partial t} = -A\phi \exp\left\{\frac{r}{\epsilon}\left(1 - \frac{1}{\theta}\right)\right\}, \quad \phi(0) = 1, \tag{3.1}$$

$$\frac{\partial \psi}{\partial t} = -\psi\sqrt{\phi} \exp\left\{\frac{1}{\epsilon}\left(1 - \frac{1}{\theta}\right)\right\}, \quad \psi(0) = 1, \tag{3.2}$$

$$\frac{\partial \theta}{\partial t} = \frac{\partial^2 \theta}{\partial x^2} + B\psi\sqrt{\phi} \exp\left\{\frac{1}{\epsilon}\left(1 - \frac{1}{\theta}\right)\right\}, \quad \theta(0, x) = \theta_0, \tag{3.3}$$

$$\frac{\partial \theta(t,0)}{\partial x} = -h(t), \quad \text{and} \quad \theta \to \theta_0 \quad \text{as} \quad x \to \infty. \tag{3.4}$$

The additional nondimensional parameters A and B appearing in (3.1) and (3.3) are defined by

$$A = \frac{\tilde{k}_d^0}{\tilde{k}_{eff}^0 \sqrt{I_0}} \quad \text{and} \quad B = \frac{M_0 q}{T_c}.$$

The role of initiator consumption in the possible inhibition of initiation is inherent in the scaling of these two parameters. If A is large, for example, we can expect that the amount of initiator will rapidly decay. This rapid decay or an insufficient quantity of initiator at the onset of the experiment will cause the reaction to stop before a thermal front can develop. Similarly, if B is small, the effect of the reaction term in (3.3) is decreased. This can result in insufficient heat to initiate and maintain propagation of the polymer chain. In the present analysis, the following scaling will be assumed:

$$A = A_0 \epsilon^{-1} \quad \text{and} \quad B = B_0 \epsilon^{-\frac{1}{2}},$$

with $A_0 = O(1)$ and $B_0 = O(1)$ with respect to ϵ. The numerical values of A, B, and ϵ depend on the choice of reactants, their kinetic properties, and the conditions of the

experiment (e.g., pressure and ambient temperature). Extensive tabulated values of activation energies, preexponential factors, and other kinetic parameters for various initiators and monomers can be found in [10]. For typical values of the physical parameters appearing in (2.7)–(2.10), the value of ϵ is expected to be in the range of 10^{-4} to 10^{-3}. Moreover, at room temperature the values of A_0 and B_0 can range between 0.01 and 10. Given the typical range of values for ϵ, this is consistent with the assumption that A_0 and B_0 are $O(1)$ with respect to ϵ.

For fixed A_0, the quantity T_c is defined by the relation

$$A = \frac{k_d(T_c)}{k_{eff}(T_c)\sqrt{I_0}}. \tag{3.5}$$

Equations (3.1) and (3.2) are separable and can be solved explicitly. We have

$$\phi(t) = \exp\left(-A\int_0^t e^{\frac{r}{\epsilon}\left(1-\frac{1}{\theta}\right)}\,ds\right),$$
$$\psi(t) = \exp\left(-\int_0^t e^{\frac{1}{\epsilon}\left(1-\frac{1}{\theta}\right)} \times e^{-\frac{A}{2}\int_0^s e^{\frac{r}{\epsilon}\left(1-\frac{1}{\theta}\right)}\,dq}\,ds\right).$$

Upon substitution of the above into the boundary value problem (3.3)–(3.4), the system reduces to one involving only a single dependent variable. In the next section, an asymptotic solution to (3.3)–(3.4) will be derived.

4. Solving for the temperature. As stated, we consider the initial temperature to be small so that the reaction terms are negligible at the onset of the experiment—during the inert heating stage. In the formulation above, this means that we take $\theta_0 < 1$ and $1 - \theta_0 = O(1)$ with respect to ϵ. This allows us to initially ignore the Arrhenius term, which is mathematically equivalent to taking the limit $\epsilon \to 0$ in (3.1)–(3.3). Let θ_I be given by

$$\theta_I(t,x) = \theta_0 + \int_0^t h(\tau)\frac{e^{\frac{-x^2}{4(t-\tau)}}}{\sqrt{\pi(t-\tau)}}\,d\tau.$$

Then θ_I solves the problem (3.3)–(3.4) in the limit $\epsilon \to 0$; we will call this the inert solution. From $0 < 1 - \theta_0$ and $1 - \theta_0 = O(1)$, it follows that initially

$$\theta = \theta_I + \text{e.s.t.},$$

where e.s.t. represents terms that are exponentially small with respect to ϵ. However, this remains valid only until such time as $\theta_I \approx 1$. In order to continue the analysis, let us define the critical time t_c to be the smallest value such that

$$1 = \theta_I(t_c, 0) = \theta_0 + \int_0^{t_c} \frac{h(\tau)}{\sqrt{\pi(t_c-\tau)}}\,d\tau.$$

For arbitrary $h(t)$, such a critical time need not exist. This suggests a restriction on the class of boundary conditions that can lead to initiation. We will assume that the imposed flux $h(t)$ given is such that this critical time does exist. Also note that the above is evaluated at $x = 0$ because θ_I attains its maximum at the end $x = 0$. The inert stage of the reaction ends in the neighborhood of $(t_c, 0)$, and the system enters a transition stage where the reaction terms first become appreciable. To further

our investigation, we perturb about this point and introduce the new independent variables

$$\xi = \frac{x}{\epsilon}, \qquad \tau = \frac{t - t_c}{\epsilon}.$$

In these inner variables, (3.1)–(3.3) become

$$\phi_\tau = -A_0 \phi \exp\left\{\frac{r}{\epsilon}\left(1 - \frac{1}{\theta}\right)\right\}, \quad \phi \to 1 \quad \text{as} \quad \tau \to -\infty, \tag{4.1}$$

$$\psi_\tau = -\epsilon \psi \sqrt{\phi} \exp\left\{\frac{1}{\epsilon}\left(1 - \frac{1}{\theta}\right)\right\}, \quad \psi \to 1 \quad \text{as} \quad \tau \to -\infty, \tag{4.2}$$

$$\epsilon\theta_\tau = \theta_{\xi\xi} + \epsilon^{3/2} B_0 \psi \sqrt{\phi} \exp\left\{\frac{1}{\epsilon}\left(1 - \frac{1}{\theta}\right)\right\}, \quad \theta \to \theta_0 \quad \text{as} \quad \tau \to -\infty, \tag{4.3}$$

$$\theta_\xi = O(\epsilon) \quad \text{for} \quad \xi = 0 \quad \text{and} \quad \tau > -\infty. \tag{4.4}$$

We note here that the conditions at $t = 0$ in the outer variables correspond asymptotically to conditions in the inner variables as $\tau \to -\infty$. The first two equations can again be solved to obtain

$$\phi(\tau) = \exp\left(-A_0 \int_{-\infty}^{\tau} e^{\frac{r}{\epsilon}\left(1 - \frac{1}{\theta}\right)}\, ds\right), \tag{4.5}$$

$$\psi(\tau) = \exp\left(-\epsilon \int_{-\infty}^{\tau} e^{\frac{1}{\epsilon}\left(1 - \frac{1}{\theta}\right)} \times e^{-\frac{A_0}{2}\int_{-\infty}^{s} e^{\frac{r}{\epsilon}\left(1 - \frac{1}{\theta}\right)}\, dq}\, ds\right). \tag{4.6}$$

Substitution of these integrals into (4.3)–(4.4) yields a single problem in the variable θ.

4.1. An asymptotic expansion. We seek an asymptotic expansion for θ of the form

$$\theta = \theta_I + \epsilon\theta^0 + \epsilon^{3/2}\theta^1 + \cdots.$$

Then we can expand θ_I about $(t_c, 0)$ and write

$$\theta_I = 1 + \epsilon a\tau - \epsilon b\xi + o(\epsilon), \tag{4.7}$$

where

$$a = \lim_{t \to t_c} \frac{\partial \theta_I}{\partial t}, \qquad b = -\lim_{x \to 0} \frac{\partial \theta_I}{\partial x}.$$

For the continued analysis, we must assume that these limits exist and that $a > 0$ and $b > 0$. The latter condition follows from requiring that h be a nonnegative function for all times corresponding to an influx of energy at the end of the test tube. The condition $a > 0$ implies that the temperature is increasing at the onset of the transition phase. Both of these are consistent with the potential for initiation.

Substitution of (4.7) into the expansion of θ yields

$$\theta = 1 + \epsilon(a\tau - b\xi + \theta^0) + \epsilon^{3/2}\theta^1 + o(\epsilon^{3/2}),$$

so that

$$\frac{1}{\epsilon}\left(1 - \frac{1}{\theta}\right) = (a\tau - b\xi + \theta^0) + o(1).$$

Combining this result with (4.5) and (4.6) and substituting into the boundary value problem (4.3)–(4.4), we arrive at the equations governing θ^0 and θ^1:

$$O(\epsilon): \quad \begin{aligned} &\theta^0_{\xi\xi} = 0, \\ &\theta^0(-\infty,\xi) = 0, \quad \theta^0_\xi(\tau,0) = 0, \end{aligned} \tag{4.8}$$

$$O(\epsilon^{3/2}): \quad \begin{aligned} &\theta^1_{\xi\xi} = -B_0 e^{a\tau - b\xi + \theta^0} \exp\left(\tfrac{-A_0}{2}\int_{-\infty}^{\tau} e^{r(as - b\xi + \theta_0)}\, ds\right), \\ &\theta^1(-\infty,\xi) = 0, \quad \theta^1_\xi(\tau,0) = 0. \end{aligned} \tag{4.9}$$

Equation (4.8) has solution

$$\theta^0(\tau,\xi) = f_0(\tau), \quad \text{where} \quad f_0(\tau) \to 0 \text{ as } \tau \to -\infty.$$

This is substituted into (4.9) to obtain

$$\theta^1(\tau,\xi) = -B_0 e^{a\tau + f_0(\tau)} \int_0^\xi \int_0^z e^{-b\hat{z}} \exp\left(\frac{-A_0}{2} e^{-rb\hat{z}} \int_{-\infty}^{\tau} e^{r(as + f_0(s))}\, ds\right) d\hat{z}\, dz + f_1(\tau),$$

with $f_1(\tau) \to 0$ as $\tau \to -\infty$.

The function $f_0(\tau)$ governs the first order correction to the inert solution in the transition stage.

4.2. The transition stage solution. In order to determine the nature of f_0 we need a matching condition for large ξ. To this end, we consider the stretched space variable

$$X = \sqrt{\epsilon}\xi.$$

Let $\hat{\Theta}$ represent the solution in the boundary layer. From (4.3) we have

$$\hat{\Theta}_\tau = \hat{\Theta}_{XX} + O(\epsilon^{1/2}).$$

Assuming that θ has the following form in the boundary layer,

$$\theta = \theta_I + \epsilon\hat{\Theta}^0 + \epsilon^{3/2}\hat{\Theta}^1 + \cdots,$$

the $O(\epsilon)$ problem is

$$\hat{\Theta}^0_\tau = \hat{\Theta}^0_{XX}, \quad \hat{\Theta}^0 \to 0 \text{ as } \tau \to -\infty.$$

Additional conditions at $X = 0$ are needed and are determined by matching to the outer solution. Observe that as $X \to 0$ and $\xi \to \infty$,

$$\epsilon\hat{\Theta}^0 + \epsilon^{3/2}\hat{\Theta}^1 + \cdots = \epsilon\theta^0 + \epsilon^{3/2}\theta^1 + \cdots, \tag{4.10}$$

$$\begin{aligned} \epsilon\hat{\Theta}^0_X + \epsilon^{3/2}\hat{\Theta}^1_X + \cdots &= 0 + \epsilon^{3/2}\theta^1_X + \cdots \\ &= 0 + \epsilon^{3/2}(\epsilon^{-1/2}\theta^1_\xi) + \cdots. \end{aligned} \tag{4.11}$$

Equating by powers in ϵ, the above implies that

$$\lim_{X\to 0} \hat{\Theta}^0(\tau,X) = \lim_{\xi\to\infty} \theta^0(\tau,\xi), \qquad \lim_{X\to 0} \hat{\Theta}_X(\tau,X) = \lim_{\xi\to\infty} \theta^1_\xi(\tau,\xi).$$

The equation that $\hat{\Theta}^0$ satisfies is

$$\begin{aligned} \hat{\Theta}^0_\tau &= \hat{\Theta}^0_{XX}, \\ \hat{\Theta}^0_X(\tau,0) &= -B_0 e^{a\tau+f_0(\tau)} \int_0^\infty e^{-bz} e^{\frac{-A_0}{2} e^{-rbz} \int_{-\infty}^{\tau} e^{r(as+f_0(s))}\, ds}\, dz \\ &\equiv J(\tau), \\ \hat{\Theta}^0 &\to 0 \text{ as } \tau \to -\infty. \end{aligned} \tag{4.12}$$

The additional condition

$$\hat{\Theta}^0(\tau,0) = f_0(\tau)$$

determines the unknown function f_0. The solution of (4.12) can be expressed in terms of the Green's function

$$\hat{\Theta}^0(X,\tau) = -\int_{-\infty}^{\tau} J(\sigma) G(X,\tau;0,\sigma)\, d\sigma,$$

where

$$G(X,\tau;0,\sigma) = \frac{1}{\sqrt{\pi(\tau-\sigma)}} e^{\frac{-X^2}{4(\tau-\sigma)}}.$$

Finally, applying the condition on $\hat{\Theta}^0$ at $X = 0$, we arrive at the nonlinear integral equation governing the temperature in the transition stage:

$$f_0(\tau) = -\int_{-\infty}^{\tau} \frac{J(\sigma)}{\sqrt{\pi(\tau-\sigma)}}\, d\sigma = \frac{B_0}{b} \int_{-\infty}^{\tau} \frac{e^{f_0(\sigma)+a\sigma}}{\sqrt{\pi(\tau-\sigma)}} Q(\sigma)\, d\sigma, \tag{4.13}$$

where

$$Q(\sigma) = \int_0^\infty b e^{-bz} \exp\left(-e^{-rbz} \frac{A_0}{2} \int_{-\infty}^{\sigma} e^{r(f_0(s)+as)}\, ds \right) dz.$$

In the next section, we will examine the integral equation (4.13). We will perform a coordinate change resulting in the appearance of an additional parameter governing the qualitative behavior of the solution. Existence considerations will be addressed, and both analytical and numerical results presented.

5. Analysis of the integral equation. The parameter r was defined as the ratio of the decomposition activation energy to the effective activation energy obtained by applying the QSSA. Typical experimental values of the activation energy for decomposition, propagation, and termination are such that $E_d \gg E_p \gg E_t$. It follows that the ratio r is roughly 2. We will consider only values of r such that $1 < r \leq 2$, with special attention given to the case $r = 2$.

The integral Q appearing in (4.13) can be expressed in terms of gamma functions. Note that

$$\int_0^\infty b e^{-bz} \exp\left(-e^{-rbz} \frac{A_0}{2} \int_{-\infty}^{\sigma} e^{r(f_0(s)+as)}\, ds \right) dz = \frac{\Gamma\left(\frac{1}{r}\right)}{r} \gamma\left(\frac{1}{r}, q(\sigma) \right),$$

where

$$q(\sigma) = \frac{A_0}{2} \int_{-\infty}^{\sigma} e^{r(f_0(s)+as)}\, ds,$$

Γ is the gamma function, and γ is the incomplete gamma function defined by

$$\gamma(\alpha, z) = \frac{z^{-\alpha}}{\Gamma(\alpha)} \int_0^z e^{-t} t^{\alpha-1}\, dt.$$

To facilitate the analysis of the integral equation, let us introduce the change of variables

$$\eta = a\tau + \log\left(\frac{B_0}{b\sqrt{a}}\right) \quad \text{and} \quad u(\eta) = f_0(\tau).$$

In these new coordinates, (4.13) takes the form

$$u(\eta) = \int_{-\infty}^{\eta} \frac{e^{u(\sigma)+\sigma}}{\sqrt{\pi(\eta-\sigma)}} F_r\left(\lambda_r \int_{-\infty}^{\sigma} e^{r(u(s)+s)}\, ds\right) d\sigma. \tag{5.1}$$

The function F_r appearing above is defined by

$$F_r(x) = \frac{\Gamma\left(\frac{1}{r}\right)}{r}\, \gamma\left(\frac{1}{r}, x\right) \quad \text{for} \quad x > 0, \quad \text{with} \quad F_r(0) = 1,$$

and the parameter λ_r is the ratio

$$\lambda_r = a^{r/2-1} \frac{A_0 b^r}{2B_0^r} \geq 0.$$

Note that in the limiting case $r = 2$,

$$F_2(x) = \frac{\sqrt{\pi}}{2} \frac{\operatorname{erf}(\sqrt{x})}{\sqrt{x}}$$

and

$$\lambda_2 = \frac{A_0 b^2}{2B_0^2}.$$

A number of observations should be made about the parameter λ_r and the function F_r defined above. First, in the limiting case, $\lambda = 0$ ($F_r \equiv 1$), equation (5.1) reduces to the integral equation derived by Liñàn and Williams [1], Kapila [2], and Olmstead [3]. It is known that this equation has a solution u that is positive and monotonically increasing, with the asymptotic behavior

$$u \sim e^{\eta} + \frac{1}{\sqrt{2}} e^{2\eta} + \cdots \quad \text{as} \quad \eta \to -\infty,$$
$$u \sim -\frac{1}{2} \log(\eta^* - \eta) + \cdots \quad \text{as} \quad \eta \to \eta^*,$$

with $\eta^* \approx -0.431$ determined numerically. Also, for every value of r, F_r is positive monotonically decreasing, with $F_r \to 0$ as its argument tends to infinity. If $r = 1$, then (5.1) is exactly that obtained by Lasseigne and Olmstead [4] governing ignition of a solid half-space with first order Arrhenius reaction and accounting for reactant consumption. They found that there is a critical value of the parameter λ_1 such that, for values less than this critical value, the solution u becomes unbounded in finite time—it is this unbounded behavior that is taken to signal the onset of ignition.

For values of λ_1 larger than this critical value, the solution remains bounded for all finite time. It is the decaying nature of F_r that serves to inhibit initiation of a polymerization front. This is the case for all r on $1 < r \leq 2$. However, for fixed x note that $(d/dr)\, F_r(x) > 0$. Hence, as r increases, F_r decays less rapidly. As will be shown in section 5.3, $r = 2$ appears to be an upper limit for the possible existence of solutions exhibiting the type of logarithmic singularity analogous to those discussed in [3] and [4].

5.1. Existence of solutions to the integral equation. We continue the analysis by establishing the existence of solutions to (5.1). This is useful because it will establish a lower bound on the time of initiation. To that end, let us consider the class of bounded functions

$$S = \{u : (-\infty, \tilde{\eta}] \to [0, N]\},$$

where $\tilde{\eta} > -\infty$ and $0 < N < \infty$. Additionally, let the integral operator T be given by

$$Tu \longmapsto \int_{-\infty}^{\eta} \frac{e^{u(\sigma)+\sigma}}{\sqrt{\pi(\eta-\sigma)}} F_r\left(\lambda_r \int_{-\infty}^{\sigma} e^{r(u(s)+s)}\, ds\right) d\sigma \quad \text{for} \quad \eta \leq \tilde{\eta},\ u \in S.$$

Conditions on $\tilde{\eta}$ and N are sought to ensure that T is a contraction on S. First, observe that, for $u \in S$,

$$Tu \leq e^N I_0(\eta; r, \lambda_r),$$

where

$$I_0(\eta; r, \lambda_r) = \int_{-\infty}^{\eta} \frac{e^{\sigma}}{\sqrt{\pi(\eta-\sigma)}} F_r\left(\frac{\lambda_r}{r} e^{r\sigma}\right) d\sigma.$$

Second, let u_1 and u_2 be elements of S. Then

$$|Tu_1 - Tu_2| \leq \sup_{u_1, u_2 \in S} |u_1 - u_2| \left\{ e^N I_0(\eta; r, \lambda_r) + e^{(r+1)N} I_1(\eta; r, \lambda_r) \right\},$$

where

$$I_1(\eta; r, \lambda_r) = \lambda_r \int_{-\infty}^{\eta} \frac{e^{(r+1)\sigma}}{\sqrt{\pi(\eta-\sigma)}} \left| F_r'\left(\frac{\lambda_r}{r} e^{r\sigma}\right) \right| d\sigma.$$

Since I_0 and I_1 are monotonic increasing in η, we can conclude that T is a contraction on S, provided

$$I_0(\tilde{\eta}; r, \lambda_r) \leq N e^{-N} \tag{5.2}$$

and

$$e^N I_0(\tilde{\eta}; r, \lambda_r) + e^{(r+1)N} I_1(\tilde{\eta}; r, \lambda_r) < 1. \tag{5.3}$$

For given λ_r, there exists a unique pair $\hat{N} < 1,\ \hat{\eta} > -\infty$ such that (5.2) and (5.3) are satisfied as equalities. That is,

$$e^{\hat{N}} I_0(\hat{\eta}; r, \lambda_r) = \hat{N},$$
$$e^{\hat{N}} I_0(\hat{\eta}; r, \lambda_r) + e^{(r+1)\hat{N}} I_1(\hat{\eta}; r, \lambda_r) = 1.$$

Inequalities (5.2) and (5.3) are satisfied for $N = \hat{N}$ and any choice of $\tilde{\eta} < \hat{\eta}$. We note that the value of $\hat{\eta}(\lambda_r)$ provides a lower bound on the time of initiation for given λ_r. Moreover, $\hat{N}$ and $\hat{\eta}$ have the following asymptotic expansions for $\lambda_r \ll 1$ and $\lambda_r \to \infty$:

$$\hat{N} \sim 1 - \frac{\lambda_r}{(r+1)^{3/2}} + \cdots,$$
$$\hat{\eta} \sim -1 + \frac{\lambda_r}{re^r(r+1)^{3/2}} + \cdots \quad \text{as} \quad \lambda \to 0$$

and

$$\hat{N} \sim \hat{N}_\infty + \cdots,$$
$$\hat{\eta} \sim \lambda_r^{2/r}\left(\frac{\pi}{4\Gamma^2(\frac{1}{r})} r^{2-2/r}\right)\hat{N}_\infty^2 e^{-2\hat{N}_\infty} + \cdots \quad \text{as} \quad \lambda \to \infty.$$

The value $\hat{N}_\infty$ is the solution to the transcendental equation $\hat{N}_\infty = (1 - \hat{N}_\infty)e^{-r\hat{N}_\infty}$. For $1 < r \le 2$, the value of $\hat{N}_\infty$ is such that $0.33 \le \hat{N}_\infty < 0.41$. Also, $r \le 2$ and $\hat{\eta} = O(\lambda_r^{2/r})$ as $\lambda_r \to \infty$ suggests that the onset of initiation can be delayed as long as desired by taking λ_r sufficiently large.

We anticipate two qualitatively different types of solutions to (5.1), depending on the value of λ_r. Self-consistent analyses for solutions that remain bounded in finite time and those that exhibit an unbounded singularity at a finite time are sought. Such solutions are interpreted as indicating noninitiation and initiation of a front, respectively. Moreover, for a given r, there is a critical value λ_r^c separating the initiation and noninitiation regimes.

5.2. Noninitiation solutions. First, we consider the existence of solutions bounded for all finite η. To this end, assume that the solution u has the following form:

$$u \sim C\eta^d \quad \text{as} \quad \eta \to \infty, \tag{5.4}$$

where C and d are constants to be determined. If $\lambda_r > 0$ and $d < 1$, then (5.4) implies

$$e^{u+\eta}F_r\left(\lambda_r \int_{-\infty}^{\eta} e^{r(u+s)}\, ds\right) \sim \frac{\Gamma(\frac{1}{r})}{r}\left(\frac{r}{\lambda_r}\right)^{1/r} \quad \text{as} \quad \eta \to \infty.$$

For each $\eta \gg 1$, we can write

$$u(\eta) = C_0 + J(\eta),$$

where J is defined as

$$J(\eta) = \frac{1}{\sqrt{\pi}} \int_0^{\eta} \frac{e^{u+\eta}}{\sqrt{\eta - \sigma}} F_r\left(\lambda_r \int_{-\infty}^{\sigma} e^{r(u+s)}\, ds\right) d\sigma.$$

Employing the asymptotic techniques given in [7], we find that, as $\eta \to \infty$,

$$J(\eta) \sim \frac{2\Gamma(\frac{1}{r})}{r\sqrt{\pi}}\left(\frac{r}{\lambda_r}\right)^{1/r} \eta^{1/2} + \cdots.$$

Hence, u has the form given in (5.4), with the constants determined as

$$C = \frac{2\Gamma(\frac{1}{r})}{r\sqrt{\pi}}\left(\frac{r}{\lambda_r}\right)^{1/r} \quad \text{and} \quad d = \frac{1}{2}.$$

Note that $d < 1$, which is consistent with our initial requirement. If λ_r is large enough so as to advance the damping effect of F_r appearing in the integrand of (5.1), the leading order behavior of the solution is expected to be square root growth. In section 5.4, numerical confirmation of this is presented.

5.3. Initiation solutions. Next, we look for solutions of (5.1) that become unbounded at some finite time value η^*. In the case $\lambda_r = 0$, we know that the solution of (5.1) has a logarithmic singularity as previously discussed. This motivates looking for behavior of the form

$$u \sim -\beta \log(\eta^* - \eta) + \cdots \quad \text{as} \quad \eta \to \eta^*, \tag{5.5}$$

where $\beta = \beta(\lambda_r)$ and $\eta^* = \eta^*(\lambda_r) < \infty$. The analysis is facilitated by translating the singularity to the point at infinity. The techniques given in [7] and [8] can then be used. In the coordinates

$$\rho = (\eta^* - \eta)^{-1}, \qquad v(\rho) = u(\eta),$$

equation (5.1) becomes

$$v(\rho) = \sqrt{\rho}\, e^{\eta^*} \int_0^\rho \frac{e^{v - s^{-1}}}{\sqrt{\pi(\rho - s)}} s^{-3/2} F_r\left(\lambda_r e^{r\eta^*} \int_0^s t^{-2} e^{rv - rt^{-1}}\, dt\right) ds, \tag{5.6}$$

and the asymptotic behavior of v is sought as ρ tends to infinity. The cases $1 < r < 2$ and $r = 2$ must be considered separately as they give rise to different matching requirements.

Suppose $1 < r < 2$ and

$$v \sim \log(\rho^{1/2}) + \log(P) + \log(1 + o(\rho^{1/2})) \quad \text{as} \quad \rho \to \infty, \tag{5.7}$$

where P is constant. Then, as $\rho \to \infty$,

$$\frac{e^{v - 1/\rho}}{\rho^{3/2}} F_r\left(\lambda_r e^{r\eta^*} \int_0^\rho t^{-2} e^{rv - rt^{-1}}\, dt\right) \sim P\rho^{-1} F_r(\lambda_r e^{r\eta^*} I_r(\infty)) + o(\rho^{-1}),$$

where

$$I_r(\infty) = \int_0^\infty \frac{e^{rv - r/t}}{t^2}\, dt < \infty.$$

By the results in [7] and [8], it follows that

$$\int_0^\rho \frac{e^{v - 1/s}}{s^{3/2}} F_r(\lambda_r e^{r\eta^*} I_r(s)) \frac{ds}{\sqrt{\pi(\rho - s)}} \sim \frac{P}{\sqrt{\pi}} F_r(\lambda_r e^{r\eta^*} I_r(\infty)) \rho^{-1/2} \log(\rho) \tag{5.8}$$

as $\rho \to \infty$. Comparison of (5.7) and (5.8) yields

$$P = \frac{\sqrt{\pi} e^{-\eta^*}}{2F_r(\lambda_r e^{r\eta^*} I_r(\infty))}.$$

Hence,

$$v \sim \frac{1}{2} \log(\rho) + O(1) \quad \text{as} \quad \rho \to \infty,$$

and, returning to the previous coordinates, we have

$$u \sim \frac{-1}{2}\log(\eta^* - \eta) + O(1) \quad \text{as} \quad \eta \to \eta^*.$$

Different initial assumptions are needed when $r = 2$. In this case, we look for the solution of (5.6) to have the asymptotic form

$$v \sim \log(\rho^\beta) + \log(1 + o(\rho^\beta)) \quad \text{as} \quad \rho \to \infty. \tag{5.9}$$

Under the assumption (5.9), observe that the integral in the argument of F_2 appearing in (5.6) is finite only if $\beta > 1/2$. That is, matching can occur only if we restrict $\beta > 1/2$; this becomes a consistency condition on the analysis. Supposing that this is the case and that (5.9) holds, we find that

$$\frac{e^{v-\rho^{-1}}}{\rho^{3/2}} F_2\left(\lambda_2 e^{2\eta^*} \int_0^\rho \frac{e^{2v-2t^{-1}}}{t^2}\, dt\right) \sim \frac{\sqrt{\pi}}{2} e^{-\eta^*} \lambda^{-1/2} \rho^{-1} \sqrt{2\beta - 1} + o(\rho^{-1})$$

as $\rho \to \infty$. Then, employing the results in [7] and [8],

$$\rho^{-1/2} e^{-\eta^*} v(\rho) \sim \frac{\lambda_2^{-1/2}}{2} e^{-\eta^*} \sqrt{2\beta - 1}\, \rho^{-1/2} \log(\rho) \quad \text{as} \quad \rho \to \infty. \tag{5.10}$$

Comparing the left- and right-hand sides of this relation and using (5.9), we arrive at the equation for β:

$$\beta(\lambda_2) = \frac{1}{4\lambda_2}\left(1 - \sqrt{1 - 4\lambda_2}\right). \tag{5.11}$$

The following observations should be made about this result. First, note that $\beta > 1/2$, as was required for the derivation. Also, we see that this result makes sense—insofar as β is real—only for values of λ_2 between 0 and 0.25. This seems to suggest an upper bound of 0.25 on the critical value of λ_2. In fact, the numerical analysis confirms this where we find that $\lambda_2^c = 0.11998$. Finally, we note that $\beta \to 1/2$ as $\lambda_2 \to 0$, which is consistent with the results for $r < 2$ and those in [1] and [3] for the $\lambda = 0$ case. In terms of the variables u and η, the asymptotic results for the initiation case are summarized:

$$u \sim -\frac{1}{2}\log(\eta^* - \eta) + \cdots \quad \text{as} \quad \eta \to \eta^*(\lambda_r)$$

for $1 < r < 2$, and

$$u(\eta) \sim -\beta(\lambda_2)\log(\eta^*(\lambda_2) - \eta) + \cdots \quad \text{as} \quad \eta \to \eta^*(\lambda_2)$$

for $r = 2$ with β given by (5.11). In both cases, the value of η^* is to be determined numerically.

5.4. Numerical analysis. Equation (5.1) was solved numerically for several values of r and λ_r. Because the lower bound of the integral is infinite, the asymptotic form of the solution u as $\eta \to -\infty$ is useful. Using the properties of the incomplete

gamma function and the identity

$$\int_{-\infty}^{\eta} \frac{e^{\alpha\sigma}}{\sqrt{\pi(\eta-\sigma)}}\,d\sigma = \frac{e^{\alpha\eta}}{\sqrt{\alpha}} \quad \text{for all} \quad \alpha > 0,$$

we have

$$u \sim e^{\eta} + \frac{1}{\sqrt{2}}e^{2\eta} + \cdots \quad \text{as} \quad \eta \to -\infty, \tag{5.12}$$

$$\int_{-\infty}^{\sigma} e^{r(u+s)}\,ds \sim \frac{1}{r}e^{r\sigma} + \frac{r}{r+1}e^{(r+1)\sigma} + \cdots \quad \text{as} \quad \sigma \to -\infty. \tag{5.13}$$

We then fix $\eta_0 > -\infty$ and assume that for all $\eta, \sigma < \eta_0$ the relations (5.12) and (5.13) hold. Substituting (5.12) and (5.13) into (5.1), we arrive at the following equation, which is solved numerically:

$$u(\eta) = e^{\eta}\text{erfc}\sqrt{\eta-\eta_0} + \frac{1}{\sqrt{2}}e^{2\eta}\text{erfc}\sqrt{2(\eta-\eta_0)} + \int_{\eta_0}^{\eta} \frac{e^{u+\sigma}}{\sqrt{\pi(\eta-\sigma)}} F_r\left(\lambda_r I_r(\sigma)\right)\,d\sigma,$$

where

$$I_r(\sigma) = \frac{1}{r}e^{r\eta_0} + \frac{r}{r+1}e^{(r+1)\eta_0} + \int_{\eta_0}^{\sigma} e^{r(u+s)}\,ds.$$

This approach is similar to that applied by Lasseigne and Olmstead [4]. Moreover, if $r = 1$, the above reduces to the integral equation considered in [4] for a first order reaction term. The accuracy of the numerical methods employed in the current work was tested by comparing the results obtained for $r = 1$ with those in [4]. The value $\eta_0 = -10$ was found to be sufficient to produce reliable results, and this was used for all numerical trials given in this paper.

6. Results and discussion. For convenience, we restate the definition of the parameter λ_r here,

$$\lambda_r = a^{r/2-1}\frac{A_0 b^r}{2B_0^r},$$

and recall that A_0 and B_0 are measures of the consumption rate of initiator and heat release due to conversion of monomer, respectively; r $(1 < r \leq 2)$ is the ratio of activation energies associated with decomposition of initiator and polymer chain propagation. We see that λ_r is small, provided that A_0 is relatively small and B_0 relatively large. Hence, we can consider large values of λ_r to indicate an inadequate amount of initiator (i.e., initiator is consumed too rapidly) or that heat release is insufficient to sustain further reaction. Conversely, small values of λ_r represent a sufficiently exothermic reaction, in which the consumption rate of initiator is small relative to the amount of initiator present in the mixture. Small λ_r values are therefore expected to lead to initiation, while large values of λ_r are not. The appearance of a and b in the ratio is the effect of the inert heating, and the values of these parameters are controlled by the choice of heat source applied. As suggested by the results in [4] and the self-consistent analyses in sections 5.2 and 5.3 of this paper, there exists a critical value of λ_r separating the initiation and noninitiation regimes.

TABLE 6.1
The critical parameter value, λ_r^c, as a function of r.

r	1.5	1.8	1.9	2
λ_r^c	0.6645	0.31086	0.21058	0.11998

TABLE 6.2
Initiation time η^* for selected values of λ_2.

λ_2	0	0.01	0.1	0.117	0.11997
η^*	-0.4310	-0.4287	-0.4088	-0.4048	-0.4037

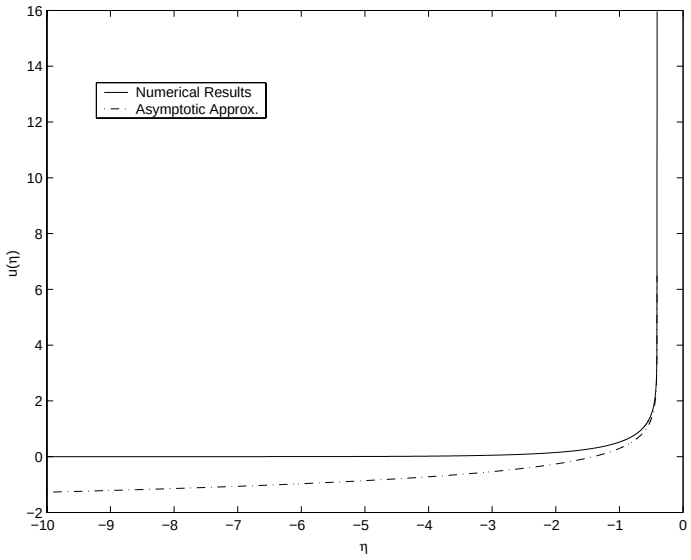


FIG. 6.1. *Initiation solution of the integral equation for $r = 2$ and $\lambda_2 = 0.1$ The solution approaches the asymptotic approximation $U = -\beta(0.1)\log(\eta^* - \eta)$ as $\eta \to \eta^* \approx -0.4088$.*

The critical parameter value, λ_r^c, was determined numerically for different r values. The results are given in Table 6.1. If $\lambda_r < \lambda_r^c$, then the solution exhibits a logarithmic singularity, with the asymptotic behavior described in section 5.3. For values of λ_r larger than λ_r^c, the solution to (5.1) exists and is finite for all η. When λ_r is only slightly larger than the critical value, the solution exhibits behavior on two time scales (see Figure 6.3). The temperature grows slowly while oscillating on a short time scale. This results from the competing effects of the exponential term appearing in (5.1) and the decaying function F_r. If λ_r is increased further, the solution has the leading form described in section 5.2.

The time at which initiation occurs for the case $r = 2$ is given in Table 6.2 for various λ_2, with the critical value found to be 0.11998. Solutions of the types described above for $r = 2$ are shown in Figures 6.1–6.4. In Figure 6.1, λ_2 is less than the critical value. The solution becomes unbounded at $\eta = -0.4088$. The asymptotic results are shown as a dashed curve for comparison. Similarly, Figure 6.2 shows the initiation solution for $\lambda_2 = 0.11997$, just slightly less than the critical value. In both cases, the singular behavior indicates that the temperature progresses beyond the transition stage, and a polymerization front is formed. In contrast, Figures 6.3 and 6.4 show the solution when λ_2 is above the critical value. In Figure 6.3, $\lambda_2 = 0.4$ and the

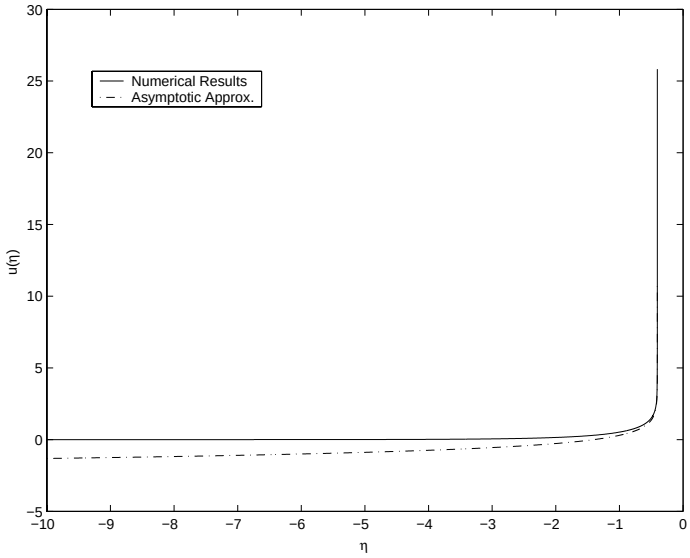


FIG. 6.2. *Initiation solution of the integral equation for* $r = 2$ *and* $\lambda_2 = 0.11997$*, just below the critical value of* 0.11998*. The solution approaches the asymptotic approximation* $U = -\beta(0.11997)\log(\eta^* - \eta)$ *as* $\eta \to \eta^* \approx -0.4037$*.*

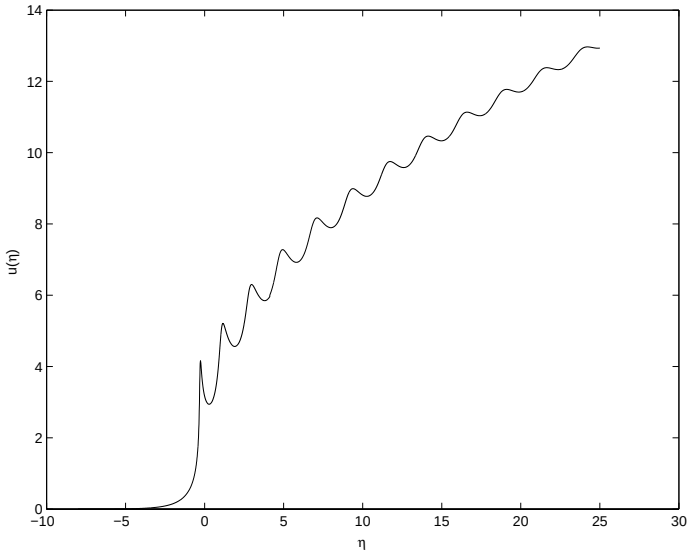


FIG. 6.3. *The noninitiation solution showing oscillation for* $r = 2$ *and* $\lambda_2 = 0.4$*.*

temperature oscillations described can be seen. However, the large scale behavior is slow growth with the oscillations damping as η increases. Figure 6.4 is a plot of the solution when $\lambda_2 = 1$. Here, the solution is monotonic with a change of concavity occurring in a neighborhood of $\eta = 0$. The temperature remains bounded, indicating that a reaction front does not form.

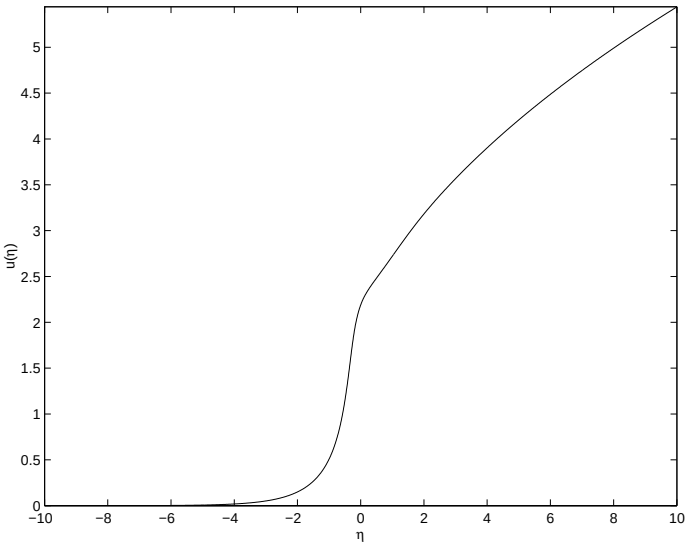


FIG. 6.4. *The noninitiation solution for* $r = 2$ *and* $\lambda_2 = 1$.

7. Summary and conclusion. A reduced system governing a five species reaction model of free radical frontal polymerization was considered, and the temperature was tracked from the inert heating to the transition stage. Through an asymptotic analysis, the integral equation (4.13) arose as the first correction to the inert solution. This integral equation was then rewritten by a change of variables as (5.1), where there appears the parameter λ_r which governs the qualitative behavior of the solution. For a fixed ratio of the activation energies, there is a critical value of the parameter λ_r^c such that the solution of (5.1) has an infinite singularity in finite time if $\lambda_r < \lambda_r^c$ but remains bounded for all time if $\lambda_r > \lambda_r^c$.

The unbounded and bounded types of solutions are taken to indicate initiation and noninitiation in the underlying system, initiation being the formation and onset of propagation of a polymerization front. In the noninitiation case, but for values of λ_r close to the critical value, an oscillatory type of solution was found numerically. The solution remains bounded in this case, and it appears that the oscillations dampen with the growth of the independent variable.

The experimental parameters can be chosen so as to ensure the onset of a thermal front. The critical temperature T_c used in the scaling can be determined by taking $A_0 = 1$ in the relation (3.5). This results in a transcendental equation for T_c,

$$\frac{E_{eff}}{R_g T_c} = \frac{k_d(T_c)}{k_{eff}(T_c)\sqrt{I_0}}.$$

Then, B_0 can be found in terms of the initial amount of monomer and the heat release parameter q, and the values a and b are given in terms of the known flux condition.

Some additional comments regarding the relationship of the integral equation (5.1) to the original system (3.1)–(3.4) and the limitations of the results are in order. First, we have shown that, under certain conditions, the solution to the integral equation (5.1) exhibits an infinite singularity at a finite time. This singular behavior is interpreted as thermal runaway and hence initiation of a polymerization front. This

does not, however, correspond to blow-up of the solution of the original system (3.1)–(3.4) of interest in this study. For the system (3.1)–(3.4), there exists a unique, global solution as indicated by the classical theory of parabolic equations. However, the Arrhenius reaction term produces a large temperature gradient at the site of initiation so that the temperature profile at the end of the tube exhibits a steep increase to the maximum temperature in a thin reaction zone. It is this sharp increase in temperature that is modeled asymptotically by the thermal runaway phenomenon of the integral equation (5.1).

Second, we note that even in the case when thermal runaway occurs in (5.1)—i.e., when the parameter values are such that $\lambda_r < \lambda_r^c$—the front formed requires a sufficiently large amount of initiator present in the mixture for propagation throughout the tube. While it is possible to induce runaway by imposing a sufficiently high level of external energy input at $x = 0$, this case is not of interest since the reaction will die off and the polymer will not be produced. Hence, for the results obtained in this paper to be of practical use, the values of a and b must be assumed to be $O(1)$ and fixed as prescribed by the externally imposed heat flux. Then, the variation in the magnitude of λ_r can be viewed as due to changes in the values of A_0 and B_0, which correspond to the physical and chemical properties of any particular mixture.

REFERENCES

[1] A. Liñàn and F. A. Williams, *Theory of ignition of a reactive solid by constant energy flux*, Combustion Sci. Tech., 3 (1971), pp. 91–98.

[2] A. K. Kapila, *Evolution and deflagration in a cold combustible subjected to a uniform energy flux*, Internat. J. Engrg. Sci., 19 (1981), pp. 495–509.

[3] W. E. Olmstead, *Ignition of a combustible half space*, SIAM J. Appl. Math., 43 (1983), pp. 1–15.

[4] D. Glenn Lasseigne and W. E. Olmstead, *Ignition of a combustible solid with reactant consumption*, SIAM J. Appl. Math., 47 (1987), pp. 332–342.

[5] P. M. Goldfeder, V. A. Volpert, V. M.Ilyashenko, A. M.Khan, J. A. Pojman, and S. E. Solovyov, *Mathematical modeling of free-radical polymerization fronts*, J. Phys. Chem. B, 101 (1997), pp. 3474–3482.

[6] C. A. Spade and V. A. Volpert, *On the steady-state approximation in thermal free radical frontal polymerization*, Chem. Engrg. Sci., 55 (2002) pp. 641–654.

[7] R. A. Handelsman and W. E. Olmstead, *Asymptotic solution to a class of nonlinear Volterra integral equations*, SIAM J. Appl. Math., 22 (1972), pp. 373–384.

[8] W. E. Olmstead and R. A. Handelsman, *Asymptotic solution to a class of nonlinear Volterra integral equations* II, SIAM J. Appl. Math., 30 (1976), pp. 180–189.

[9] G. B. Manelis, L. P. Smirnov, and N. I. Peregudov, *Nonisothermal kinetics of polymerization processes. Finite cylindrical reactor*, Combustion Explosion Shock Waves, 13 (1977), pp. 389–383.

[10] J. Brandrup and E. H. Immergut, *Polymer Handbook*, Wiley, New York, 1989.

SIAM J. APPL. MATH.
Vol. 63, No. 6, pp. 1849–1888

EFFECTIVE EQUATIONS FOR SOUND AND VOID WAVE PROPAGATION IN BUBBLY FLUIDS*

NIANQING WANG† AND PETER SMEREKA‡

Abstract. Effective equations that describe both sound wave and void wave propagation for bubbly flows at high Reynolds numbers are derived in this paper. First ideal bubble flows are considered, and a new method for solving Laplace's equation for the velocity potential is presented. This approach is based on a generalization of the method of images and also yields a precise definition of the ambient field experienced by a bubble. With the velocity potential known, the Lagrangian is then computed, and equations of motion for a finite number of bubbles using the Euler–Lagrange equations are derived. The continuum limit is then used to obtain our effective equations. Our expressions for the sound wave and void wave speeds agree well with previous investigations. The effects of gravity and viscosity on void waves are considered. Viscous effects are incorporated using a dissipation function. The steady rise speed and void wave speed for a column of rising bubbles are computed and found to agree well with experiments.

Key words. bubbly flow, potential flow, void waves

AMS subject classifications. 76T10, 76B07

DOI. 10.1137/S0036139902413052

1. Introduction.

In this paper, we derive effective equations for sound and void wave propagation in bubbly fluids. Sound propagation was studied by Carstensen and Foldy [11], who derived the speed of sound using a linear scattering theory developed by Foldy [13]. Iordanskii [21] and van Wijngaarden [48] derived effective equations including nonlinear effects. For review of the literature on acoustic waves in bubbly liquids the reader is referred to the article by van Wijngaarden [49]. Later, Caflisch et al. [9] provided an alternate derivation that clarified the range of validity of the effective equations derived by Iordanskii and van Wijngaarden. These equations are valid when the volume fraction of bubbles is very small. This is, in part, because in these investigations it was assumed that bubbles would undergo only radial motion. In order to develop equations valid at higher volume fraction, one must include the effects of bubble translation. This has been investigated by Crespo [12], Noordzij and van Wijngaarden [32], Caflisch et al. [10], and Sangani [38], among others. Crespo used volume averaged equations of motion, which are valid for low frequency perturbations. He then linearized these equations of motion and computed the speed of sound waves, the results of which were found to be in good agreement with experiments. Caflisch et al. [10] linearized the equations of motion and used a multiple-scale method. Their computation of the wave speed is valid only for small frequencies and was in agreement with the results of Crespo. Sangani also linearized the equations of motion and then performed ensemble averaging. He also computed the wave speed, and his expression was valid over a wide range of wave frequencies. Sangani's results compare well to the experimental results of Silberman [40]. In this work we shall derive equations of motion that are fully nonlinear and valid over a wide range of frequencies. Our

*Received by the editors August 13, 2002; accepted for publication (in revised form) February 16, 2003; published electronically August 15, 2003. This work was supported, in part, by the National Science Foundation through a Career Award.
http://www.siam.org/journals/siap/63-6/41305.html
†Epic Systems Corporation, 5301 Tokay Blvd., Madison, WI 53711 (nianqing@epicsystems.com).
‡Department of Mathematics, University of Michigan, Ann Arbor, MI 48109 (psmereka@umich.edu).

calculation of the sound speed agrees with those of previous investigators.

There has been considerable work on void wave propagation for ideal bubbly flows where the bubbles are considered to be rigid spheres surrounded by an incompressible, inviscid, irrotational fluid. For example, Biesheuvel and van Wijngaarden [8], Geurst [17, 18], Biesheuvel and Gorissen [6], Wallis [47], Pauchon and Smereka [34], Zhang and Prosperetti [56], and Park, Drew, and Lahey [33], among others, have derived effective equations using various types of averaging. The motion of individual bubbles was studied numerically by Sangani and Didwania [39] and Smereka [42]. Smereka used the point bubble approximation combined with Euler–Lagrange equations to obtain explicit equations of motion which were integrated numerically. Sangani and Didwania used a multipole method together with Newton's law to simulate bubble motion. Both investigations observed that in many situations a spatially uniform mixture of bubbles moving with approximately the same velocity was unstable and the bubbles would form clusters. In should also be mentioned that van Wijngaarden [53] predicted that bubbly flows would have a tendency to cluster. We speculate that the instability of the spatially uniform mixture of bubbles is consistent with the ill-posed effective equations found by Geurst [17], Wallis [47], and Pauchon and Smereka [34] for dilute bubbly flows.

It was also found by Sangani and Didwania [39] and Smereka [42] that, if gravity and viscosity were included, the clustering was much more pronounced. Smereka constructed a Lyapunov function and showed that bubbles have a strong tendency to maximize their virtual mass in the direction of motion. This means that the bubbles will form pancake-shaped clusters. The dynamics of clustering has been studied in more detail by Yurkovetsky and Brady [55] and Galper and Miloh [16]. Many aspects of bubbly flows with potential flow interaction have also been discussed in the review article by Koch and Hill [24].

In an effort to understand the numerical simulations of Sangani and Didwania and Smereka, Russo and Smereka [37] wrote the equations of motion for individual bubbles in Hamiltonian form (using the point bubble approximation) and then deduced a kinetic equation for the probability density of the bubbles in phase space. They proved that the spatially uniform case was unstable, provided that the variance of the bubble's velocity was sufficiently small. On the other hand, they proved that, if the variance of the bubble's velocity was sufficiently large, the spatially uniform bubbly fluid was stable. Similar results were obtained by Spelt and Sangani [45]. Herrero, Lucquin-Desreux, and Perthame [20] were able to provide a more rigorous derivation of the equations derived by Russo and Smereka and remove an important assumption.

The effects of liquid viscosity have also been considered by van Wijngaarden and Kapteyn [52] and van Wijngaarden [53]. Van Wijngaarden and Kaptyen computed the drag force on a pair of bubbles using an energy dissipation argument. They then computed averaged equations using ensemble averages over pairs of bubbles. The results were used to compute the profile of a wave of steady shape. Van Wijngaarden used the results of van Wijngaarden and Kapteyn to compute the rise speed of a mixture of bubbles. The result is in good agreement with experimental data.

More recently, Lammers and Biesheuvel [26] studied void waves in bubbly flows using a theory of Batchelor [4]. They find that the speed c of void waves is given by

$$c = U_0(\beta) + \beta U_0'(\beta), \tag{1}$$

where U_0 is the rise speed of the bubbles and β is the void fraction. They measure both c and U_0 and find that they agree well with (1). In the current work we also

obtain (1) by a different approach. In addition, we calculate U_0 and find that it agrees closely with the experimental results of Lammers and Biesheuvel.

2. Outline. We shall derive effective equations by first computing the equations of motion of N bubbles surrounded by an ideal liquid of infinite extent. We assume that the bubbles are spherical but may change their radius. To fix ideas let us first consider a single spherical gas bubble in a liquid which is at rest at infinity. The equations of motion are well known; they are

$$R\ddot{R} + \frac{3}{2}\dot{R}^2 - \frac{1}{4}|\mathbf{U}|^2 + \frac{p_\infty - P}{\rho_\ell} = 0, \tag{2}$$

$$\frac{1}{3}\dot{\mathbf{U}} + \frac{\dot{R}}{R}\mathbf{U} = 0, \tag{3}$$

where $R(t)$ is the bubble radius, $\mathbf{U}(t)$ is the bubble velocity, ρ_ℓ is the density of the liquid, p_∞ is the pressure at infinity, and P is the pressure inside the bubble. Surface tension is neglected for the purpose of simplicity. Equation (2) with $\mathbf{U} = 0$ can be found in Lamb [25], for example. The inclusion of (3) can be found in Hermans [19], for example.

Next we consider the situation in which the surrounding liquid is uniformly accelerated. The equations of motion are

$$R\ddot{R} + \frac{3}{2}\dot{R}^2 - \frac{1}{4}|\mathbf{U} - \mathbf{v}|^2 + \frac{p_\infty - p_g}{\rho_\ell} = 0, \tag{4}$$

$$\frac{1}{3}\dot{\mathbf{U}} - \dot{\mathbf{v}} + \frac{\dot{R}}{R}(\mathbf{U} - \mathbf{v}) = 0, \tag{5}$$

where $\mathbf{v}$ is the velocity of the liquid at infinity. The equations of motion in this case are derived by first obtaining the pressure at the bubble surface from Bernoulli's law and then demanding that the average pressure on the surface be equal to the pressure inside the bubble and that total force on the bubble be zero. Equation (5) can be found in Batchelor [3, p. 455].

Let us consider the situation with N bubbles surrounded by a fluid of infinite extent initially at rest. The bubbles are then set into motion. It is plausible to think that each bubble moves only according to certain "ambient" fields. Therefore we write the equations of motion, heuristically, for the kth bubble as

$$R_k\ddot{R}_k + \frac{3}{2}\dot{R}^2 - \frac{1}{4}|\mathbf{U}_k - \mathbf{v}_A(k)|^2 + \frac{p_A(k) - P_k}{\rho_\ell} = 0, \tag{6}$$

$$\frac{1}{3}\dot{\mathbf{U}}_k - \frac{D_?}{Dt}\mathbf{v}_A(k) + \frac{\dot{R}_k}{R_k}(\mathbf{U}_k - \mathbf{v}_A(k)) = 0, \tag{7}$$

where $\mathbf{v}_A(k)$ and $p_A(k)$ are the ambient liquid velocity and the ambient pressure of the kth bubble, which must be determined. $\frac{D_?}{Dt}$ denotes a material derivative associated to a velocity field yet to be determined. One of the key results of this paper is a systematic way to determine these ambient fields.

2.1. Summary and approach. Our approach is as follows. First consider a finite number of bubbles in an infinite expanse of fluid. We assume that the fluid motion is irrotational, inviscid, incompressible, and at rest at infinity. We also assume that the bubbles are spherical. We then derive equations of motion for this finite collection of bubbles using Lagrange's variational principal as outlined in Lamb [25]

or Milne-Thompson [30], for example. This requires one to compute the velocity potential. We develop a new method to solve for the velocity potential. The method is an extension of the method of images for two bubbles (e.g., Lamb [25]) to multiple bubbles. We prove that this method is convergent.

With the velocity potential known, we can calculate the Lagrangian for a finite number of bubbles. In principle we could calculate the exact equations of motion; however, they would be extremely complex. Instead we truncate our equations of motion and include only terms involving monopoles and dipoles. In addition, we are able to systematically deduce the ambient fields. We then take the continuum limit of our discrete equations of motion and find the following effective equations:

$$R\frac{d^2R}{dt^2} + \frac{3}{2}\left(\frac{dR}{dt}\right)^2 - \frac{1}{4}|\mathbf{U}-\mathbf{v}|^2 + \frac{p-p_g(R)}{\rho_\ell} = 0, \tag{8}$$

$$\frac{1}{3}\frac{d\mathbf{U}}{dt} - \frac{D\mathbf{v}}{Dt} + \frac{1}{R}\frac{dR}{dt}(\mathbf{U}-\mathbf{v}) + (\mathbf{U}-\mathbf{v})\times(\nabla\times\mathbf{v}) = 0, \tag{9}$$

where $\frac{d}{dt} = \partial_t + \mathbf{U}\cdot\nabla$ and $\frac{D}{Dt} = \partial_t + \mathbf{v}\cdot\nabla$. The dependent variables are now functions of space and time (e.g., $R = R(\mathbf{x}, t)$). The ambient pressure p and ambient liquid velocity $\mathbf{v}$ are related as follows:

$$\mathbf{v} = \nabla\psi \quad \text{and} \quad p = p_\infty - \frac{\partial\psi}{\partial t} - \frac{1}{2}|\mathbf{v}|^2,$$

where p_∞ is the pressure at infinity and ψ is the ambient velocity potential. An explicit expression for ψ is given by (42). We also establish that, to leading order, $\mathbf{v}$ is the volume averaged liquid velocity.

We also consider the effects of gravity and liquid viscosity. We include the viscous effects by using the energy dissipation method. Our approach is similar to that of van Wijngaarden and Kapteyn [52] and van Wijngaarden [53] except that we do not use the assumption of pairwise interaction. When we pass to the continuum limit we find that, to leading order, the drag force on a bubble, in the case of zero volume flux, is

$$12\pi\mu R(\mathbf{U} - 2\mathbf{v} - \mathbf{w}).$$

The expression for $\mathbf{w}$ is given by (79) in section 5. The above formula in one space dimension can be written as

$$12\pi\mu R(1 + \beta + \beta^2)(U - v).$$

This formula is also valid for all cases when the volume flux is constant in time.

We include this formula in our model along with effects of gravity to study the propagation of void waves in bubbly flow. Work of Sangani and Didwania [39] and Smereka [42] suggest that the potential flow approximation cannot be used for void wave propagation in bubbly flows since it predicts strong clustering of the bubbles, which is something not observed in experiments. Recent experiments of Zenit, Koch, and Sangani [58] show that there is some clustering but not to the extent predicted in [39, 42].

We show that our model has a steady solution which corresponds to a spatially homogeneous bubble mixture; the computed velocity is in good agreement with experiments. Furthermore, we demonstrate that this steady solution is unstable, which is in agreement with [39, 42]. We then argue that for naturally occurring perturbations the instability is rather weak. The propagation of these perturbations corresponds to void waves. We calculate the speed of these waves and find that our calculations agree closely with experimental results.

3. Equations of motion. We neglect liquid viscosity and gravity in this chapter. The total energy, the sum of the kinetic energy of the liquid and potential energy stored in the bubbles, is conserved. The Lagrangian is calculated, and the Euler–Lagrange equations give the equations of motion. The velocity potential, given as a convergent series, and its combinatorial properties play an important role in the derivation.

3.1. Velocity potential. For a flow with N disjoint spherical bubbles, we want to find the velocity potential satisfying

$$\triangle\phi = 0 \quad \text{outside the bubbles,} \tag{10}$$

$$\frac{\partial\phi}{\partial n} = \mathbf{U}_i \cdot \mathbf{n} + \dot{R}_i \quad \text{on the surface of bubble } i,\ i = 1, \dots, N, \tag{11}$$

$$\nabla\phi = 0 \quad \text{at infinity,} \tag{12}$$

where $\dot{R}_i$, $\mathbf{U}_i$ are radial and translational velocities of bubble i, $\mathbf{n}$ is the unit normal vector pointing toward the liquid phase on the surface, and

$$\frac{\partial\phi}{\partial n} = \mathbf{n} \cdot \nabla\phi$$

is the directional derivative along $\mathbf{n}$. We have the following result concerning the solution of (10)–(12).

THEOREM 3.1. *Let*

$$\phi_i(\mathbf{r}) = -\frac{R_i^2 \dot{R}_i}{|\mathbf{r} - \mathbf{x}_i|} + \frac{1}{2} R_i^3 \nabla_r \left(\frac{1}{|\mathbf{r} - \mathbf{x}_i|} \right) \cdot \mathbf{U}_i. \tag{13}$$

Here ϕ_i is the solution if only the ith bubble is present. The solution of (10)–(12) *can be written as the uniformly convergent series in the liquid,*

$$\phi = \sum_{i_1=1}^{N} \phi_{i_1} + \sum_{\substack{i_1, i_2 = 1 \\ i_1 \neq i_2}}^{N} I_{i_1} \phi_{i_2} + \cdots + \sum_{\substack{i_1, \dots, i_k = 1 \\ i_\ell \neq i_{\ell+1}}}^{N} I_{i_1} I_{i_2} \cdots I_{i_{k-1}} \phi_{i_k} + \cdots, \tag{14}$$

where I_i refers to the image potential operator with respect to bubble i. I_i is defined as follows: let $f(\mathbf{x})$ be a harmonic function inside the ith bubble and let $g(\mathbf{x})$ be a harmonic function outside the ith bubble such that

$$\frac{\partial f}{\partial n} = -\frac{\partial g}{\partial n} \tag{15}$$

on the surface of the ith bubble; then g is called the image potential of f with respect to bubble i with the notation $g = I_i f$.

We note that this is a generalization of the method of images used to solve the motion of two spheres as outlined in Lamb [25].

Next we define the ambient velocity potential experienced by the jth bubble to be

$$\psi_j = \sum_{i_1 \neq j}^{N} \phi_{i_1} + \sum_{\substack{i_1, i_2 = 1 \\ i_1 \neq j, i_1 \neq i_2}}^{N} I_{i_1} \phi_{i_2} + \cdots + \sum_{\substack{i_1, \dots, i_k = 1 \\ i_1 \neq j, i_\ell \neq i_{\ell+1}}}^{N} I_{i_1} I_{i_2} \cdots I_{i_{k-1}} \phi_{i_k} + \cdots. \tag{16}$$

The ambient liquid velocity experienced by bubble j is defined as $\mathbf{v}_j = \nabla\psi_j$. In the expression for ψ_j we see that the final image reflection of each term is not with respect to bubble j. This means that ψ_j is harmonic inside bubble j and $I_j\psi_j$ is well defined.

It is easy to derive two useful expressions,

$$\phi = \phi_j + \psi_j + I_j\psi_j \tag{17}$$

and

$$\psi_j = \sum_{j\neq k} \left(\phi_k + I_k\psi_k\right). \tag{18}$$

Most of the proof of Theorem 3.1 will be in Appendix A. Here we only outline the proof of convergence and show that ϕ satisfies the boundary condition (11). To prove the latter, we first notice that

$$\frac{\partial\phi_i}{\partial n} = \mathbf{U}_i \cdot \mathbf{n} + \dot{R}_i$$

on the surface of bubble i. Thus ϕ_i is the exact solution of (10)–(12) for one bubble. In the case of multiple bubbles, it follows from (17) that

$$\frac{\partial\phi}{\partial n} = \frac{\partial\phi_i}{\partial n} + \frac{\partial\psi_i}{\partial n} + \frac{\partial I_i\psi_i}{\partial n}.$$

From the definition of the operator I, we have

$$\frac{\partial\psi_i}{\partial n} = -\frac{\partial I_i\psi_i}{\partial n}$$

at the surface of the bubble i. Hence we find

$$\frac{\partial\phi}{\partial n} = \frac{\partial\phi_i}{\partial n} = \mathbf{U}_i \cdot \mathbf{n} + \dot{R}_i.$$

To prove convergence we first separate the series into N subseries (one for each bubble), each of which is harmonic in the region exterior to the corresponding bubble. We then prove that this series converges in the energy norm. This allows us to prove that the velocity potential converges on the surface of the corresponding bubble. This enables us to find the velocity potential on the bubble surface. The Poisson kernel for the exterior Dirichlet problem is used to prove that the series converges at each point in the liquid to a solution of Laplace's equation. The details can be found in Appendix A.

3.1.1. Example. To understand some of the preceding formulas more easily it is useful to write them explicitly for the case $N = 2$. We begin with

$$\phi = \phi_1 + \phi_2 + I_1\phi_2 + I_2\phi_1 + I_2I_1\phi_2 + I_1I_2\phi_1 + I_2I_1I_2\phi_1 + I_1I_2I_1\phi_2 + \cdots.$$

The ambient velocity potential for the first bubble is

$$\psi_1 = \phi_2 + I_2\phi_1 + I_2I_1\phi_2 + I_2I_1I_2\phi_1 + \cdots.$$

3.1.2. The image operator. The following theorem provides an explicit formula for the image potential with respect to a bubble centered at $\mathbf{x} = \mathbf{p}$ with radius R (denoted B).

THEOREM 3.2. *If $f(\mathbf{x})$ is harmonic inside a bubble centered at $\mathbf{p}$ with radius R, then*

$$I_B f(\mathbf{x}) = \sum_{n=1}^{\infty} \frac{(-1)^n R^{2n+1} \nabla^n f(\mathbf{p}) \cdot \nabla_{\mathbf{x}}^n (\frac{1}{|\mathbf{x}-\mathbf{p}|})}{(n-1)!(n+1)(2n-1)!!}$$
$$= -\frac{1}{2} R^3 \nabla f(\mathbf{p}) \cdot \nabla_{\mathbf{x}} \left(\frac{1}{|\mathbf{x}-\mathbf{p}|} \right) + \textit{(higher order harmonics)}.$$

The proof is given in Appendix B.

This theorem expresses the image potential as an expansion of spherical harmonics centered at $\mathbf{p}$. With this theorem, (14) can be written as a multipole series, and it becomes a natural extension of the twin spherical expansion method, which is commonly used when solving Laplace's equation outside two spheres (e.g., Ross [36] or Jeffrey [22]). ∇^n is the nth order matrix of partial derivatives, i.e.,

$$\nabla^n = \frac{\partial^n}{\partial x_{k_1} \partial x_{k_2} \cdots \partial x_{k_n}},$$

where $k_j = 1, 2$, or 3, with $j = 1, \ldots, n$. $\nabla^n f \cdot \nabla^n g$ denotes the scalar product of two nth order matrices.

3.2. Kinetic energy and potential energy. The kinetic energy of the liquid is

$$K = \frac{1}{2}\rho_\ell \int_{V_\ell} |\nabla\phi|^2 dv = \frac{1}{2}\rho_\ell \int_{V_\ell} \nabla \cdot (\phi \nabla\phi) dv = -\frac{1}{2}\rho_\ell \sum_{j=1}^{N} \int_{S_j} \phi \frac{\partial \phi}{\partial n} ds, \tag{19}$$

where ρ_ℓ is the density of the liquid, V_ℓ is the volume occupied by the liquid, and S_i is the surface of the ith bubble. From (17), we have

$$K = -\frac{1}{2}\rho_\ell \sum_{j=1}^{N} \int_{S_j} (\phi_j + \psi_j + I_j\psi_j)(\dot{R}_j + \mathbf{U}_j \cdot n) ds.$$

Substituting (13) into the above expression, we obtain

$$K = -\frac{1}{2}\rho_\ell \sum_{j=1}^{N} \int_{S_j} \left(-R_j \dot{R}_j - \frac{1}{2} R_j \mathbf{U}_j \cdot \mathbf{n} + \psi_j + I_j\psi_j \right) \cdot (R_j + \mathbf{U}_j \cdot \mathbf{n}) ds.$$

Next, we use the expressions in Appendix C, and the kinetic energy of the liquid becomes

$$K = 2\pi\rho_\ell \sum_{i=1}^{N} \left(R_i^3 \dot{R}^2 + \frac{1}{6} R_i^3 \mathbf{U}_i^2 - R_i^2 \dot{R}_i \psi_i(\mathbf{x}_i) - \frac{1}{2} R_i^3 \mathbf{U}_i \cdot \mathbf{v}_i(\mathbf{x}_i) \right). \tag{20}$$

The first two terms in the parentheses correspond to the energy generated by the motion of each individual bubble, as if no other bubbles exist. The last two terms are caused by the interactions between the bubbles through the ambient fields ψ_i and $\mathbf{v}_i$.

We assume that there is no heat transfer involved in the radial oscillation of the bubbles and that the adiabatic constant for the gas is γ. Therefore the pressure inside the bubble is

$$p_g(R) = p_\infty \left(\frac{\rho_0}{\rho}\right)^\gamma = p_\infty \left(\frac{R_0}{R}\right)^{3\gamma}.$$

Therefore the potential energy for a single bubble is

$$\begin{aligned} -\int_{R_0}^{R} 4\pi R^2 (p_g(R) - p_\infty) dR &= -\int_{R_0}^{R} 4\pi R^2 p_\infty \left(\left(\frac{R_0}{R}\right)^{3\gamma} - 1\right) dR \\ &= 4\pi p_\infty \left(\frac{R_0^{3\gamma} R^{-3\gamma+3}}{3\gamma - 3} + \frac{1}{3}R^3 - \frac{R_0^3}{3\gamma-3} - \frac{1}{3}R_0^3\right). \end{aligned}$$

The total potential energy of the bubbles is

$$\mathcal{U}_g = 4\pi p_\infty \sum_{i=1}^{N} \left(\frac{R_0^{3\gamma} R_i^{-3\gamma+3}}{3\gamma - 3} + \frac{1}{3}R_i^3 - \frac{R_0^3}{3\gamma-3} - \frac{1}{3}R_0^3\right). \tag{21}$$

We remark that the assumption that the bubbles behave adiabatically can be removed by using the approach outlined by Smereka [44].

3.3. Euler–Lagrange equations. With the kinetic and potential energy obtained in the last two sections, we can write the Lagrangian of the system as

$$\mathcal{L} = \mathcal{K} - \mathcal{U}_g.$$

The Euler–Lagrange equations are

$$\begin{aligned} \frac{d}{dt}\frac{\partial \mathcal{L}}{\partial \dot{R}_i} - \frac{\partial \mathcal{L}}{\partial R_i} &= 0, \\ \frac{d}{dt}\frac{\partial \mathcal{L}}{\partial \mathbf{U}_i} - \frac{\partial \mathcal{L}}{\partial \mathbf{x}_i} &= 0. \end{aligned}$$

To simplify notation, $\psi_i, \mathbf{v}_i, \nabla\psi_i$, etc. will be used to refer to their function values at $\mathbf{x}_i$ in the subsequent equations of this section. The following formulas are derived in Appendix D:

$$\frac{\partial K}{\partial \dot{R}_i} = 2\pi\rho_\ell (2R_i^3\dot{R} - 2R_i^2\psi_i), \tag{22}$$

$$\frac{\partial K}{\partial \mathbf{U}_i} = 2\pi\rho_\ell \left(\frac{1}{3}R_i^3\mathbf{U}_i - R_i^3\mathbf{v}_i\right), \tag{23}$$

$$\frac{\partial K}{\partial R_i} = 2\pi\rho_\ell \left(3R_i^2\dot{R}_i^2 + \frac{1}{2}R_i^2|\mathbf{U}_i|^2 - 4R_i\dot{R}_i\psi_i - 3R_i^3\mathbf{U}_i\cdot\mathbf{v}_i + \frac{3}{2}R_i^2|\mathbf{v}_i|^2 + F\right), \tag{24}$$

$$\frac{\partial K}{\partial \mathbf{x}_i} = 2\pi\rho_\ell (-2R_i^2\dot{R}_i\mathbf{v}_i + R_i^3 (\nabla\mathbf{v}_i)^T \cdot (\mathbf{v}_i - \mathbf{U}_i) + G), \tag{25}$$

where F and G involve only $\nabla^2\psi_i, \nabla^3\psi_i, \ldots$, which correspond to spherical harmonics of higher order than a dipole. From the above formulas, we obtain the equations of motion

$$\frac{d}{dt}\left(2R_i^3\dot{R}_i - 2R_i^2\psi_i\right) - \left(3R_i^2\dot{R}_i^2 + \frac{1}{2}R_i^2|\mathbf{U}_i|^2 - 4R_i\dot{R}_i\psi_i - 3R_i^3\mathbf{U}_i\cdot\mathbf{v}_i + \frac{3}{2}R_i^2|\mathbf{v}_i|^2 + F\right) + 2\frac{p_\infty}{\rho_\ell}\left(-R_0^{3\gamma}R_i^{-3\gamma+2} + R_i^2\right) = 0, \tag{26}$$

$$\frac{d}{dt}\left(\frac{1}{3}R_i^3\mathbf{U}_i - R_i^3\mathbf{v}_i\right) - \left(-2R_i^2\dot{R}_i\mathbf{v}_i + R_i^3\left(\nabla\mathbf{v}_i\right)^T\cdot(\mathbf{v}_i - \mathbf{U}_i) + G\right) = 0, \tag{27}$$

$$\psi_i = \sum_{i\neq k}\left(\phi_k(\mathbf{x}_i) + I_k\psi_k(\mathbf{x}_i)\right), \tag{28}$$

$$\mathbf{v}_i = \nabla\psi_i. \tag{29}$$

As it stands, the above system is rather intractable for analysis. To proceed further we must make a simplifying approximation, which will be to keep only terms that arise from monopoles and dipoles. Therefore, the F and G terms in (26) and (27) will be ignored. We will also use this approximation to simplify (28) and (29) as follows: from Theorem 3.2 we have

$$I_k\psi_k(\mathbf{x}_i) = -\frac{1}{2}R_k^3\nabla\psi_k(\mathbf{x}_k)\cdot\nabla_{\mathbf{x}_k}\left(\frac{1}{|\mathbf{x}_i - \mathbf{x}_k|}\right) + (\text{higher order harmonics}),$$

which when used with (29) gives

$$I_k\psi_k(\mathbf{x}_i) = -\frac{1}{2}R_k^3\mathbf{v}_k\cdot\nabla_{\mathbf{x}_k}\left(\frac{1}{|\mathbf{x}_i - \mathbf{x}_k|}\right) + (\text{higher order harmonics}).$$

Next, we combine the above formula with (28) to obtain

$$\psi_i = \sum_{k\neq i}\left[-\frac{R_k^2\dot{R}_k}{|\mathbf{x}_i - \mathbf{x}_k|} + \frac{1}{2}R_k^3\nabla_{\mathbf{x}_i}\left(\frac{1}{|\mathbf{x}_i - \mathbf{x}_k|}\right)\cdot(\mathbf{U}_k - \mathbf{v}_k)\right] + (\text{higher order harmonics}).$$

Thus by ignoring the terms caused by spherical harmonics of orders higher than dipole, (26)–(29) simplify to

$$\frac{d}{dt}(R_i^3\dot{R}_i - R_i^2\psi_i) - \left(\frac{3}{2}R_i^2\dot{R}_i^2 + \frac{1}{4}R_i^2|\mathbf{U}_i|^2 - 2R_i\dot{R}_i\psi_i - \frac{3}{2}R_i^3\mathbf{U}_i\cdot\mathbf{v}_i + \frac{3}{4}R_i^2|\mathbf{v}_i|^2\right) + \frac{p_\infty}{\rho_\ell}(-R_0^{3\gamma}R_i^{-3\gamma+2} + R_i^2) = 0, \tag{30}$$

$$\frac{d}{dt}\left(\frac{1}{3}R_i^3\mathbf{U}_i - R_i^3\mathbf{v}_i\right) - \left(-2R_i^2\dot{R}_i\mathbf{v}_i + R_i^3\left(\nabla\mathbf{v}_i\right)^T\cdot(\mathbf{v}_i - \mathbf{U}_i)\right) = 0, \tag{31}$$

$$\psi_i = \sum_{k\neq i} -\frac{R_k^2\dot{R}_k}{|\mathbf{x}_i - \mathbf{x}_k|} + \frac{1}{2}R_k^3\nabla_{\mathbf{x}_i}\left(\frac{1}{|\mathbf{x}_i - \mathbf{x}_k|}\right)\cdot(\mathbf{U}_k - \mathbf{v}_k), \tag{32}$$

$$\mathbf{v}_i = \nabla_{\mathbf{x}_i}\psi_i = \sum_{k\neq i}\nabla_{\mathbf{x}_i}\left(-\frac{R_k^2\dot{R}_k}{|\mathbf{x}_i - \mathbf{x}_k|} + \frac{1}{2}R_k^3\nabla_{\mathbf{x}_i}\left(\frac{1}{|\mathbf{x}_i - \mathbf{x}_k|}\right)\cdot(\mathbf{U}_k - \mathbf{v}_k)\right). \tag{33}$$

These equations simplify greatly if we define the following ambient pressure motivated using Bernoulli's law. The ambient pressure associated with the ith bubble is defined as

$$\frac{p_i}{\rho_\ell} = \frac{p_\infty}{\rho_l} - \frac{\partial \psi_i}{\partial t} - \frac{1}{2}|\mathbf{v}_i|^2. \tag{34}$$

Then (30) and (31) can be written as

$$R_i\ddot{R}_i + \frac{3}{2}\dot{R}_i^2 - \frac{1}{4}|\mathbf{U}_i - \mathbf{v}_i|^2 + \frac{1}{\rho_\ell}(p_i - p_{g,i}) = 0, \tag{35}$$

$$\frac{1}{3}\dot{\mathbf{U}}_i - \frac{D\mathbf{v}_i}{Dt} + \frac{\dot{R}_i}{R_i}(\mathbf{U}_i - \mathbf{v}_i) = (\mathbf{v}_i - \mathbf{U}_i) \times (\nabla \times \mathbf{v}_i), \tag{36}$$

where

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + \mathbf{v}_i \cdot \nabla \quad \text{and} \quad p_{g,i} = p_\infty \left(\frac{R_0}{R_i}\right)^{3\gamma}.$$

In arriving at (36) we have used the identity

$$-(\mathbf{u} \cdot \nabla)\mathbf{v} + (\nabla \mathbf{v})^T \cdot \mathbf{u} = \mathbf{u} \times (\nabla \times \mathbf{v}).$$

In deriving these equations of motion we have not assumed that $\nabla \times \mathbf{v}_i = 0$; however, it is apparent from (33) that $\nabla \times \mathbf{v}_i = 0$ since $\mathbf{v}_i = \nabla \psi_i$. This means the right-hand side of (36) is zero. Nevertheless, we shall leave (36) in the form we have written since, as we shall see later, the continuum limit of $\mathbf{v}_i$ is not necessarily curl free.

3.3.1. Bubble motion in nonuniform flows. The motion of a bubble in a nonuniform potential flow has been the subject of much interest. If we consider a massless rigid spherical bubble with velocity $\mathbf{U}$ moving in liquid with an ambient velocity $\mathbf{v}$, then the equation of motion is

$$\frac{1}{2}\dot{U} = \frac{3}{2}\left(\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \nabla)\mathbf{v}\right), \tag{37}$$

where $\mathbf{v}$ is derived from a potential flow, thus $\nabla \times \mathbf{v} = 0$. This equation has been derived by Voinov, Voinov, and Petrov [54], Landweber and Miloh [27], van Wijngaarden [51], and Galper and Miloh [14]. Galper and Miloh [14, 15] also derive extensions of this formula for more complex problems. If $\dot{R}_i = 0$, then we see that (36) reduces to (37), where we have used $\nabla \times \mathbf{v}_i = 0$.

Therefore we see that our computation is in agreement with well-known results. The main contribution of our work is in finding a self-consistent expression for the ambient liquid velocity produced by the motion of the other bubbles.

4. Continuum limit. In the previous section, we derived the equations of motion for a finite number of bubbles. In this section, we will take the continuum limit to obtain our effective equations. This approach is similar to that used by solid state physics to obtain effective equations; see, for example, Batteh and Powell [5] or Rosenau [35]. It also very close to the approach used by Caflisch et al. [9]. This approach is expected to give a faithful approximation, provided that the wavelength of interest is considerably longer than the distance between particles. One of the important parts of this section is taking the continuum limit of (32). This is obtained by approximating the summation by an integration. Let us explain this with the following example.

4.0.1. Example. Consider the situation with N point charges located at $\mathbf{x}_j$ with charge q_j, where $j = 1, \ldots, N$. The ambient electric potential at $\mathbf{x}_i$ is

$$\psi_i(\mathbf{x}_i) = \sum_{i \neq j}^{N} \frac{q_j}{|\mathbf{x}_i - \mathbf{x}_j|}. \tag{38}$$

We suppose that there exists a smooth function $q(\mathbf{x})$ such that $q_j = q(\mathbf{x}_j)$; then this summation can be approximated as

$$\psi(\mathbf{x}) = \int \frac{\rho(\mathbf{y})q(\mathbf{y})}{|\mathbf{x} - \mathbf{y}|} d\mathbf{y}, \tag{39}$$

where $\rho(\mathbf{x})$ is the number of particles per unit volume. Since $-(4\pi|\mathbf{x}|)^{-1}$ is the fundamental solution of Laplace's equation in three space dimensions, then it follows that (39) is equivalent to

$$\triangle \psi = -4\pi q(\mathbf{x})\rho(\mathbf{x}). \tag{40}$$

4.1. Effective equations for bubbly flows. Our equations of motion for a finite number of bubbles are given by (35) and (36), with the ambient field determined from (32), (33), and (34). When passing to the continuum limit, we first assume that there exist functions $R(\mathbf{x}, t)$ and $\mathbf{U}(\mathbf{x}, t)$ such that $R_k(t) = R(\mathbf{x}_k, t)$ and $\mathbf{U}_k(t) = \mathbf{U}(\mathbf{x}_k, t)$. We note that this assumption indicates that we assume that, on length scales less than the wavelength of interest, the bubbles are all "doing the same thing." This assumption implies that neighboring bubbles are oscillating coherently and moving with the same velocity. If we assume that nearby bubbles have the same velocity, then this means that we are studying "cold" bubbly flows; in other words, we are ignoring effects of the fluctuation of the bubbles' velocity. These effects have been studied in simpler models of bubbly flows by Russo and Smereka [37] and Herrero, Lucquin-Desreux, and Perthame [20]. The effects of incoherent bubble oscillations have been considered by, for example, Carstensen and Foldy [11] and Smereka [44] for models that ignore the effects of bubble translation. We also point out that this same assumption was used by Zhang and Prosperetti [57] in what they call sharply peaked probability distributions. As pointed out in [44], this same assumption was used by van Wijngaarden [48] and Caflisch et al. [10].

Now we take the continuum limit of (32), following the approach outlined in the example above, and we find

$$\psi(\mathbf{x}, t) = \int \rho \left(-\frac{R^2(\mathbf{y}, t)\dot{R}(\mathbf{y}, t)}{|\mathbf{x} - \mathbf{y}|} + \frac{1}{2} R^3(\mathbf{y}, t) \nabla_{\mathbf{x}} \left(\frac{1}{|\mathbf{x} - \mathbf{y}|} \right) \cdot (\mathbf{U}(\mathbf{y}, t) - \mathbf{v}(\mathbf{y}, t)) \right) d\mathbf{y},$$

where $\dot{}$ denotes $\frac{d}{dt} = \partial_t + \mathbf{U} \cdot \nabla$ and $\rho = \rho(\mathbf{y}, t)$ is the number of bubbles per unit volume. ρ satisfies the following conservation equation,

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{U}) = 0, \tag{41}$$

and is a statement of particle conservation.

Using integration by parts on our expression for ψ, we have

$$\psi(\mathbf{x}, t) = \int \left(-\frac{\rho R^2 \dot{R}}{|\mathbf{x} - \mathbf{y}|} + \frac{1}{2} \frac{\nabla \cdot \left(\rho R^3 (\mathbf{U} - \mathbf{v}) \right)}{|\mathbf{x} - \mathbf{y}|} \right) d\mathbf{y},$$

which is equivalent to

$$\triangle\psi = -4\pi\left(-\rho R^2\dot{R} + \frac{1}{2}\nabla\cdot\left(\rho R^3(\mathbf{U}-\mathbf{v})\right)\right).$$

If we use the void fraction $\beta = \frac{4}{3}\pi R^3\rho$ instead of ρ, the above equation becomes

$$\triangle\psi = \frac{3\beta}{R}\dot{R} - \frac{3}{2}\nabla\left(\cdot\beta(\mathbf{U}-\mathbf{v})\right). \tag{42}$$

In a similar way we see that the continuum limit of (33) is

$$\mathbf{v} = \int\left(-R^2\dot{R}\nabla_{\mathbf{x}}\left(\frac{1}{|\mathbf{x}-\mathbf{y}|}\right) + \frac{1}{2}R^3\nabla^2_{\mathbf{x}}\left(\frac{1}{|\mathbf{x}-\mathbf{y}|}\right)\cdot(\mathbf{U}-\mathbf{v})d\mathbf{y}\right). \tag{43}$$

Caution has to be taken in evaluating the second part of the above integral, as

$$\int\nabla^2_{\mathbf{x}}\left(\frac{1}{|\mathbf{x}-\mathbf{y}|}\right)\cdot\mathbf{p}d\mathbf{y}$$

is singular and the integrand is not integrable. We take the principal value, which is defined. This is justified because bubble i is not in the original sum of (33). From Smereka [43, (27)] we have

$$⨍\nabla^2_{\mathbf{x}}\left(\frac{1}{|\mathbf{x}-\mathbf{y}|}\right)\cdot\mathbf{p}(\mathbf{y})d\mathbf{y} = \frac{4\pi}{3}\mathbf{p}(x) + \nabla_{\mathbf{x}}\int_V\nabla_{\mathbf{x}}\left(\frac{1}{|\mathbf{x}-\mathbf{y}|}\right)\cdot\mathbf{p}(\mathbf{y})d\mathbf{y}, \tag{44}$$

where $⨍$ is the principal value integral. Using the above formula in (43), we obtain

$$\mathbf{v} = \nabla\psi + \frac{\beta}{2}(\mathbf{U}-\mathbf{v}). \tag{45}$$

To obtain an expression for β we start with (41) and use $\beta = \frac{4}{3}\pi R^3\rho$ to obtain

$$\frac{d}{dt}\left(\frac{3\beta}{4\pi R^3}\right) + \left(\frac{3\beta}{4\pi R^3}\right)\nabla\cdot U = 0,$$

which can be simplified to

$$\frac{d\beta}{dt} - \frac{3\beta}{R}\frac{dR}{dt} + \beta\nabla\cdot\mathbf{U} = 0. \tag{46}$$

This is the conservation of volume for the gas phase. If we take the gradient of (45) and substitute it into (42), we find

$$\nabla\cdot(\beta\mathbf{U} + (1-\beta)\mathbf{v}) - \frac{3\beta}{R}\frac{dR}{dt} = 0. \tag{47}$$

This is exactly the conservation of total volume. This indicates that, to the level of our approximation, $\mathbf{v}$ is also the volume averaged liquid velocity. We can make this point even more transparent by subtracting (46) from (47), to obtain

$$\frac{\partial(1-\beta)}{\partial t} + \nabla\cdot((1-\beta)\mathbf{v}) = 0. \tag{48}$$

This is a statement of conservation of liquid volume.

The continuum limit of (35) and (36) are obtained by realizing that $\frac{d}{dt}$ is the material derivative. Collecting our results, we find the following set of effective equations:

$$R\frac{d^2R}{dt^2} + \frac{3}{2}\left(\frac{dR}{dt}\right)^2 - \frac{1}{4}|\mathbf{U}-\mathbf{v}|^2 + \frac{p-p_g}{\rho_\ell} = 0, \tag{49}$$

$$\frac{1}{3}\frac{d\mathbf{U}}{dt} - \frac{D\mathbf{v}}{Dt} + \frac{1}{R}\frac{dR}{dt}(\mathbf{U}-\mathbf{v}) + (\mathbf{U}-\mathbf{v})\times(\nabla\times\mathbf{v}) = 0, \tag{50}$$

$$\triangle\psi - \frac{3\beta}{R}\frac{dR}{dt} + \frac{3}{2}\nabla\cdot(\beta(\mathbf{U}-\mathbf{v})) = 0, \tag{51}$$

$$\mathbf{v} - \nabla\psi - \frac{\beta}{2}(\mathbf{U}-\mathbf{v}) = 0, \tag{52}$$

$$\frac{1}{2}\mathbf{v}^2 + \frac{\partial\psi}{\partial t} + \frac{p-p_\infty}{\rho_\ell} = 0, \tag{53}$$

$$\frac{\partial\beta}{\partial t} + \nabla\cdot(\beta\mathbf{U}) - \frac{3\beta}{R}\frac{dR}{dt} = 0, \tag{54}$$

$$p_g - p_\infty\left(\frac{R_0}{R}\right)^{3\gamma} = 0, \tag{55}$$

where

$$\frac{d}{dt} = \frac{\partial}{\partial t} + \mathbf{U}\cdot\nabla \quad \text{and} \quad \frac{D}{Dt} = \frac{\partial}{\partial t} + \mathbf{v}\cdot\nabla.$$

To summarize, in the above set of effective equations, (49) and (50) are the continuum versions of (35) and (36). Equations (51), (52), and (53) are the continuum versions of (32), (33), and (34), respectively. Finally, (54) is a statement of bubble number conservation (equivalently, conservation of liquid volume), and (55) is the equation of state for the gas contained in the bubbles. The equation for the conservation of liquid volume, (48), shows that the ambient velocity is well approximated by the average liquid velocity.

We observe from (52) that

$$\nabla\times\mathbf{v} = \nabla\psi\times\nabla\left(\frac{2}{2+\beta}\right) + \nabla\times(\beta\mathbf{U}),$$

which is not necessarily zero. This may seem strange since it appears from (33) that $\mathbf{v}_i$ is curl free. However, it is important to note that $\mathbf{v}_i(\mathbf{x})$ has singularities when $\mathbf{x} = \mathbf{x}_k$. In the discrete case these singularities are not important, as is clear from (33). However, when we pass to the continuum limit, these singularities become source terms which cause the continuum limit to have a nonzero curl.

A simple example of this behavior can be understood using the example given in subsection 4.01. It is clear that (38) is a harmonic function of $\mathbf{x}_i$ ($\triangle_{\mathbf{x}_i}\psi_i = 0$) except when $\mathbf{x}_i = \mathbf{x}_k$. The expression for (38) is singular at $\mathbf{x}_i = \mathbf{x}_k$, corresponding to the location of the charges. It is evident from (40) that the continuum limit, (39), of (38) is not harmonic due the presence of the charges.

In a similar way, it follows that while the discrete ambient velocity field $\mathbf{v}_i$ is curl free, the same is not true of the continuum limit. This behavior is not uncommon; for example, in the theory of dielectrics the local electric field is curl free, whereas the ambient field of a continuum of dipoles is not curl free (see, for example Lorrain

and Corson [29]). We are not claiming that this flow has vorticity. The ambient liquid velocity has a nonzero curl, whereas the local liquid velocity is curl free. This may sound contradictory, but it is not; the ambient liquid velocity does not satisfy the Euler equation. Another way to look at this term is that it reflects the vorticity present on the surface of each bubble, and the ambient liquid velocity is a homogenized liquid velocity that accounts for the bubbles. In fact, the net vorticity produced by a single bubble is zero. Therefore if we had a homogeneous suspension of bubbles all moving with same velocity, then $\mathbf{v}$ would be constant in space and $\nabla \times \mathbf{v} = 0$.

It is interesting to compare (50) with work of Auton, Hunt, and Prud'homme [1]. In this work the authors compute the force on a bubble in a nonuniform flow with vorticity. From their work it follows that the equation of motion of the bubble (in our notation) is given by

$$C_M \left(\frac{\partial \mathbf{U}}{\partial t} + \mathbf{U} \cdot \nabla \mathbf{U} \right) = (1 + C_M) \left(\frac{\partial \mathbf{v}}{\partial t} + \mathbf{v} \cdot \nabla \mathbf{v} \right) - C_L (\mathbf{U} - \mathbf{v}) \times \boldsymbol{\omega}, \tag{56}$$

where C_M is the added mass coefficient, C_L is the rotational lift coefficient, and $\boldsymbol{\omega}$ is the liquid vorticity. If we choose $C_M = \frac{1}{2}$ and $C_L = \frac{3}{2}$, then we see that our result is the same as theirs (using $\dot{R} = 0$). A priori we should not expect any close relation between (56) and (50), since in (50), $\nabla \times \mathbf{v}$ is not the vorticity of the liquid. In fact, Auton, Hunt, and Prud'homme argue for a spherical bubble that $C_L = \frac{1}{2}$, whereas we find $C_L = \frac{3}{2}$.

4.2. Sound speed. We now study sound propagation for one dimensional flows, and therefore we assume that our dependent variables depend on only one space coordinate, which we take to be x. We linearize (49) through (55) around the equilibrium

$$R = R_0, \quad U = v = \psi = 0, \quad \beta = \beta_0,$$

and obtain the linearized equations

$$\begin{aligned}
&\frac{\partial^2 R}{\partial t^2} + \omega_0^2 R - \frac{1}{R_0} \frac{\partial \psi}{\partial t} = 0, \\
&\frac{\partial U}{\partial t} - 3 \frac{\partial v}{\partial t} = 0, \\
&\frac{\partial^2 \psi}{\partial x^2} = \frac{3\beta_0}{R_0} \frac{\partial R}{\partial t} - \frac{3}{2} \beta_0 \frac{\partial (U - v)}{\partial x}, \\
&v = \frac{\partial \psi}{\partial x} + \frac{\beta_0}{2} (U - v), \\
&\frac{\partial \beta}{\partial t} - \frac{3\beta_0}{R_0} \frac{\partial R}{\partial t} + \beta_0 \frac{\partial U}{\partial x} = 0,
\end{aligned}$$

where the unsubscripted variables represent perturbations from equilibrium and

$$\omega_0 = \sqrt{\frac{3\gamma P_\infty}{\rho_\ell R_0^2}}$$

is the natural frequency of a single bubble in an unbounded fluid.

We let $(R, U, v, \psi, \beta) = (A, B, C, D, E)e^{i(\omega t - kx)}$ and find

$$\begin{pmatrix} -\omega^2+\omega_0^2 & 0 & 0 & -\frac{i\omega}{R_0} & 0 \\ 0 & i\omega & -3i\omega & 0 & 0 \\ -\frac{3i\omega\beta_0}{R_0} & -\frac{3ik\beta_0}{2} & \frac{3ik\beta_0}{2} & -k^2 & 0 \\ 0 & -\frac{\beta_0}{2} & 1+\frac{\beta_0}{2} & ik & 0 \\ -\frac{3\beta_0 i\omega}{R_0} & -ik\beta_0 & 0 & 0 & i\omega \end{pmatrix} \begin{pmatrix} A \\ B \\ C \\ D \\ E \end{pmatrix} = 0.$$

The above equations will have nontrivial solutions when the determinant is zero. Thus, we obtain the dispersion relation

$$3\omega^2\beta_0(1-\beta_0) = k^2R_0^2(1+2\beta_0)(\omega_0^2-\omega^2).$$

The effective sound speed is $c = \frac{\omega}{k}$ and

$$c^2 = \frac{R_0^2(1+2\beta_0)(\omega_0^2-\omega^2)}{3\beta_0(1-\beta_0)}. \tag{57}$$

If we let $\omega \to 0$, we have

$$c_0^2 = \frac{R_0^2(1+2\beta_0)\omega_0^2}{3\beta_0(1-\beta_0)},$$

which is the same as the expression given by Crespo [12] and Caflisch et al. [10]. The sound speed in (57) is also in agreement with Sangani [38]. We note that we have assumed that the liquid is incompressible; therefore to compare with other investigations one must consider the case $C_\ell \to \infty$, where C_ℓ is the speed of sound in the liquid region.

4.3. Void waves. Void waves have been observed, and various properties such as wave speed have been measured. Typically void waves travel at speeds much slower than sound waves. This means that void waves and sound waves interact weakly, and we will not make any significant error if we assume the bubble radius is fixed. In this case our system of equations becomes

$$\frac{1}{3}\frac{d\mathbf{U}}{dt} - \frac{D\mathbf{v}}{Dt} + (\mathbf{U}-\mathbf{v}) \times (\nabla \times \mathbf{v}) = 0, \tag{58}$$

$$\triangle\psi + \frac{3}{2}\nabla \cdot (\beta(\mathbf{U}-\mathbf{v})) = 0, \tag{59}$$

$$\mathbf{v} - \nabla\psi - \frac{\beta}{2}(\mathbf{U}-\mathbf{v}) = 0, \tag{60}$$

$$\frac{\partial\beta}{\partial t} + \nabla \cdot (\beta\mathbf{U}) = 0. \tag{61}$$

These equations simplify greatly if we consider flows in one spatial dimension. In this situation (59) and (60) become

$$\frac{\partial^2\psi}{\partial x^2} + \frac{3}{2}\frac{\partial}{\partial x}(\beta(U-v)) = 0 \quad \text{and} \quad v - \frac{\partial\psi}{\partial x} - \frac{\beta}{2}(U-v) = 0.$$

We eliminate ψ to obtain

$$\frac{\partial}{\partial x}\left((1-\beta)v + \beta U\right) = 0.$$

Since we are in the situation where the volume flux is zero, the above equation indicates that

$$v = \frac{-\beta U}{(1-\beta)}. \tag{62}$$

In the one dimensional case, (58) and (61) become

$$\frac{1}{3}\left(\frac{\partial U}{\partial t} + U\frac{\partial U}{\partial x}\right) - \left(\frac{\partial v}{\partial t} + v\frac{\partial v}{\partial x}\right) = 0, \tag{63}$$

$$\frac{\partial \beta}{\partial t} + \frac{\partial}{\partial x}(\beta U) = 0. \tag{64}$$

If we multiply (63) by 3/2 and use (62), we find

$$\frac{\partial}{\partial t}\left(\frac{(1+2\beta)}{2(1-\beta)}U\right) + \frac{\partial}{\partial x}\left(\frac{(1-2\beta-2\beta^2)}{4(1-\beta)^2}U^2\right) = 0. \tag{65}$$

Next we consider a new variable

$$M = \frac{\beta U}{h(\beta)}, \quad \text{where} \quad h(\beta) = \frac{2\beta(1-\beta)}{1+2\beta}; \tag{66}$$

then (64) and (65) can be written as the following system:

$$\begin{aligned} &\frac{\partial \beta}{\partial t} + \frac{\partial}{\partial x}(h(\beta)M) = 0, \\ &\frac{\partial M}{\partial t} + \frac{\partial}{\partial x}\left(\frac{1}{2}h'(\beta)M^2\right) = 0. \end{aligned} \tag{67}$$

This set of conservation laws can be written as

$$\frac{\partial}{\partial t}\begin{pmatrix} \beta \\ M \end{pmatrix} + \mathsf{A}\frac{\partial}{\partial x}\begin{pmatrix} \beta \\ M \end{pmatrix} = 0, \quad \text{where} \quad \mathsf{A} = \begin{pmatrix} h'(\beta)M & h(\beta) \\ \frac{1}{2}h''(\beta)M^2 & h'(\beta)M \end{pmatrix}. \tag{68}$$

The eigenvalues of A are

$$\lambda = M\left(h'(\beta) \pm \sqrt{\frac{1}{2}h(\beta)h''(\beta)}\right).$$

Upon substituting the expression for $h(\beta)$, we find

$$\lambda = \frac{2M}{(1+2\beta)^2}\left(1-2\beta-2\beta^2 \pm i\sqrt{3\beta(1-\beta)}\right). \tag{69}$$

This dispersion relation shows that the Fourier modes will increase in magnitude at a rate proportional to wave number. This indicates that a spatially uniform bubbly flow is unstable to all perturbations. In fact, the initial value problem for (68) is ill-posed. This is consistent with the bubble clustering observed in the numerical simulations done by Sangani and Didwania [39] and Smereka [42]. In the next section, we will compute the dispersion relation when viscosity and gravity are considered. We will also offer an explanation of the discrepancy between experiments and numerical simulations.

4.4. Comparison with previous work. Geurst [17, 18] derived a set of equations for two-phase flow using a variational method. The same set of equations were derived by Wallis [47] and Pauchon and Smereka [34] using different approaches. The equations contained one phenomenological relation, denoted $m_G(\beta)$, which Wallis calls the exteria. Pauchon and Smereka showed that Geurst's equations simplify greatly in the frame of reference where the volume flux is zero. If we use $m_G(\beta) = \beta/2$ in Guerst's equations (as written by Pauchon and Smereka), we find that they are identical to (67). (Note, however, that Pauchon and Smereka use $\Gamma(\beta)$, which is $1/h(\beta)$). It should also be noted that from the work of Smereka and Milton [41] one can show that the exteria, $m_G(\beta)$, is related to the virtual mass in the zero volume flux frame $m(\beta)$, as follows:

$$m_G(\beta) = \frac{m(\beta)}{\rho_\ell}(1-\beta) - \beta^2. \tag{70}$$

The virtual mass in the zero volume flux frame has been calculated by Zuber [59], van Wijngaarden [50], Biesheuvel and Spolestra [7], Wallis [47], and Smereka and Milton [41]. Zuber's result was

$$m(\beta) = \rho_\ell \frac{\beta}{2}\left(\frac{1+2\beta}{1-\beta}\right).$$

The results of the other investigators were similar. Smereka and Milton showed that Zuber's result was exact for a certain type of bubbly flow. If we substitute Zuber's result into (70), then we find $m_G(\beta) = \beta/2$. Thus we conclude that (67) is the same as that derived by Geurst when Zuber's expression for the virtual mass is used.

5. Effects of liquid viscosity and gravity. In this section, we consider the effects of liquid viscosity and gravity in an effort to understand the dynamics of void waves. We shall assume that the bubble radii are unchanging and of identical sizes.

The effective equations are derived, and the void wave speed obtained is in good agreement with experimental data. We also offer an explanation of why bubble clustering is not observed in experiments.

5.1. Equations of motion. We shall proceed in a fashion similar to that in section 3. The Lagrangian is given by

$$\mathcal{L} = \mathcal{K} - \mathcal{U},$$

where $\mathcal{K}$ is given by (20). The potential energy is modified by gravity as follows:

$$\mathcal{U} = \mathcal{U}_g + \rho_\ell g \sum_{k=1}^{N} \frac{4}{3}\pi R_k^3 z_k,$$

where z_k is the z coordinate of the kth bubble and $\mathcal{U}_g$ is given by (21). We shall include effects of liquid viscosity by using a dissipation function denoted as $\mathcal{D}$. The equations of motion are now

$$\frac{d}{dt}\frac{\partial \mathcal{L}}{\partial \mathbf{U}_i} - \frac{\partial \mathcal{L}}{\partial \mathbf{x}_i} = \frac{1}{2}\frac{\partial \mathcal{D}}{\partial \mathbf{U}_i}. \tag{71}$$

The amount of energy dissipated is given by

$$\mathcal{D} = \int_{V_\ell} D d\mathbf{x}, \qquad \text{where} \quad D = 2\mu \varepsilon_{ij} \cdot \varepsilon_{ij},$$

where ε_{ij} is the rate-of-strain tensor.

We shall assume that the Reynolds number is high, so the flow is close to potential flow except in the thin boundary layer wrapped around each bubble. We shall further assume that no significant amount of energy dissipates in the boundary layer. The verification of this assumption for one bubble can be found in Moore [31]. With this assumption one finds that

$$\varepsilon_{ij} = \frac{\partial^2 \phi}{\partial x_i \partial x_j}.$$

We can check that

$$D = \mu \Delta E, \qquad \text{where} \quad E = \nabla\phi \cdot \nabla\phi.$$

Using Green's theorem, we have

$$\mathcal{D} = \int_{V_\ell} \mu \Delta E dv = -\mu \int_S \nabla E \cdot \mathbf{n} ds = -\mu \int_S \frac{\partial E}{\partial n} ds = -\mu \int_S \frac{\partial (\nabla\phi \cdot \nabla\phi)}{\partial n} ds, \tag{72}$$

where the integral is taken on the surface of the bubbles, and $\mathbf{n}$ is an outward normal vector.

5.2. Calculation of drag. In this section we calculate the drag force of a bubble in the presence of a finite number of bubbles. We begin with the case of a single bubble.

5.2.1. Single bubble. We consider a single bubble, moving with a fixed radius and a translational velocity $\mathbf{U}$, in a fluid with a constant ambient velocity $\mathbf{v}_\infty$. The velocity potential in this case is

$$\phi = \frac{1}{2} R^3 \nabla_r \left(\frac{1}{|\mathbf{r} - \mathbf{x}|} \right) \cdot (\mathbf{U} - \mathbf{v}_\infty) + \mathbf{v}_\infty \cdot \mathbf{r}.$$

The energy dissipation is computed using (72) and found to be

$$\mathcal{D} = 12\pi\mu R |\mathbf{U} - \mathbf{v}_\infty|^2. \tag{73}$$

The drag on the bubble is then given by

$$\mathbf{F} = \frac{1}{2} \frac{\partial \mathcal{D}}{\partial \mathbf{U}} = 12\pi\mu R (\mathbf{U} - \mathbf{v}_\infty). \tag{74}$$

Levich [28] derived (74) using the method outlined here. Moore [31] determined the drag force by computing the pressure distribution around the bubble. Kang and Leal [23] and Stone [46] provide alternate derivations.

5.2.2. Two bubbles. For the case of two bubbles in an infinite liquid, van Wijngaarden and Kapteyn [52], using the energy dissipation argument, derived

$$\mathbf{F} = -12\pi\mu R (\mathbf{U} - 2\mathbf{v}_{ind}), \tag{75}$$

where $\mathbf{v}_{ind}$ is the velocity generated by the other bubble.

5.2.3. N bubbles. With the expression of the velocity potential (14), we calculate $\mathcal{D}$ in Appendix E. After neglecting terms caused by spherical harmonics of orders higher than dipole, we have

$$\mathcal{D} = 12\pi\mu R \sum_{i=1}^{N} |\mathbf{U}_i - \mathbf{v}_i(\mathbf{x}_i)|^2. \tag{76}$$

From this equation we see that the dissipation due to the ith bubble is

$$12\pi\mu R|\mathbf{U}_i - \mathbf{v}_i(\mathbf{x}_i)|^2.$$

This compares closely with (73), with $\mathbf{v}_\infty$ replaced by the ambient field of the ith bubble ($\mathbf{v}_i$). Nevertheless, the drag force will be different from (74) since the ambient velocity of the ith bubble will depend on the velocities of all of the bubbles. In Appendix E we compute the drag force and find

$$\frac{1}{2}\frac{\partial \mathcal{D}}{\partial \mathbf{U}_i} = 12\pi\mu R(\mathbf{U}_i - 2\mathbf{v}_i(\mathbf{x}_i) - \mathbf{w}_i(\mathbf{x}_i)), \tag{77}$$

where

$$\mathbf{w}_i(\mathbf{r}) = -\sum_{j\neq i} \frac{1}{2}R^3\nabla^2\left(\frac{1}{|\mathbf{r}-\mathbf{x}_j|}\right)\cdot\mathbf{v}_j.$$

This compares closely with the result by van Wijngaarden and Kapteyn [52].

The continuum limit of (77) is

$$12\pi\mu R(\mathbf{U} - 2\mathbf{v} - \mathbf{w}),$$

where $\mathbf{v}$ is given by (59) and (60); $\mathbf{w}$ is determined from

$$\triangle\chi - \frac{3}{2}\nabla\cdot(\beta\mathbf{v}) = 0, \tag{78}$$

$$\mathbf{w} - \nabla\chi + \frac{\beta\mathbf{v}}{2} = 0. \tag{79}$$

In one space dimension, (78) and (79) simplify to

$$\frac{\partial^2\chi}{\partial x^2} = \frac{3}{2}\frac{\partial\beta v}{\partial x} \quad \text{and} \quad w = \frac{\partial\chi}{\partial x} - \frac{1}{2}\beta v,$$

from which it follows that

$$\frac{\partial w}{\partial x} - \frac{\partial(\beta v)}{\partial x} = 0.$$

Since w must vanish if v vanishes, then we find

$$w = \beta v.$$

Therefore the drag force is

$$12\pi\mu R(U - (2+\beta)v). \tag{80}$$

Using the expression for v given by (62), we find that the drag force is

$$12\pi\mu RU\frac{1+\beta+\beta^2}{1-\beta}. \tag{81}$$

We can write this formula in a different form by noticing that the bubble's velocity relative to the ambient liquid velocity is

$$U - v = \frac{U}{1-\beta}.$$

Therefore we can rewrite the drag force as

$$12\pi\mu R(1+\beta+\beta^2)(U-v).$$

We have shown, by following a procedure similar to that outlined in Appendix E, that the formula above is valid for any value of the volume flux.

5.3. Void waves. With the drag force computed, we can now modify our model for void waves to include gravity and liquid viscosity. Following the approach previously outlined, we find that (71) becomes, in the continuum limit,

$$\frac{1}{3}\left(\frac{\partial U}{\partial t}+U\frac{\partial U}{\partial x}\right)-\left(\frac{\partial v}{\partial t}+v\frac{\partial v}{\partial x}\right)=\frac{2}{3}g\left(1-\frac{U}{U_\infty}\frac{1+\beta+\beta^2}{1-\beta}\right), \tag{82}$$

$$\frac{\partial \beta}{\partial t}+\frac{\partial\left(\beta U\right)}{\partial x}=0, \tag{83}$$

where v is given by (62) and

$$U_\infty=\frac{R^2\rho_\ell g}{9\mu}$$

is the steady speed of a single bubble rising in an infinite fluid under the force of gravity. Next we multiply (82) by 3/2, use the expression for v, and rewrite (82) as

$$\frac{\partial}{\partial t}\left(\frac{(1+2\beta)}{2(1-\beta)}U\right)+\frac{\partial}{\partial x}\left(\frac{(1-2\beta-2\beta^2)}{4(1-\beta)^2}U^2\right)=g-\frac{gU}{U_\infty}\left(\frac{(1+\beta+\beta^2)}{1-\beta}\right). \tag{84}$$

Our model for void wave propagation including dissipation is then given by (83) and (84).

It is easy to verify that (83) and (84) have the equilibrium solution $\beta=\beta_0$ and $U=U_0$, where

$$U_0=\frac{1-\beta_0}{1+\beta_0+\beta_0^2}U_\infty. \tag{85}$$

This corresponds to a spatially uniform mixture of bubbles rising due to gravity in the zero volume flux frame of reference. The rise speed of the bubbles is given by (85). The prediction of (85) is in good agreement with the experimental data reported by Lammers and Biesheuvel [26] as shown in Figure 1 below. This result seems somewhat paradoxical; it has been shown by Sangani and Didwania [39], Smereka [42], and van Wijngaarden [53] that, in the context of potential theory, there is not a stable steady homogeneous distribution of rising bubbles; the key word here is stable. We shall now show that this steady solution that we have calculated is in fact unstable, in agreement with [39, 42, 53]. In fact we conjecture that this steady state is only weakly unstable, which is why (85) is in good agreement with experimental data.

Next we wish to examine the stability of this equilibrium solution. For this purpose it is useful to use the change of variables described in section 3. That is, we consider $M=\beta U/h(\beta)$; then (83) and (84) can be written as the following system:

$$\begin{aligned}&\frac{\partial \beta}{\partial t}+\frac{\partial}{\partial x}\left(h(\beta)M\right)=0,\\ &\frac{\partial M}{\partial t}+\frac{\partial}{\partial x}\left(\frac{1}{2}h'(\beta)M^2\right)=g\left(1-\frac{M}{M_0(\beta)}\right),\end{aligned} \tag{86}$$

where $h(\beta)$ is given by (66) and

$$M_0(\beta)=U_\infty\frac{1+2\beta}{2(1+\beta+\beta^2)}.$$

In these variables the equilibrium solution is $(\beta, M) = (\beta_0, M_0(\beta_0))$. If we linearize (86) about the equilibrium solution, we obtain the following linear system, with M and β now being the linearized variables:

$$\frac{\partial}{\partial t}\begin{pmatrix} \beta \\ M \end{pmatrix} + \mathsf{A}_0 \frac{\partial}{\partial x}\begin{pmatrix} \beta \\ M \end{pmatrix} = \mathsf{R}\begin{pmatrix} \beta \\ M \end{pmatrix}, \tag{87}$$

where

$$\mathsf{A}_0 = \begin{pmatrix} h'(\beta_0)M_0 & h(\beta_0) \\ \frac{1}{2}h''(\beta_0)M_0^2 & h'(\beta_0)M_0 \end{pmatrix} \quad \text{and} \quad \mathsf{R} = \frac{g}{M_0(\beta_0)}\begin{pmatrix} 0 & 0 \\ M_0'(\beta_0) & -1 \end{pmatrix}.$$

Next we look for solutions of the form $(\beta, M) = (A, B)e^{i(\omega t - kx)}$. We find that there will be solutions of this form, provided that the following dispersion relationship is satisfied:

$$z^2 + \frac{g}{ikM_0(\beta_0)}z - h(\beta_0)\left(\frac{1}{2}h''(\beta_0)M_0^2(\beta_0) + \frac{gM_0'(\beta_0)}{ikM_0(\beta_0)}\right) = 0, \tag{88}$$

with $z = c - h'(\beta_0)M_0(\beta_0)$, where $c = \frac{\omega}{k}$ is the phase speed. Solving the above equation for z reveals that z always has complex roots, indicating that the initial value problem for (87) is ill-posed. Hence, the growth rate of a Fourier mode is proporational to its wave number. This seems consistent with numerical simulations of bubble clustering. In these simulations a spatially uniform distribution of bubbles quickly assembles into horizontal clusters of bubbles (see Sangani and Didwania [39] and Smereka [42]). However, in experiments such behavior is not observed and is inconsistent with experimental observations of void wave propagation. Nevertheless, we shall see below that our model can predict some phenomena seen in experiments. To this end, let us then consider situations where the wavelengths tend to be large and therefore k is small. For small k the dispersion relationship has the two solutions

$$c_1 = \frac{ig}{M_0(\beta)k} + h'(\beta_0)M_0(\beta_0) - h(\beta_0)M_0'(\beta_0) + ic_I + O(k^2),$$

$$c_2 = (h(\beta_0)M_0(\beta_0))' - ic_I + O(k^2),$$

where

$$c_I = \frac{kh(\beta_0)M_0(\beta_0)}{2g}\left(2h(\beta_0)(M_0'(\beta_0))^2 - M_0^2(\beta_0)h''(\beta_0)\right).$$

Substituting in our expressions for h and M_0, these become

$$c_1 = i\frac{2g}{kU_\infty}\frac{1+\beta_0+\beta_0^2}{1+2\beta_0} + U_\infty(1 - 6\beta_0 + O(\beta_0^2)) + ic_I,$$

$$c_2 = c_R - ic_I,$$

where

$$c_R = \frac{1 - 2\beta_0 - 2\beta_0^2}{(1+\beta_0+\beta_0^2)^2}U_\infty \tag{89}$$

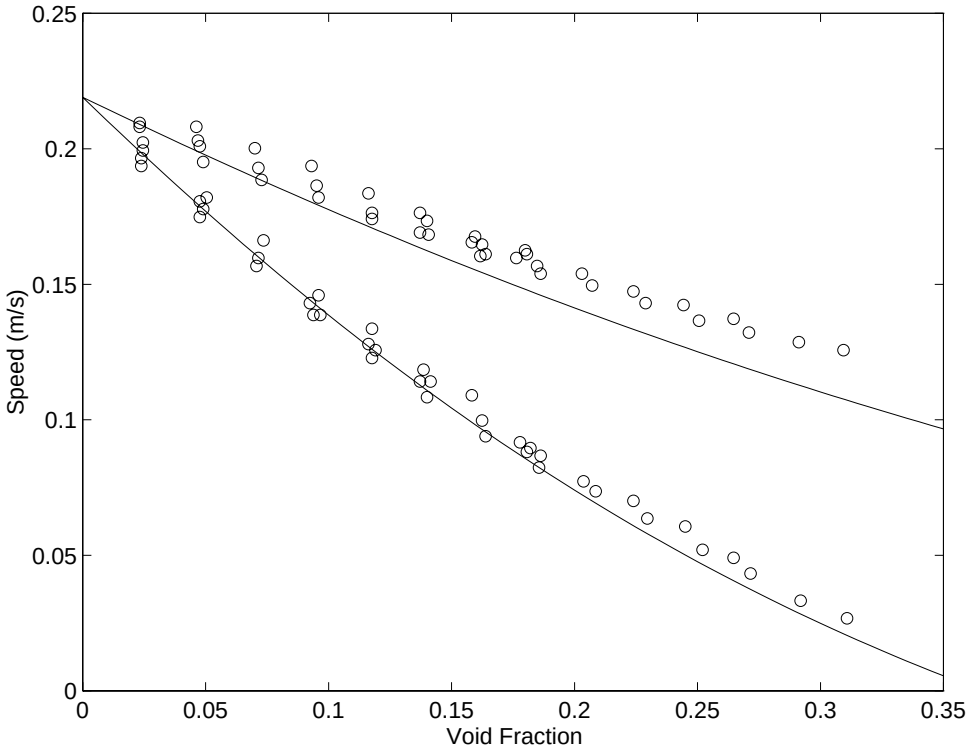


FIG. 1. *The upper curve shows a plot of the predicted bubble rise speed using* (85), *and the lower curve shows the predicted void wave speed using* (89). *We have used* $U_\infty = .219$ *m/sec. The circles are the experimental findings of Lammers and Biesheuvel* [26], *as is the value for* U_∞.

and

$$c_I = \frac{U_\infty^3 k}{2g}(3\beta_0 + O(\beta_0^2)). \tag{90}$$

In the expression for c_1 we observe that its imaginary part is positive. This indicates that this mode decays. This corresponds to the relaxation of the bubble's speed to the equilibrium speed. The second mode corresponds to void waves. It predicts that the void waves will move with speed c_R and will grow with a rate given by $c_I k$. Figure 1 shows a plot of c_R as function of the volume fraction along with experimental data. The agreement is good.

In the experiments of Lammers and Biesheuvel [26] the rise speed of a single bubble was approximately 25 cm/sec, the void fraction was in the range 0 to 0.4, and the frequency of naturally occurring void waves was approximately 1 Hz (see Biesheuvel and Gorissen [6]). Using our expression for the real part of the wave speed, we can estimate that this corresponds to disturbances with a wavelength of approximately 27 cm ($k \approx .2$). The growth rate of these disturbances (using (90)) is

$$c_I k \approx \frac{3\beta_0 U_\infty^3 k^2}{2g}.$$

If we use the experimental parameters given above, we find that the growth rate is approximately β_0. This suggests that the void waves do not have time to grow

significantly in normal experimental settings. Therefore our model predicts that the observed void waves should grow slightly and travel with a phase speed given by c_R. Recent experiments of Zenit, Koch, and Sangani [58] demonstrate a small amount of clustering.

Therefore, it appears that our model provides an accurate description of long wavelength disturbances. It is possible that the model breaks down for small wavelength disturbances and that there are regularizing effects which, when included in our model, will result in a well-posed model. Another possibility is that Sangani and Didwania [39] and Smereka [42] overestimate the amount of cluster formation observed in experiments. This could be because in these numerical simulations the computational domain was a cube with a size of only a few centimeters. Thus they were exciting modes of a much smaller wavelength than those observed in experiments, due to the periodic boundary conditions.

Finally, we remark that, if we assume the wave frequencies and wave numbers are small, then it follows from (84) that

$$g - \frac{gU}{U_\infty}\left(\frac{(1+\beta+\beta^2)}{1-\beta}\right) \approx 0,$$

which implies that

$$U \approx U_0(\beta),$$

where U_0 is given by (85). We can rewrite (83) using the above expression as

$$\frac{\partial \beta}{\partial t} + (U_0(\beta) + \beta U_0'(\beta))\frac{\partial \beta}{\partial x} \approx 0.$$

This equation was obtained by Lammers and Biesheuvel [26]. Using (85), we note that the above equation can be written as

$$\frac{\partial \beta}{\partial t} + c_R \frac{\partial \beta}{\partial x} \approx 0,$$

where c_R is given by (89). Thus we see, again, that when the frequency and wavenumber of the waves are small, the void wave should travel with a speed given by c_R.

6. Summary and conclusions. In this paper we have developed a new method for solving Laplace's equation for the velocity potential in a liquid with a finite number of bubbles. This method is a generalization of the method of images. Our approach also allows us to define the ambient velocity and ambient pressure associated with a particular bubble. The velocity potential is then used to calculate the total kinetic energy of the liquid. We then use the Euler–Lagrange equation to compute exact equations of motion for a finite collection of bubbles, which are a set of ordinary differential equations. We then make a simplifying approximation, which is to keep only terms arising from monopoles and dipoles. We then take the continuum limit of the equations of motion to obtain a set of partial differential equations that represent our effective equations for ideal bubbly fluids. Our model includes both sound and void wave propagation, includes nonlinear effects, and is valid over a wide range of wave numbers. We show that our model captures the results for the speed of sound waves from Caflisch et al. [10], Crespo [12], and Sangani [38]. We also show that our model reduces to Geurst's model [17, 18] when we consider void wave propagation.

We then consider the effects of liquid viscosity and gravity on void wave propagation. The effects of liquid viscosity are incorporated by using an energy dissipation function. We apply this technique for finite collection of bubbles, thus extending the work of van Wijngaarden and Kapteyn [52] for the two-bubble problem. We then compute the drag force. The continuum limit of the drag force is found, and our effective equations for void waves including gravity and liquid viscosity are formulated. We then observe that our model has a steady-state solution which corresponds to a mixture of bubbles rising with a steady speed. The calculated bubble rise speed is in good agreement with experimental values. We also compute the speed of void waves and find good agreement with experimental results. Our computations show that these waves are unstable, but, using the experimental parameters, we find that the instability can be small.

Appendix A. Proof of Theorem 3.1. Theorem 3.1 states that the method of images, when generalized to N spheres, results in a converging sequence which is the solution to Laplace's equation with the correct boundary conditions.

We begin our proof with some definitions. Let $B = B(\mathbf{p}, R)$ be a ball of radius R centered at the point $\mathbf{p}$. We define the energy norms:

$$||f||_B = \int_B |\nabla f|^2 d\mathbf{x} \quad \text{and} \quad ||f||_{\overline{B}} = \int_{\overline{B}} |\nabla f|^2 d\mathbf{x},$$

where $\overline{B}$ is the region exterior to B. We will also use the L_2 norm on the surface of B (∂B):

$$||f||_{\partial B} = \left(\int_{\partial B} f^2 ds \right)^{\frac{1}{2}}.$$

We shall make use of Weiss' sphere theorem, which, in the notation of our paper, is the following: *If $f(\mathbf{x})$ is harmonic inside $B = B(\mathbf{p}, R)$, then the image operator with respect to B is*

$$I_B f(r, \theta, \psi) = \frac{1}{R} \int_0^{\frac{R^2}{r}} w \frac{\partial f}{\partial w}(w, \theta, \psi) dw.$$

This can be found, for example, in Milne-Thompson [30, p. 520].

Remark. If the closest singularity of f has distance d from $\mathbf{p}$, then $I_B f$ is harmonic for all points outside of $B^*(\mathbf{p}, \frac{R^2}{d})$, which is a sphere smaller than $B(\mathbf{p}, R)$.

Our proof begins with the following three lemmas.

LEMMA A.1. *$B^m = B^m(\mathbf{p}, mR)$, $B = B(\mathbf{p}, R)$, and $B^M = B^M(\mathbf{p}, MR)$ are three concentric spheres with*

$$m < 1 < M, \quad c < 1, \quad 1 < cMm;$$

f is a harmonic function inside B^M; and $I_B f$ is harmonic exterior to B^m. Then we have

$$||I_B f||_{\overline{B^m}} < c||f||_{B^M},$$

where $\overline{B^m}$ is the region exterior to B^m.

Proof. Assume that $\mathbf{p}$ is at the origin. We can write f in terms of spherical harmonics:

$$f(r,\theta,\psi) = f(0) + \sum_{k=1}^{\infty}\sum_{j=1}^{2k+1} c_{k,j}h_{k,j}(\theta,\psi)r^k,$$

where $c_{k,j}$ are constants and $h_{k,j}$ is a set of spherical harmonics which satisfy the orthogonality condition

$$\int_{\text{unit ball}} h_{k,j}h_{m,i}dS = \delta_{km}\delta_{ij}.$$

From Weiss' sphere theorem, we have

$$\begin{aligned} I_B f &= \frac{1}{R}\int_0^{\frac{R^2}{r}} w\frac{\partial f}{\partial w}(w,\theta,\psi)dw \\ &= \frac{1}{R}\sum_{k=1}^{\infty}\sum_{j=1}^{2k+1}\frac{k}{k+1}c_{k,j}h_{k,j}R^{k+1}\Big|_0^{\frac{R^2}{r}} \\ &= \sum_{k=1}^{\infty}\sum_{j=1}^{2k+1}\frac{kR^{2k+1}}{(k+1)r^{k+1}}c_{k,j}h_{k,j}. \end{aligned}$$

We can use the divergence theorem and the expansion of f in spherical harmonics to obtain

$$||f||_{B^M} = \int_{\partial B^M} f\frac{\partial f}{\partial n}ds = \sum_{k=1}^{\infty}\sum_{j=1}^{2k+1} z_{k,j}^{(1)} \tag{91}$$

with

$$z_{k,j}^{(1)} = k(MR)^{2k+1}c_{k,j}^2.$$

In a similar fashion,

$$||I_B f||_{\overline{B^m}} = -\int_{\partial B^m} I_B f\frac{\partial I_B f}{\partial n}ds = \sum_{k=1}^{\infty}\sum_{j=1}^{2k+1} z_{k,j}^{(2)}, \tag{92}$$

where

$$z_{k,j}^{(2)} = \frac{k^2R^{4k+2}}{(k+1)(mR)^{2k+1}}c_{k,j}^2.$$

Therefore one has

$$\frac{z_{k,j}^{(2)}}{z_{k,j}^{(1)}} = \frac{k}{k+1}\left(\frac{1}{M^{2k+1}m^{2k+1}}\right).$$

Since $cmM < 1$, it follows that

$$\frac{z_{k,j}^{(2)}}{z_{k,j}^{(1)}} < c.$$

Hence

$$||I_B f||_{\overline{B^m}} < c||f||_{B^M}.$$

This completes the proof of Lemma A.1.

LEMMA A.2. *If f is harmonic outside $B(p,R)$, then*

$$||f||^2_{\partial B} \leq R||f||_{\overline{B}},$$

where $\overline{B}$ is the region outside of B.

Proof. After expanding

$$f(r,\theta,\psi) = \sum_{k=1}^{\infty}\sum_{m=1}^{2k-1} \frac{c_{k,m}h_{k,m}(\theta,\psi)}{r^{k+1}},$$

we have

$$||f||^2_{\partial B} = \sum_{k=1}^{\infty}\sum_{m=1}^{2k-1} \frac{c^2_{k,m}}{R^{2k}} \leq R\sum_{k=1}^{\infty}\sum_{m=1}^{2k-1} \frac{kc^2_{k,m}}{R^{2k+1}} = R||f||_{\overline{B}}.$$

This completes the proof of Lemma A.2.

LEMMA A.3. *If $f_k, k = 1,2,\ldots,$ are harmonic functions outside $B(\mathbf{p},R)$ and*

$$||f_k||_{\overline{B}} < Yc^k, \qquad c < 1,\ Y \text{ are constants},$$

then there exists a function f such that, at all points $\mathbf{x}$ outside B, f is harmonic and

$$\lim_{k\to\infty}\sum_{j=1}^{k} f_j(\mathbf{x}) = f(\mathbf{x}),$$

$$\lim_{\mathbf{x}\to\infty} f(\mathbf{x}) = 0.$$

Furthermore, the convergence is uniform outside $B^*(\mathbf{p},R^*)$ for any $R^* > R$.

Proof. We assume that $\mathbf{p}$ is at the origin. From Lemma A.2, we have

$$||f_k||^2_{\partial B} < R||f_k||_{\overline{B}} < RYc^k.$$

Hence

$$||f_k||_{\partial B} < \sqrt{RY}(\sqrt{c})^k.$$

Therefore $\{\sum_{i=1}^{k} f_i\}_k$ is a Cauchy sequence in the $||\cdot||_{\partial B}$ norm. Since the L^2 space on ∂B is complete, there exists a function $f \in L^2(\partial B)$ such that

$$\lim_{k\to\infty}\left\|\sum_{i=1}^{k} f_i - f\right\|_{\partial B} = 0.$$

We define f, for any $\mathbf{x}$ outside B, by using

$$f(\mathbf{x}) = \int_{\partial B} P(\mathbf{x},\mathbf{y})f(\mathbf{y})dS_{\mathbf{y}}, \tag{93}$$

where $P(\mathbf{x},\mathbf{y})$ is the Poisson kernel. From Axler, Bourdon, and Ramey [2], we have

$$P(\mathbf{x},\mathbf{y}) = \frac{1}{4\pi R}\frac{|\mathbf{x}|^2 - R^2}{|\mathbf{x}-\mathbf{y}|^3}.$$

Therefore

$$\left| f(\mathbf{x}) - \sum_{i=1}^{k} f_i(\mathbf{x}) \right| = \left| \int_{\partial B} P(\mathbf{x},\mathbf{y}) \left(f(\mathbf{y}) - \sum_{i=1}^{k} f_i(\mathbf{y}) \right) dS_{\mathbf{y}} \right|$$
$$\leq ||P(\mathbf{x},\mathbf{y})||_{\partial B} \cdot \left\| f - \sum_{i=1}^{k} f_i \right\|_{\partial B},$$

where the last inequality comes from the Cauchy–Schwarz inequality We also have a uniform bound on P for all $\mathbf{y} \in \partial B$, when $\mathbf{x}$ is a fixed point strictly outside B,

$$P(\mathbf{x},\mathbf{y}) \leq \frac{1}{4\pi R}\frac{|\mathbf{x}|^2 - R^2}{(|\mathbf{x}| - |\mathbf{y}|)^3} = \frac{1}{4\pi R}\frac{|\mathbf{x}| + R}{(|\mathbf{x}| - R)^2}.$$

Hence

$$\lim_{k\to\infty} \left| f(\mathbf{x}) - \sum_{i=1}^{k} f_i(\mathbf{x}) \right| = 0. \tag{94}$$

Furthermore, if $|\mathbf{x}| \geq R^* > R$, then

$$P(\mathbf{x},\mathbf{y}) \leq \frac{1}{4\pi R}\frac{R^* + R}{(R^* - R)^2}.$$

Thus the convergence in (94) is uniform for $|\mathbf{x}| \geq R^*$.

In this context, f is harmonic because, from (93) and the fact that $P(\mathbf{x},\mathbf{y})$ is harmonic in $\mathbf{x}$ if $\mathbf{y}$ is fixed,

$$\triangle f = \int_{\partial B} \triangle_{\mathbf{x}} P(\mathbf{x},\mathbf{y}) f(\mathbf{y}) dS_{\mathbf{y}} = 0.$$

Since $P(\mathbf{x},\mathbf{y})$ vanishes when $\mathbf{x} \to \infty$, then so does f because of (93). This completes the proof of Lemma A.3.

Proof of Theorem 3.1. Since all the spheres $B_i(\mathbf{x}_i, R_i)$, $i = 1,\ldots,N$, do not intersect and from the remark after the statement of Weiss' sphere theorem, we conclude that $I_i I_j \cdots \phi_k$ is harmonic not only outside $B_i(\mathbf{x}_i, R_i)$, but also outside a smaller sphere $B^*(\mathbf{x}_i, R_i^2/d_{ij})$, where d_{ij} is the distance between $\mathbf{x}_i$ and the closest point on ∂B_j because $I_j \cdots \phi_k$ has singularities only inside B_j. Therefore, we can find constant m, M, c such that

- $m < 1 < M$, $c < 1$, $1 < cMm$,
- $B_i^M(p_i, MR_i)$ do not intersect,
- $I_i \cdots I_j \phi_k$ is harmonic outside $B_i^m(p_i, mR_i)$.

To achieve this we let 2ε be the smallest distance between the surface of any of the N spheres. Let $q = \min_i \frac{R_i+\varepsilon}{R_i}$; then the choice $M = q^{\frac{3}{4}}$, $m = q^{-\frac{1}{4}}$, and $c = m$ will work (for example).

Next, we write

$$\phi = \sum_{j=1}^{N} \Omega_j,$$

where

$$\Omega_j = \phi_j + \sum_{\substack{i=i_1 \\ i_1 \neq j}}^{N} I_j \phi_i + \cdots + \sum_{\substack{i_1,\ldots,i_k=1 \\ i_1 \neq j, i_\ell \neq i_{\ell+1}}}^{N} I_j I_{i_1} I_{i_2} \cdots I_{i_{k-1}} \phi_{i_k} \cdots . \tag{95}$$

From the remarks above we know that Ω_j is harmonic outside $B_j^m(\mathbf{x}_j, mR_j)$.

Our plan is to prove that each term of Ω_j satisfies an estimate of the form given in Lemma A.3. Therefore we must estimate each term in the series. We denote the kth term of (95) as T_k and let

$$Y = \sum_{i=1}^{N} ||\phi_i||_{\overline{B_i^m}}.$$

Then it is obvious that

$$||T_0||_{\overline{B_j^m}} = ||\phi_j||_{\overline{B_j^m}} < Y. \tag{96}$$

To estimate the second term we appeal to Lemma A.1, and we have for $i_1 \neq j$

$$||I_j \phi_{i_1}||_{\overline{B_j^m}} < c||\phi_{i_1}||_{B_j^M} < c||\phi_{i_1}||_{\overline{B_{i_1}^m}}.$$

The second inequality follows since B_j^M is contained in $\overline{B_{i_1}^m}$. Now, we sum the above inequality over i_1 to obtain

$$||T_1||_{\overline{B_j^m}} < c \sum_{\substack{i=i_1 \\ i_1 \neq j}}^{N} ||\phi_{i_1}||_{\overline{B_{i_1}^m}} < cY. \tag{97}$$

Now, we look at the kth term and use Lemma A.1 to obtain the following estimate:

$$||T_k||_{\overline{B_j^m}} < c \sum_{\substack{i_1,\ldots,i_k=1 \\ i_1 \neq j, i_\ell \neq i_{\ell+1}}}^{N} ||I_{i_1} I_{i_2} \cdots I_{i_{k-1}} \phi_{i_k}||_{B_j^M}.$$

Next we use the above estimate and the fact that B_j^M is contained in $\overline{B_{i_1}^m}$ for $i_1 \neq j$ to find

$$||T_k||_{\overline{B_j^m}} < c \sum_{\substack{i_1,\ldots,i_k=1 \\ i_1 \neq j, i_\ell \neq i_{\ell+1}}}^{N} ||I_{i_1} I_{i_2} \cdots I_{i_{k-1}} \phi_{i_k}||_{\overline{B_{i_1}^m}}. \tag{98}$$

Applying Lemma A.1 again, we find

$$||T_k||_{\overline{B_j^m}} < c^2 \sum_{\substack{i_1,\ldots,i_k=1 \\ i_1 \neq j, i_\ell \neq i_{\ell+1}}}^{N} ||I_{i_2} \cdots I_{i_{k-1}} \phi_{i_k}||_{B_{i_1}^M}. \tag{99}$$

Since $i_1 \neq i_2$, it then follows that all $B_{i_1}^M$ are contained in $\overline{B_{i_2}^m}$, and we find from the above inequality

$$||T_k||_{\overline{B_j^m}} < c^2 \sum_{\substack{i_2,\ldots,i_k=1 \\ i_\ell \neq i_{\ell+1}}}^{N} ||I_{i_2} \cdots I_{i_{k-1}} \phi_{i_k}||_{\overline{B_{i_2}^m}}. \tag{100}$$

This is of the same form as (98), so therefore we repeat the same steps as were used to obtain (99) and (100) and thereby obtain the estimate

$$||T_k||_{\overline{B_j^m}} < c^k Y.$$

Next we combine Lemma A.2 and Lemma A.3 to conclude that Ω_j is harmonic and uniformly convergent outside bubble j. It then follows that ϕ is also harmonic and uniformly convergent in the liquid region. The completes our proof of Theorem 3.1.

Appendix B. Proof of Theorem 3.2. We begin with the following formula. Suppose $\mathbf{r} = (x_1, x_2, x_3)$; then we have

(101)
$$(-1)^n \left(\nabla^n \left(\frac{1}{|\mathbf{r}|}\right)\right)_{i_1 \cdots i_n} = \frac{(2n-1)!! x_{i_1} \cdots x_{i_n}}{|\mathbf{r}|^{2n+1}} + \sum_{j=1}^{N_2} \frac{(-1)^j (2n-2j-1)!! A_j}{|\mathbf{r}|^{2n-2j+1}},$$

with $i_1, i_2, \ldots, i_n = 1, 2, 3$; N_2 is the integer part of $\frac{n}{2}$; and

$$A_j = \sum \delta_{k_1 k_2} \cdots \delta_{k_{2j-1} k_{2j}} x_{k_{2j+1}} \cdots x_{k_n},$$

where the sum is over all possible j pairs $(k_1, k_2) \cdots (k_{2j-1}, k_{2j})$ from i_1 to i_n.

This formula can be proven by induction. Some examples of (101) are

$$\left(\nabla \left(\frac{1}{|\mathbf{r}|}\right)\right)_i = -\frac{x_i}{|\mathbf{r}|^3},$$
$$\left(\nabla^2 \left(\frac{1}{|\mathbf{r}|}\right)\right)_{ij} = \frac{3x_i x_j}{|\mathbf{r}|^5} - \frac{\delta_{ij}}{|\mathbf{r}|^3},$$
$$\left(\nabla^3 \left(\frac{1}{|\mathbf{r}|}\right)\right)_{ijk} = \frac{-15 x_i x_j x_k}{|\mathbf{r}|^7} + \frac{3(\delta_{ij} x_k + \delta_{ik} x_j + \delta_{jk} x_i)}{|\mathbf{r}|^5},$$

where i, j, and k run from 1 to 3.

One can use (101) to prove the following:

(102)
$$\frac{(-1)^k}{(2k-1)!!} \nabla^k f \cdot \nabla^k \left(\frac{1}{|\mathbf{r}|}\right) = \nabla^k f \cdot \frac{\mathbf{n}^k}{|\mathbf{r}|^{k+1}},$$

where $k \geq 2$ and

$$\mathbf{n} = \frac{\mathbf{r}}{|\mathbf{r}|}.$$

To prove (102) is straightforward. We first notice, since f is harmonic, that

$$\nabla^k f \cdot \delta_{ij} = 0.$$

Combining this result with (101), we obtain (102). We are now ready to prove Theorem 3.2; without losing the generality, we assume that $\mathbf{p}$ is the origin. According to Weiss's sphere theorem, we have

$$I_B f(\mathbf{r}) = \frac{1}{R} \int_0^{\frac{R^2}{|\mathbf{r}|}} w \nabla f(w\mathbf{n}) \cdot \mathbf{n} dw.$$

Using a Taylor series expansion for $f(\mathbf{r})$, we can rewrite the above expression as

$$I_B f(\mathbf{r}) = \frac{1}{R}\int_0^{\frac{R^2}{|\mathbf{r}|}} \sum_{k=1}^{\infty} \frac{1}{(k-1)!} w^k \nabla^k f(0)\cdot \mathbf{n}^k dw.$$

Integrating term by term, we obtain

$$I_B f(\mathbf{r}) = \sum_{k=1}^{\infty} \frac{R^{2k+1}}{(k-1)!(k+1)|\mathbf{r}|^{k+1}} \nabla^k f(0)\cdot \mathbf{n}^k. \tag{103}$$

Next we combine (102) with (103) to obtain Theorem 3.2. This completes the proof of Theorem 3.2.

Appendix C. Useful formulas. If $g(\mathbf{x})$ is a harmonic function in $B(\mathbf{p}, R)$ and $\mathbf{d}$, $\mathbf{e}$ are constant vectors, then

$$\int_{\partial B} g(\mathbf{x}) ds = 4\pi R^2 g(\mathbf{p}),$$

$$\int_{\partial B} (\mathbf{d}\cdot\mathbf{n}) g(\mathbf{x}) ds = \frac{4\pi R^3}{3} \mathbf{d}\cdot\nabla g(\mathbf{p}),$$

$$\int_{\partial B} (\mathbf{d}\cdot\mathbf{n})(\mathbf{e}\cdot\mathbf{n}) ds = \frac{4\pi}{3} R^2 \mathbf{d}\cdot\mathbf{e},$$

$$\int_{\partial B} I_B g(\mathbf{x}) ds = 0,$$

$$\text{and} \quad \int_{\partial B} (\mathbf{d}\cdot\mathbf{n})\, I_B g(\mathbf{x}) ds = \frac{2\pi R^3}{3} \mathbf{d}\cdot\nabla g(\mathbf{p}).$$

These formulas are established using the orthogonality properties of spherical harmonics. Theorem 3.2 is used in the proof of the last two equations.

Appendix D. Derivative calculations. In this section we shall compute the derivatives of K that appear in the Euler–Lagrange equations. When computing these derivatives, we will not use the expression for K given by (20). Instead we will use the integral form of K, which we rewrite here:

$$K = -\frac{1}{2}\rho_\ell \sum_{j=1}^{N} \int_{S_j} \phi \frac{\partial \phi}{\partial n} ds. \tag{104}$$

We have found it useful, when computing these derivatives, to introduce the operator J_i, which is defined as follows: if we have N spheres $B_1, B_2, \ldots, B_N$ and a function $f(\mathbf{x})$, which is harmonic outside of B_i, then we say that

$$g(\mathbf{x}) = J_i f(\mathbf{x})$$

if

$$\begin{aligned} &\frac{\partial g}{\partial n} = \frac{\partial f}{\partial n} \quad \text{at } \partial B_i, \\ &\frac{\partial g}{\partial n} = 0 \qquad \text{at } \partial B_j \text{ when } j \neq i. \end{aligned} \tag{105}$$

We note that $g(\mathbf{x})$ will be harmonic outside all of the spheres.

The operator can be expressed in terms of the image operator (defined in Theorem 3.1) as follows:

$$J_i(f) = f + \sum_{i_1=1,i_1\neq i}^{N} I_{i_1} f + \cdots + \sum_{i_1,\ldots,i_k=1,i_k\neq i,i_j\neq i_{j+1}}^{N} I_{i_1} I_{i_2} \cdots I_{i_k} f + \cdots. \tag{106}$$

This can be seen by applying Theorem 3.1 with $\phi_i = f$ and $\phi_j = 0$ if $j \neq i$.

It is easy to verify that the solution to (11) is

$$\phi = \sum_{i=1}^{N} J_i \phi_i, \tag{107}$$

where ϕ_i is defined in Theorem 3.1.

One property associated with operator J that we will use, if f and g are harmonic, is

$$\int_S J_i f \frac{\partial g}{\partial n} ds = \int_S \frac{\partial J_i f}{\partial n} g ds = \int_{S_i} \frac{\partial f}{\partial n} g ds, \tag{108}$$

where the first equality is Green's theorem and the second equality follows from the definition of the operator J.

Proof of (22). We have from (104)

$$\begin{aligned} K &= -\frac{\rho_\ell}{2} \sum_{j=1}^{N} \int_{S_j} \phi \frac{\partial \phi}{\partial n} ds \\ &= -\frac{\rho_\ell}{2} \rho_\ell \sum_{j=1}^{N} \int_{S_j} \phi (\dot{R}_j + \mathbf{U}_j \cdot \mathbf{n}) ds. \end{aligned}$$

Thus we have

$$-\frac{2}{\rho_\ell} \frac{\partial K}{\partial \dot{R}_i} = \int_{S_i} \phi ds + \sum_{j=1}^{N} \int_{S_j} \frac{\partial \phi}{\partial \dot{R}_i} \frac{\partial \phi}{\partial n} ds. \tag{109}$$

We will consider the two terms on the right-hand side of (109) separately. One has

$$\int_{S_i} \phi ds = \int_{S_i} (\phi_i + \psi_i + I_i \psi_i) ds.$$

It follows from Appendix C that this becomes

$$\int_{S_i} \phi ds = 4\pi R_i^2 (-R_i \dot{R}_i + \psi_i(\mathbf{x}_i)). \tag{110}$$

Turning to the second term, we have

$$\begin{aligned} \sum_{j=1}^{N} \int_{S_j} \frac{\partial \phi}{\partial \dot{R}_i} \frac{\partial \phi}{\partial n} ds &= \int_S \frac{\partial \phi}{\partial \dot{R}_i} \frac{\partial \phi}{\partial n} ds \\ &= \sum_{j=1}^{N} \int_S \frac{\partial J_j \phi_j}{\partial \dot{R}_i} \frac{\partial \phi}{\partial n} ds, \end{aligned}$$

where we have used (107). Next we observe that $J_j\phi_j$ does not depend on $\dot{R}_i$ unless $i = j$; thus we have

$$\sum_{j=1}^{N}\int_{S_j}\frac{\partial\phi}{\partial\dot{R}_i}\frac{\partial\phi}{\partial n}ds = \int_S \frac{\partial J_i\phi_i}{\partial\dot{R}_i}\frac{\partial\phi}{\partial n}ds.$$

Furthermore, the operator J_i does not depend on $\dot{R}_i$; thus the previous expression becomes

$$\sum_{j=1}^{N}\int_{S_j}\frac{\partial\phi}{\partial\dot{R}_i}\frac{\partial\phi}{\partial n}ds = \int_S J_i\left(\frac{\partial\phi_i}{\partial\dot{R}_i}\right)\frac{\partial\phi}{\partial n}ds. \tag{111}$$

Applying (108), we find

$$\int_S J_i\left(\frac{\partial\phi_i}{\partial\dot{R}_i}\right)\frac{\partial\phi}{\partial n}ds = \int_{S_i}\frac{\partial}{\partial n}\left(\frac{\partial\phi_i}{\partial\dot{R}_i}\right)\phi ds = \int_{S_i}\phi ds. \tag{112}$$

Combining (109), (110), (111), and (112), we find

$$\frac{\partial K}{\partial\dot{R}_i} = 2\pi\rho_\ell(2R_i^3\dot{R} - 2R_i^2\psi_i(\mathbf{x}_i)).$$

Thus (22) is proven. The proof for (23) is similar to that above.

Proof of (24). When calculating $\frac{\partial K}{\partial R_i}$, we notice that R_i enters into K in (104) in three ways:

- the integration region depends on R_i,
- ϕ_i depends on R_i,
- the image potential operator I_i depends on R_i.

Therefore we have

$$\frac{\partial K}{\partial R_i} = \frac{\partial K}{\partial R_i^S} + \frac{\partial K}{\partial R_i^\phi} + \frac{\partial K}{\partial R_i^I},$$

where R_i^S, R_i^ϕ, and R_i^I represent R_i in the integration region, in ϕ_i, and in I_i, respectively.

Step 1. We first assume that ϕ_i and I_i are fixed and consider only the effect of changing the integration region. We have

$$-\frac{2}{\rho_\ell}\frac{\partial K}{\partial R_i^S} = \frac{\partial}{\partial R_i^S}\int_S\phi\frac{\partial\phi}{\partial n}ds = \frac{\partial}{\partial R_i^S}\int_{S_i}\phi\frac{\partial\phi}{\partial n}ds.$$

It follows from the above result and (17) that

$$-\frac{2}{\rho_\ell}\frac{\partial K}{\partial R_i^S} = \frac{\partial}{\partial R_i^S}\int_{S_i}(\phi_i + \psi_i + I_i\psi_i)\frac{\partial\phi}{\partial n}ds.$$

We continue our calculation by applying Theorem 3.2 to expand $I_i\psi_i$ and using a Taylor expansion for ψ_i to obtain

$$\begin{aligned}-\frac{2}{\rho_\ell}\frac{\partial K}{\partial R_i^S} = \frac{\partial}{\partial R_i^S}\int_{S_i}\Bigg[&-\frac{R_i^2\dot{R}_i}{|\mathbf{r}-\mathbf{x}_i|} + \frac{1}{2}R_i^3\nabla_r\left(\frac{1}{|\mathbf{r}-\mathbf{x}_i|}\right)\cdot\mathbf{U}_i\\ &+\psi_i(\mathbf{x}_i) + \mathbf{v}_i(\mathbf{x}_i)\cdot(\mathbf{r}-\mathbf{x}_i) - \frac{1}{2}R_i^3\mathbf{v}_i(\mathbf{x}_i)\cdot\nabla_{\mathbf{r}}\left(\frac{1}{|\mathbf{r}-\mathbf{x}_i|}\right)\\ &+(\text{higher order harmonics})\Bigg](\dot{R}_i + \mathbf{U}_i\cdot\mathbf{n})ds.\end{aligned}$$

Evaluating the integrand on S_i ($|\mathbf{r}| = R_i$) and using the orthogonality properties of spherical harmonics, we find

$$-\frac{2}{\rho_\ell}\frac{\partial K}{\partial R_i^S} = \frac{\partial}{\partial R_i^S}\int_{S_i}\Bigg[-\frac{R_i^2\dot{R}_i}{R_i^S} - \frac{R_i^3}{2(R_i^S)^2}\mathbf{U}_i\cdot\mathbf{n} + \psi_i(\mathbf{x}_i) + R_i^S\mathbf{v}_i(\mathbf{x}_i)\cdot\mathbf{n} + \frac{R_i^3}{2(R_i^S)^2}\mathbf{v}_i(\mathbf{x}_i)\cdot\mathbf{n}\Bigg](\dot{R}_i + \mathbf{U}_i\cdot\mathbf{n})ds.$$

The integration over S_i is performed using the results in Appendix C, and we obtain

$$-\frac{2}{\rho_\ell}\frac{\partial K}{\partial R_i^S} = \frac{\partial}{\partial R_i^S}\Bigg[4\pi(R_i^S)^2\left(-\frac{R_i^2\dot{R}_i^2}{R_i^S} + \psi_i(\mathbf{x}_i)\dot{R}_i\right) + \frac{4\pi(R_i^S)^2}{3}\mathbf{U}_i\cdot\left(-\frac{R_i^3\mathbf{U}_i}{2(R_i^S)^2} + R_i^S\mathbf{v}_i(\mathbf{x}_i) + \frac{R_i^3}{2(R_i^S)^2}\mathbf{v}_i(\mathbf{x}_i)\right)\Bigg].$$

Next, we take the derivative of the above expression and evaluate it at $R_i^S = R_i$ to obtain

$$-\frac{2}{\rho_\ell}\frac{\partial K}{\partial R_i^S} = -4\pi R_i^2\dot{R}_i^2 + 8\pi R_i\dot{R}_i\psi_i(\mathbf{x}_i) + 4\pi R_i^2\mathbf{U}_i\cdot\mathbf{v}_i(\mathbf{x}_i). \tag{113}$$

Step 2. We now assume that the integration region and I_i are fixed, and that only R_i in ϕ_i is changing. We need to calculate

$$-\frac{2}{\rho_\ell}\frac{\partial K}{\partial R_i^\phi} = \frac{\partial}{\partial R_i^\phi}\int_S \phi\frac{\partial\phi}{\partial n}ds.$$

Applying (107) and using the fact that only ϕ_i is changing, we have

$$-\frac{2}{\rho_\ell}\frac{\partial K}{\partial R_i^\phi} = \frac{\partial}{\partial R_i^\phi}\int_S J_i(\phi_i)\frac{\partial\phi}{\partial n}ds.$$

In this case neither J_i nor $\frac{\partial\phi}{\partial n}$ depend on R_i^ϕ; thus we may write the above equation as

$$-\frac{2}{\rho_\ell}\frac{\partial K}{\partial R_i^\phi} = \int_S J_i\left(\frac{\partial\phi_i}{\partial R_i^\phi}\right)\frac{\partial\phi}{\partial n}ds.$$

Substituting our expression for ϕ_i given by (13), we have

$$-\frac{2}{\rho_\ell}\frac{\partial K}{\partial R_i^\phi} = \int_S J_i\left(-\frac{2R_i^\phi\dot{R}_i}{|\mathbf{r}-\mathbf{x}_i|} + \frac{3}{2}(R_i^\phi)^2\nabla_r\left(\frac{1}{|\mathbf{r}-\mathbf{x}_i|}\right)\cdot\mathbf{U}_i\right)\frac{\partial\phi}{\partial n}ds.$$

It follows from (108) and (17) that we can write the previous expression as

$$-\frac{2}{\rho_\ell}\frac{\partial K}{\partial R_i^\phi} = \int_{S_i}\frac{\partial}{\partial n}\left(-\frac{2R_i^\phi\dot{R}_i}{|\mathbf{r}-\mathbf{x}_i|} + \frac{3}{2}(R_i^\phi)^2\nabla_r\left(\frac{1}{|\mathbf{r}-\mathbf{x}_i|}\right)\cdot\mathbf{U}_i\right)(\phi_i + \psi_i + I_i\psi_i)ds.$$

Evaluating the normal derivatives at the bubble surface and setting $R_i^\phi = R_i$, we obtain

$$-\frac{2}{\rho_\ell}\frac{\partial K}{\partial R_i^\phi} = \int_{S_i}\left(\frac{2\dot{R}_i}{R_i} + \frac{3\mathbf{U}_i\cdot\mathbf{n}}{R_i}\right)(\phi_i + \psi_i + I_i\psi_i)ds.$$

Next, (13) and the results of Appendix C are used to deduce

$$-\frac{2}{\rho_\ell}\frac{\partial K}{\partial R_i^\phi} = -8\pi R_i^2\dot{R}_i^2 + 8\pi R_i\dot{R}_i\psi_i(\mathbf{x}_i) - 2\pi R_i^2|\mathbf{U}_i|^2 + 6\pi R_i^2\mathbf{U}_i\cdot\mathbf{v}_i(\mathbf{x}_i). \tag{114}$$

Step 3. Finally, we wish to calculate $\frac{\partial K}{\partial R_i}$, where only R_i in the operator I_i is changing. We start with

$$-\frac{2}{\rho_\ell}\frac{\partial K}{\partial R_i^I} = \frac{\partial}{\partial R_i^I}\int_S \phi\frac{\partial\phi}{\partial n}ds.$$

Since $\frac{\partial\phi}{\partial n}$ and the region of integration does not depend on R_i^I, we then have

$$-\frac{2}{\rho_\ell}\frac{\partial K}{\partial R_i^I} = \int_S \frac{\partial\phi}{\partial R_i^I}\frac{\partial\phi}{\partial n}ds. \tag{115}$$

The task at hand then is to calculate $\frac{\partial\phi}{\partial R_i^I}$. This will be done using (14). Some terms of ϕ in (14) have multiple occurrences of I_i. For these terms, we separate each into several terms, so that after the separation, each term only has one changing I_i, denoted by $\tilde{I}_i$. For example, we have

$$\frac{\partial}{\partial R_1^I}I_1I_2I_1I_3I_4I_1\phi_3 = \frac{\partial}{\partial R_1^I}\left(\tilde{I}_1I_2I_1I_3I_4I_1\phi_3 + I_1I_2\tilde{I}_1I_3I_4I_1\phi_3 + I_1I_2I_1I_3I_4\tilde{I}_1\phi_3\right).$$

By doing this, we find

$$\frac{\partial\phi}{\partial R_i^I} = \frac{\partial}{\partial R_i^I}\left(\left(Id + \sum_{j_l\neq j_{l+1}, j_m\neq i} I_{j_1}\cdots I_{j_m}\right)\tilde{I}_i\left(\sum_{i\neq k_1, k_l\neq k_{l+1}} I_{k_1}\cdots I_{k_n}\phi_{k_{n+1}}\right)\right),$$

where Id is the identity operator. From the expressions of J_i and ψ_i in (106) and (16), we obtain

$$\frac{\partial\phi}{\partial R_i^I} = \frac{\partial J_i\tilde{I}_i\psi_i}{\partial R_i^I}.$$

Since J_i does not depend on R_i^I, we can write the previous equation as

$$\frac{\partial\phi}{\partial R_i^I} = J_i\left(\frac{\partial\tilde{I}_i\psi_i}{\partial R_i^I}\right).$$

We use the above equation to rewrite (115) as

$$-\frac{2}{\rho_\ell}\frac{\partial K}{\partial R_i^I} = \int_S J_i\left(\frac{\partial\tilde{I}_i\psi_i}{\partial R_i^I}\right)\frac{\partial\phi}{\partial n}ds.$$

It follows from (108) that

$$-\frac{2}{\rho_\ell}\frac{\partial K}{\partial R_i^I} = \int_{S_i} \phi \frac{\partial}{\partial n}\left(\frac{\partial \tilde{I}_i \psi_i}{\partial R_i^I}\right) ds$$
$$= \int_{S_i} (\phi_i + \psi_i + I_i\psi_i) \frac{\partial}{\partial n}\left(\frac{\partial \tilde{I}_i \psi_i}{\partial R_i^I}\right) ds.$$

If we expand $\tilde{I}_i\psi_i$ using (103), we can show

$$\frac{\partial}{\partial n}\left(\frac{\partial \tilde{I}_i \psi_i}{\partial R_i^I}\right) = W = \sum_{k=1}^{\infty} \frac{-(2k+1)(R_i^I)^{k-2}}{(k-1)!} \nabla^k \psi_i(\mathbf{x}_i)\cdot \mathbf{n}^k.$$

We also expand $I_i\psi_i$ using (103) and expand ψ_i in a Taylor series to obtain

$$-\frac{2}{\rho_\ell}\frac{\partial K}{\partial R_i^I} = \int_{S_i} \Bigg[- R_i\dot{R}_i - \frac{R_i}{2}\mathbf{U}_i\cdot\mathbf{n} + \sum_{k=0}^{\infty} \frac{1}{k!} R_i^k \nabla^k \psi_i(\mathbf{x}_i)\cdot \mathbf{n}^k$$
$$+ \sum_{k=1}^{\infty} \frac{R_i^{2k+1}}{(k-1)!(k+1)R_i^{k+1}} \nabla^k \psi_i(\mathbf{x}_i)\cdot \mathbf{n}^k \Bigg] W ds$$

$$= \int_{S_i} \left(-\frac{R_i}{2}\mathbf{U}_i\cdot\mathbf{n} + \sum_{k=1}^{\infty} \frac{2k+1}{(k+1)!} R_i^k \nabla^k \psi_i(\mathbf{x}_i)\cdot \mathbf{n}^k \right) W ds$$

$$= 2\pi R_i^2 \mathbf{U}_i\cdot \mathbf{v}_i(\mathbf{x}_i) - 6\pi R_i^2 |\mathbf{v}_i(\mathbf{x}_i)|^2 + F, \tag{116}$$

where we have let $R_i^I = R_i$. F contains terms involving $\nabla^k\psi\cdot\nabla^k\psi$ with $k>1$. We have used the orthogonality of spherical harmonics and the results in Appendix C to arrive at (116). Combining all three parts ((113), (114), and (116)), we have

$$\frac{\partial K}{\partial R_i} = 2\pi\rho_\ell \left(3R_i^2\dot{R}_i^2 + \frac{1}{2}R_i^2|\mathbf{U}_i|^2 - 4R_i\dot{R}_i\psi_i(\mathbf{x}_i) - 3R_i^2\mathbf{U}_i\cdot\mathbf{v}_i(\mathbf{x}_i) + \frac{3}{2}R_i^2|\mathbf{v}_i|^2 + F \right),$$

which is (24).

Proof of (25). We first expand the monopole and dipole terms at $\mathbf{x}$ in Laurent series around $\mathbf{y}$:

$$\frac{1}{|\mathbf{r}-\mathbf{x}|} = \sum_{n=0}^{\infty} \frac{1}{n!} \frac{|\mathbf{x}-\mathbf{y}|^{2n+1}}{|\mathbf{r}-\mathbf{y}|^{2n+1}} \nabla_{\mathbf{y}}^n \left(\frac{1}{|\mathbf{x}-\mathbf{y}|}\right)\cdot(\mathbf{r}-\mathbf{y})^n$$
$$= \frac{1}{|\mathbf{r}-\mathbf{y}|} + \frac{(\mathbf{x}-\mathbf{y})\cdot(\mathbf{r}-\mathbf{y})}{|\mathbf{r}-\mathbf{y}|^3} + \cdots,$$
$$\nabla_{\mathbf{r}}\left(\frac{1}{|\mathbf{r}-\mathbf{x}|}\right) = \sum_{n=0}^{\infty} \frac{-(2n+1)}{n!} \frac{|\mathbf{x}-\mathbf{y}|^{2n+1}}{|\mathbf{r}-\mathbf{y}|^{2n+3}} \nabla_{\mathbf{y}}^n \left(\frac{1}{|\mathbf{x}-\mathbf{y}|}\right)\cdot(\mathbf{r}-\mathbf{y})^{n+1}$$
$$+ \frac{1}{(n-1)!} \frac{|\mathbf{x}-\mathbf{y}|^{2n+1}}{|\mathbf{r}-\mathbf{y}|^{2n+1}} \nabla_{\mathbf{y}}^n \left(\frac{1}{|\mathbf{x}-\mathbf{y}|}\right)\cdot(\mathbf{r}-\mathbf{y})^{n-1}$$
$$= \nabla_{\mathbf{r}}\left(\frac{1}{|\mathbf{r}-\mathbf{y}|}\right) + \cdots.$$

We follow the same procedure as in the previous section to calculate $\frac{\partial K}{\partial \mathbf{x}_i}$. If $\mathbf{x}_i^I$, $\mathbf{x}_i^\phi$, and $\mathbf{x}_i^I$ are used to represent $\mathbf{x}_i$ in the integration region, in ϕ_i, and in I_i, respectively, we can make our calculation by using the same arguments as in the previous section and the two Laurent series above. Without writing the details, we obtain

$$\frac{\partial}{\partial \mathbf{x}_i^S}\int_S \phi\frac{\partial \phi}{\partial n}ds = 4\pi R_i^2\dot{R}_i\mathbf{v}_i + \frac{4\pi}{3}R_i^2\dot{R}_i\mathbf{U}_i + \frac{4\pi}{3}R_i^2 + (\nabla\mathbf{v}_i)^T\cdot\mathbf{U}_i,$$

$$\frac{\partial}{\partial \mathbf{x}_i^\phi}\int_S \phi\frac{\partial \phi}{\partial n}ds = -\frac{4\pi}{3}R_i^2\dot{R}_i\mathbf{U}_i + 4\pi R_i^2\dot{R}_i\mathbf{v}_i + 2\pi R_i^3(\nabla\mathbf{v}_i)^T\cdot\mathbf{U}_i,$$

$$\frac{\partial}{\partial \mathbf{x}_i^I}\int_S \phi\frac{\partial \phi}{\partial n}ds = \frac{2\pi}{3}R_i^3(\nabla\mathbf{v}_i)^T\cdot\mathbf{U}_i - 4\pi R_i^3 + (\nabla\mathbf{v}_i)^T\cdot\mathbf{v}_i + G,$$

where $\mathbf{v}_i = \mathbf{v}_i(\mathbf{x}_i)$ and G will have only terms involving $\nabla^k\psi_i(\mathbf{x}_i)\cdot\nabla^{k+1}\psi_i(\mathbf{x}_i)$, with $k > 1$. Adding all three parts together, we have

$$\frac{\partial K}{\partial \mathbf{x}_i} = 2\pi\rho_\ell\left(-2R_i^2\dot{R}_i\mathbf{v}_i + R_i^3(\nabla\mathbf{v}_i)^T\cdot(\mathbf{v}_i - \mathbf{U}_i) + G\right).$$

Appendix E. Energy dissipation and drag force. In this appendix, we will calculate energy dissipation of bubbly flow with a finite number of bubbles. We then calculate the drag force on each bubble.

Energy dissipation. From (72), we have

$$\mathcal{D} = -\mu\int_S \frac{\partial(\nabla\phi\cdot\nabla\phi)}{\partial r}ds.$$

Using (17) in the above expression, we have

$$\mathcal{D} = -\mu\sum_{i=1}^N\int_{S_i}\frac{\partial}{\partial n}\left|\nabla\left(\phi_i + \psi_i + I_i\psi_i\right)\right|^2 ds.$$

Applying Theorem 3.2 and expanding ψ_i, we obtain

$$\begin{aligned}\mathcal{D} =& -\mu\sum_{i=1}^N\int_{S_i}\frac{\partial}{\partial n}\Bigg|\nabla\Bigg(-\frac{R_i^2\dot{R}_i}{|\mathbf{r}-\mathbf{x}_i|} + \frac{1}{2}R_i^3\nabla_{\mathbf{r}}\left(\frac{1}{|\mathbf{r}-\mathbf{x}_i|}\right)\cdot\mathbf{U}_i \\ &+ \psi_i(\mathbf{x}_i) + \mathbf{v}_i(\mathbf{x}_i)\cdot(\mathbf{r}-\mathbf{x}_i) - \frac{1}{2}R_i^3\mathbf{v}_i(\mathbf{x}_i)\cdot\nabla_{\mathbf{r}}\left(\frac{1}{|\mathbf{r}-\mathbf{x}_i|}\right) \\ &+ (\text{higher order harmonics})\Bigg)\Bigg|^2 ds \\ =& -\mu\sum_{i=1}^N\int_{S_i}\frac{\partial}{\partial n}\Bigg|-R_i^2\dot{R}_i\nabla_r\left(\frac{1}{|\mathbf{r}-\mathbf{x}_i|}\right) + \frac{1}{2}R_i^3\nabla_{\mathbf{r}}^2\left(\frac{1}{|\mathbf{r}-\mathbf{x}_i|}\right)\cdot(\mathbf{U}_i - \mathbf{v}_i) \\ &+ \mathbf{v}_i + (\text{higher order harmonics})\Bigg|^2 ds.\end{aligned}$$

Since spherical harmonics of different order are orthogonal to each other, we calculate each term in the above equation separately. For the monopole term, we have

$$\begin{aligned}\int_{S_i}\frac{\partial}{\partial n}\left|\nabla\left(-\frac{R_i^2\dot{R}_i}{|\mathbf{r}-\mathbf{x}_i|}\right)\right|^2 ds &= \int_{S_i}\frac{\partial}{\partial n}\left[\frac{R_i^4\dot{R}_i^2}{|\mathbf{r}-\mathbf{x}_i|^4}\right] ds\\ &= \int_{S_i}\frac{-4\dot{R}_i^2}{|\mathbf{r}-\mathbf{x}_i|} ds\\ &= -16\pi R_i\dot{R}_i^2.\end{aligned}$$

A direct calculation from Mathematica shows us that

$$\begin{aligned}&\int_{S_i}\frac{\partial}{\partial n}\left(\frac{\partial^2}{\partial x^2}\left(\frac{1}{|\mathbf{r}-\mathbf{x}_i|}\right)\right)^2 ds = -\frac{96\pi}{5R_i^5},\\ &\int_{S_i}\frac{\partial}{\partial n}\left(\frac{\partial^2}{\partial x^2}\left(\frac{1}{|\mathbf{r}-\mathbf{x}_i|}\right)\frac{\partial^2}{\partial y^2}\left(\frac{1}{|\mathbf{r}-\mathbf{x}_i|}\right)\right) ds = \frac{48\pi}{5R_i^5},\\ &\int_{S_i}\frac{\partial}{\partial n}\left(\frac{\partial^2}{\partial x^2}\left(\frac{1}{|\mathbf{r}-\mathbf{x}_i|}\right)\frac{\partial^2}{\partial x\partial y}\left(\frac{1}{|\mathbf{r}-\mathbf{x}_i|}\right)\right) ds = 0,\\ &\int_{S_i}\frac{\partial}{\partial n}\left(\frac{\partial^2}{\partial x^2}\left(\frac{1}{|\mathbf{r}-\mathbf{x}_i|}\right)\frac{\partial^2}{\partial y\partial z}\left(\frac{1}{|\mathbf{r}-\mathbf{x}_i|}\right)\right) ds = 0,\\ &\int_{S_i}\frac{\partial}{\partial n}\left(\frac{\partial^2}{\partial x\partial y}\left(\frac{1}{|\mathbf{r}-\mathbf{x}_i|}\right)\frac{\partial^2}{\partial x\partial y}\left(\frac{1}{|\mathbf{r}-\mathbf{x}_i|}\right)\right) = -\frac{72\pi}{5R_i^5},\\ &\int_{S_i}\frac{\partial}{\partial n}\left(\frac{\partial^2}{\partial x\partial y}\left(\frac{1}{|\mathbf{r}-\mathbf{x}_i|}\right)\frac{\partial^2}{\partial x\partial z}\left(\frac{1}{|\mathbf{r}-\mathbf{x}_i|}\right)\right) = 0.\end{aligned}$$

With these results, we can calculate dipole terms in energy dissipation and find

$$\begin{aligned}&\int_{S_i}\frac{\partial}{\partial n}\left|\nabla\left(\frac{1}{2}R_i^3\nabla_{\mathbf{r}}\left(\frac{1}{|\mathbf{r}-\mathbf{x}_i|}\right)\cdot(\mathbf{U}_i-\mathbf{v}_i(\mathbf{x}_i))\right)\right|^2 ds\\ &= -\frac{1}{4}R_i^6\left(\frac{96\pi}{5R_i^5}+\frac{72\pi}{5R_i^5}+\frac{72\pi}{5R_i^5}\right)|\mathbf{U}_i-\mathbf{v}_i(\mathbf{x}_i)|^2\\ &= -12\pi R_i|\mathbf{U}_i-\mathbf{v}_i(\mathbf{x}_i)|^2.\end{aligned}$$

Hence one finds

$$\begin{aligned}\mathcal{D} = \sum_{i=1}^N &\mu\pi R_i\left(16\dot{R}_i^2+12|\mathbf{U}_i-\mathbf{v}_i(\mathbf{x}_i)|^2\right)\\ &+(\text{terms caused by higher order harmonics}).\end{aligned}$$

Drag force. We assume there is no radial oscillation and all bubbles have same radius R. Then

$$\begin{aligned}\mathbf{F}_i &= -\frac{1}{2}\frac{\partial\mathcal{D}}{\partial\mathbf{U}_i} = -6\mu\pi R\frac{\partial}{\partial\mathbf{U}_i}\sum_{j=1}^N|\mathbf{U}_j-\mathbf{v}_j|^2\\ &= -6\mu\pi R\frac{\partial}{\partial\mathbf{U}_i}\sum_{j=1}^N(|\mathbf{U}_j|^2-2\mathbf{U}_j\cdot\mathbf{v}_j(\mathbf{x}_j)+|\mathbf{v}_j(\mathbf{x}_j)|^2)\\ &= -12\mu\pi R\left(\mathbf{U}_i-\frac{\partial}{\partial\mathbf{U}_i}\sum_{j=1}^N\mathbf{U}_j\cdot\mathbf{v}_j(\mathbf{x}_j)+\frac{1}{2}\frac{\partial}{\partial\mathbf{U}_i}\sum_{j=1}^N|\mathbf{v}_j(\mathbf{x}_j)|^2\right).\end{aligned}\tag{117}$$

From the expression for the kinetic energy K in (20), we have, when radial oscillations are absent,

$$\sum_{j=1}^{N} \mathbf{U}_j \cdot \mathbf{v}_j(\mathbf{x}_j) = -\frac{K}{\pi \rho_\ell R^3} + \frac{1}{3}\sum_{j=1}^{N} |\mathbf{U}_j|^2.$$

Hence one finds

$$\frac{\partial}{\partial \mathbf{U}_i}\sum_{j=1}^{N} \mathbf{U}_j \cdot \mathbf{v}_j(\mathbf{x}_j) = -\frac{1}{\pi \rho_\ell R^3}\frac{\partial K}{\partial \mathbf{U}_i} + \frac{2}{3}\mathbf{U}_i.$$

Using (23), we have

$$\frac{\partial}{\partial \mathbf{U}_i}\sum_{j=1}^{N} \mathbf{U}_j \cdot \mathbf{v}_j(\mathbf{x}_j) = 2\mathbf{v}_i(\mathbf{x}_i). \tag{118}$$

Unfortunately we cannot calculate exactly

$$\frac{\partial}{\partial \mathbf{U}_i}\sum_{j=1}^{N} |\mathbf{v}_j(\mathbf{x}_j)|^2,$$

and we will use (14) to provide an approximate calculation. Using the first term in (14), we have

$$\phi \approx \sum_{j=1}^{N} \phi_i,$$

and it follows that

$$\psi_j \approx \sum_{k\neq j, k=1}^{N} \phi_k.$$

With the expression for ϕ_j in (13), we have

$$\frac{\partial \mathbf{v}_j(\mathbf{x}_j)}{\partial \mathbf{U}_i} = \frac{\partial \nabla \psi_j(\mathbf{x}_j)}{\partial \mathbf{U}_i} \approx \begin{cases} \frac{1}{2}R_i^3 \nabla^2 \left(\frac{1}{|\mathbf{x}_i - \mathbf{x}_j|}\right), & i \neq j, \\ 0, & i = j. \end{cases}$$

Therefore we find (with $R_i = R$)

$$\frac{\partial}{\partial \mathbf{U}_i}\sum_{j=1}^{N} |\mathbf{v}_j(\mathbf{x}_j)|^2 = 2\sum_{j=1}^{N} \frac{\partial \mathbf{v}_j(\mathbf{x}_j)}{\partial \mathbf{U}_i} \cdot \mathbf{v}_j(\mathbf{x}_j) = \sum_{j=1, j\neq i}^{N} R^3 \nabla^2 \left(\frac{1}{|\mathbf{x}_i - \mathbf{x}_j|}\right) \cdot \mathbf{v}_j(\mathbf{x}_j). \tag{119}$$

Using (117), (118), and (119), we obtain

$$\mathbf{F}_i \approx -12\pi\mu R(\mathbf{U}_i - 2\mathbf{v}_i(\mathbf{x}_i) - \mathbf{w}_i(\mathbf{x}_i)),$$

where

$$\mathbf{w}_i(\mathbf{r}) = -\sum_{j\neq i} \frac{1}{2} R^3 \nabla^2 \left(\frac{1}{|\mathbf{r} - \mathbf{x}_j|}\right) \cdot \mathbf{v}_j(\mathbf{x}_j).$$

REFERENCES

[1] T. R. Auton, J. C. R. Hunt, and M. Prud'homme, *The force on a body in inviscid unsteady nonuniform rotational flow*, J. Fluid Mech., 197 (1988), pp. 241–257.

[2] S. Axler, P. Bourdon, and W. Ramey, *Harmonic Function Theory*, Springer, New York, 1992.

[3] G. K. Batchelor, *An Introduction to Fluid Dynamics*, Cambridge University Press, London, 1967.

[4] G. K. Batchelor, *A new theory of the instability of a uniform fluidized bed*, J. Fluid Mech., 193 (1988), pp. 75–110.

[5] J. H. Batteh and J. D. Powell, *Solitary wave propagation in the three-dimensional lattice*, Phys. Rev. B, 20 (1979), pp. 1398–1409.

[6] A. Biesheuvel and W. C. M Gorissen, *Void fraction disturbances in a uniform bubbly liquid*, Int. J. Multiphase Flow, 18 (1990), pp. 211–231.

[7] A. Biesheuvel and S. Spolestra, *The added mass coefficient of a dispersion of spherical gas bubble in liquid*, Int. J. Multiphase Flow, 15 (1989), p. 911–924.

[8] A. Biesheuvel and L. van Wijngaarden, *Two-phase flow equations for a dilute dispersion of gas bubbles in liquid*, J. Fluid Mech., 148 (1984), pp. 301–318.

[9] R. E. Caflisch, M. J. Miksis, G. C. Papanicolaou, and L. Ting, *Effective equations for wave propagation in bubbly flow*, J. Fluid Mech., 153 (1985), pp. 259–272.

[10] R. E. Caflisch, M. J. Miksis, G. C. Papanicolaou, and L. Ting, *Wave propagation in bubbly liquids at finite volume fraction*, J. Fluid Mech., 160 (1985), pp. 1–14.

[11] E. L. Carstensen and L. L. Foldy, *Propagation of sound through a liquid containing bubbles*, J. Acoust. Soc. Am., 19 (1947), p. 481–501.

[12] A. Crespo, *Sound and shock waves in liquids containing bubbles*, Phys. Fluids, 12 (1969), pp. 2274–2282.

[13] L. L. Foldy, *The multiple scattering of waves*, Phys. Rev., 67 (1945), p. 107–119.

[14] A. Galper and T. Miloh, *Generalized Kirchoff equations for a deformable body moving in a weakly uniform flow field*, Proc. Roy. Soc. London A, 446 (1994), pp. 169–193.

[15] A. Galper and T. Miloh, *Dynamic equations of motion for a rigid or deformable body in an arbitrary non-uniform potential flow field*, J. Fluid Mech., 295 (1995), pp. 91–120.

[16] A. R. Galper and T. Miloh, *Motion stability of a deformable body in an ideal fluid with applications to the N spheres problem*, Phys. Fluids, 10 (1998), pp. 119–130.

[17] J. A. Geurst, *Virtual mass in two-phase flow*, Phys. A, 129 (1985), p. 233–261.

[18] J. A. Geurst, *Variational principles and two-fluid hydrodynamics of bubbly liquid/gas mixtures*, Phys. A, 135 (1986), pp. 455–486.

[19] W. A. H. J. Hermans, *On the Instability of a Translating Gas Bubble Under the Influence of a Pressure Step*, Philips Res. Repts., Suppl., 1973.

[20] H. Herrero, B. Lucquin-Desreux, and B. Perthame, *On the motion of dispersed balls in a potential flow: A kinetic description of the added mass effect*, SIAM J. Appl. Math., 60 (1999), pp. 61–83.

[21] S. V. Iordanskii, *On the equations of motion for a liquid containing gas bubbles*, Zh. Prikl. Mekh. i Tekhn. Fiz., 3 (1960), pp. 102–110.

[22] D. J. Jeffrey, *Conduction through random suspension of spheres*, Proc. Roy. Soc. London A, 335 (1973), p. 355–367.

[23] I. S. Kang and L. G. Leal, *The drag coefficient for a spherical bubble in a uniform streaming-flow*, Phys. Fluids, 31 (1988), pp. 233–237.

[24] D. L. Koch and R. J. Hill, *Interial effects in suspension and porous-media flows*, Annu. Rev. Fluid Mech., 33 (2001), pp. 619–647.

[25] H. Lamb, *Hydrodynamics*, Dover, New York, 1932.

[26] J. H. Lammers and A. Biesheuvel, *Concentration waves and the instability of bubbly flows*, J. Fluid Mech., 328 (1996), pp. 67–93.

[27] L. Landweber and T. Miloh, *Unsteady Lagally theorem for multipoles and deformable bodies*, J. Fluid Mech., 96 (1980), p. 33–46.

[28] V. G. Levich, *Physiochemical Hydrodynamics*, Prentice–Hall, Englewood Cliffs, NJ, 1962.

[29] P. Lorrain and D. Corson, *Electromagnetic Fields and Waves*, 2nd ed., W. H. Freeman, San Francisco, 1970.

[30] L. M. Milne-Thompson, *Theoretical Hydrodynamics*, Macmillan, New York, 1969.

[31] D. W. Moore, *The boundary layer on a spherical gas bubble*, J. Fluid Mech., 16 (1963), pp. 161–176.

[32] L. Noordzij and L. van Wijngaarden, *Relaxation effects, caused by relative motion, on shock waves in gas-bubble/liquid mixtures*, J. Fluid Mech., 66 (1974), p. 115–143.

[33] J. W. Park, D. A. Drew, and R. T. Lahey, *The analysis of void wave propagation in adiabatic monodispersed bubbly two-phase flows using an ensemble-averaged model*, Int. J. Multiphase Flow, 24 (1998), pp. 1205–1244.
[34] C. Pauchon and P. Smereka, *Momentum interactions in dispersed flow*, Int. J. Multiphase Flow, 18 (1992), pp. 65–87.
[35] P. Rosenau, *Dynamics of dense lattices*, Phys. Rev. B, 36 (1987), pp. 5868–5876.
[36] D. K. Ross, *The potential due to two point charges each at the center of a spherical cavity and embedded in a dielectric medium*, Austral. J. Phys., 21 (1968), p. 817–822.
[37] G. Russo and P. Smereka, *Kinetic theory for bubbly flow* I: *Collisionless case*, SIAM J. Appl. Math., 56 (1996), pp. 327–357.
[38] A. S. Sangani, *A pairwise interaction theory for determining the linear acoustic properties of dilute bubbly liquids*, J. Fluid Mech., 232 (1991), pp. 221–284.
[39] A. S. Sangani and A. K. Didwania, *Dynamic simulations of flows of bubbly liquids at large Reynolds numbers*, J. Fluid Mech., 250 (1993), pp. 307–337.
[40] E. Silberman, *Sound velocity and attenuation in bubbly mixtures measured in standing wave tubes*, J. Acoust. Soc. Amer., 18 (1957), p. 925–933.
[41] P. Smereka and G. Milton, *Bubbly flow and its relation to conduction in composites*, J. Fluid Mech., 233 (1991), pp. 65–81.
[42] P. Smereka, *On the motion of bubbles in a periodic box*, J. Fluid Mech., 254 (1993), pp. 79–112.
[43] P. Smereka, *A Vlasov description of the Euler equation*, Nonlinearity, 9 (1996), pp. 1361–1386.
[44] P. Smereka, *A Vlasov equation for pressure wave propagation in bubbly fluids*, J. Fluid Mech., 454 (2002), pp. 287–325.
[45] P. D. Spelt and A. S. Sangani, *Properties and averaged equations for flows of bubbly liquids*, Appl. Sci. Res., 58 (1998), pp. 337–386.
[46] H. A. Stone, *An interpretation of the translation of drops and bubbles at high Reynolds-numbers in terms of the vorticity field*, Phys. Fluids A, 5 (1993), pp. 2567–2569.
[47] G. B. Wallis, *Interial coupling in two-phase flows: Macroscopic properties of suspensions in an inviscid fluid*, Multiphase Sci. Tech., 5 (1989), p. 239–261.
[48] L. van Wijngaarden, *On the equation of motion for mixtures of fluids and gas bubbles*, J. Fluid Mech., 33 (1968), pp. 465–474.
[49] L. van Wijngaarden, *One-dimensional flow of liquids containing small gas bubbles*, Ann. Rev. Fluid Mech., 4 (1974), pp. 369–396.
[50] L. van Wijngaarden, *Hydrodynamical interaction between gas bubbles in liquid*, J. Fluid Mech., 77 (1976), p. 27–44.
[51] L. van Wijngaarden, *On the motion of gas bubbles in a perfect liquid*, Arch. Mech., 34 (1982), pp. 343–349.
[52] L. van Wijngaarden and C. Kapteyn, *Concentration waves in dilute bubble/liquid mixtures*, J. Fluid Mech., 212 (1990), pp. 111–137.
[53] L. van Wijngaarden, *The mean rise velocity of pairwise-interacting bubbles in liquid*, J. Fluid Mech., 251 (1993), pp. 55–78.
[54] V. V. Voinov, O. V. Voinov, and A. G. Petrov, *Hydrodynamic interaction of bodies in an ideal incompressible liquid and their movement in inhomogeneous flows*, Prikl. Math Mech., 37 (1973), p. 680–689.
[55] Y. Yurkovetsky and J. Brady, *Statistical mechanics of bubbly liquids*, Phys. Fluids, 8 (1996), pp. 881–895.
[56] D. Z. Zhang and A. Prosperetti, *Averaged equations for inviscid dispersed* 2*-phase flow*, J. Fluid Mech., 267 (1994), pp. 185–219.
[57] D. Z. Zhang and A. Prosperetti, *Ensemble phase-averaged equations for bubbly flows*, Phys. Fluids, 6 (1994), pp. 2956–2970.
[58] R. Zenit, D. L. Koch, and A. S. Sangani, *Measurements of the average properties of a suspension of bubbles rising in a vertical channel*, J. Fluid Mech., 429 (2001), pp. 307–342.
[59] N. Zuber, *On the dispersed two-phase flow in the laminar regime*, Chem. Engrg. Sci., 19 (1964), p. 897–917.

SIAM J. APPL. MATH.
Vol. 63, No. 6, pp. 1889–1910

CRITICAL THRESHOLDS IN 2D RESTRICTED EULER–POISSON EQUATIONS*

HAILIANG LIU† AND EITAN TADMOR‡

Abstract. We provide a complete description of the critical threshold phenomenon for the two-dimensional localized Euler–Poisson equations, introduced by the authors in [*Comm. Math. Phys.*, 228 (2002), pp. 435–466]. Here, the questions of global regularity vs. finite-time breakdown for the two-dimensional (2D) restricted Euler–Poisson solutions are classified in terms of precise explicit formulae, describing a remarkable variety of critical threshold surfaces of initial configurations. In particular, it is shown that the 2D critical thresholds depend on the relative sizes of three quantities: the initial density, the initial divergence, and the initial spectral gap, that is, the difference between the two eigenvalues of the 2×2 initial velocity gradient.

Key words. critical thresholds, restricted Euler–Poisson dynamics, spectral gap

AMS subject classifications. Primary, 35Q35; Secondary, 35B30

DOI. 10.1137/S0036139902416986

1. Introduction and statement of main results.

We are concerned with the critical threshold phenomenon in multidimensional Euler–Poisson equations. In this paper we consider a localized version of the following two-dimensional (2D) Euler–Poisson equations:

$$\partial_t \rho + \nabla \cdot (\rho U) = 0, \quad x \in \mathbb{R}^2, \quad t \in \mathbb{R}^+, \tag{1.1}$$

$$\partial_t (\rho U) + \nabla \cdot (\rho U \otimes U) = -k\rho \nabla \phi, \tag{1.2}$$

$$-\Delta \phi = \rho - c, \quad x \in \mathbb{R}^2, \tag{1.3}$$

which are the usual statements of the conservation of mass, Newton's second law, and the Poisson equation defining, say, the electric field in terms of the charge. Here $k > 0$ is a scaled physical constant, which signifies the property of the underlying repulsive forcing (avoiding the case of an attractive force with $k < 0$), and c denotes the constant "background" state. The unknowns are the local density $\rho = \rho(x,t)$, the velocity field $U = (u,v)(x,t)$, and the potential $\phi = \phi(x,t)$. It follows that, as long as the solution remains smooth, the velocity U solves a forced transport equation

$$\partial_t U + U \cdot \nabla U = F, \quad F = -k\nabla \phi, \tag{1.4}$$

with ϕ being governed by Poisson's equation (1.3).

This hyperbolic-elliptic coupled system (1.1)–(1.3) describes the dynamic behavior of many important physical flows, including charge transport [25], plasma with collision [15], cosmological waves [3], and the expansion of cold ions [13]. Let us men-

*Received by the editors November 2, 2002; accepted for publication (in revised form) March 5, 2003; published electronically August 15, 2003.

http://www.siam.org/journals/siap/63-6/41698.html

†Department of Mathematics, Iowa State University, Ames, IA 50011 (hliu@iastate.edu). The research of this author was supported in part by NSF grant DMS01-07917.

‡Department of Mathematics, Center for Scientific Computation and Mathematical Modeling (CSCAMM), and Institute for Physical Science and Technology (IPST), University of Maryland, College Park, MD 20742 (tadmor@cscamm.umd.edu). The research of this author was supported in part by ONR grant N00014-91-J-1076 and by NSF grant DMS01-07917.

tion that the Euler–Poisson equations could also be realized as the semiclassical limit of the Schrödinger–Poisson equation and are found in the "cross section" of Vlasov–Poisson equations. These relations have been the subject of a considerable amount of work in recent years, and we refer to [11, 7] and references therein for further details.

To put our study in the proper perspective we recall a few of the references from the considerable amount of literature available on the behavior of the Euler–Poisson and related problems. Let us mention the local existence in the small H^s-neighborhood of a steady state, e.g., [21, 26, 10]; the global existence of weak solutions with geometrical symmetry [6]; the two-carrier types in one dimension [32]; and the relaxation limit for the weak entropy solution (consult [24] for the isentropic case, and [16] for the isothermal case). Recently, the global existence of time-dependent sheaths with geometric symmetry was established in [14] by studying the Euler–Poisson system (1.1)–(1.3) with $c = e^{-\phi}$, the so-called Boltzmann relation.

For the question of global behavior of strong solutions, however, the choice of the initial data and/or damping forces is decisive. The nonexistence results in the case of attractive forces, $k < 0$, have been obtained by Makino and Perthame [23], and for repulsive forces by Perthame [27]. For research on the singularity formation in the model with diffusion and relaxation, consult [33]. In all these cases, the finite lifespan is due to a *global* condition of large enough initial (generalized) energy, staying outside a critical threshold ball. Using the characteristic-based method, Engelberg [8] gave local conditions for the finite-time loss of smoothness of solutions in Euler–Poisson equations. Global existence due to damping relaxation and with nonzero background can be found in [30, 31, 17]. For the model without damping relaxation the global existence was obtained by Guo [12], assuming that the flow is irrotational. His result applies to H_2-small neighborhoods of constant state.

When dealing with the questions of time regularity for Euler–Poisson equations without damping, one encounters several limitations of the classical stability analysis. Among other issues, we mention that

(i) stability analysis does not tell us how large perturbations can be before losing stability—indeed, the smallness of the initial perturbation is essential to making the energy method work (see, e.g., [12]);

(ii) the steady solution may be only conditionally stable due to the weak dissipation in the system, say, in the one-dimensional (1D) Euler–Poisson equations [9].

In order to address these difficulties, we advocated, in [9], a new notion of critical threshold (CT), which describes the conditional stability of the 1D Euler–Poisson equations, where the answer to the question of global vs. local existence depends on whether the initial configuration crosses an intrinsic $O(1)$ critical threshold. Little or no attention has been paid to this remarkable phenomenon, and our goal is to bridge the gap of previous studies on the behavior in Euler–Poisson solutions, a gap between the regularity of Euler–Poisson solutions "in the small" and their finite-time breakdown "in the large." The CT in the 1D Euler–Poisson system was completely characterized in terms of the relative size of the initial velocity slope and the initial density. Moving to the multidimensional setup, one has first to identify the proper quantities which govern the critical threshold phenomenon. In [19] we have shown that these quantities depend in an essential manner on the *eigenvalues* of the gradient velocity matrix, ∇u. In order to trace the evolution of $M := \nabla U$, we differentiate (1.4), obtaining formally

$$\partial_t M + U \cdot \nabla M + M^2 = -k(\nabla \otimes \nabla)\phi = kR[\rho - c], \tag{1.5}$$

where $R[\ ]$ is the 2×2 Risez matrix operator, defined as

$$R[f] =: \nabla\otimes\nabla\Delta^{-1}[f] = \mathcal{F}^{-1}\left\{\frac{\xi_j\xi_k}{|\xi|^2}\hat{f}(\xi)\right\}_{j,k=1,2}.$$

The above system is complemented by its coupling with the density ρ, which is governed by

$$\partial_t\rho + U\cdot\nabla\rho + \rho trM = 0. \tag{1.6}$$

Passing to the Lagrangian coordinates, that is, using the change of variables $\alpha\mapsto x(\alpha,t)$ with $x(\alpha,t)$ solving

$$\frac{dx}{dt} = U(x,t), \quad x(\alpha,0)=\alpha,$$

Euler–Poisson equations are recast into the coupled system

$$\frac{d}{dt}M + M^2 = kR[\rho - c], \tag{1.7}$$

$$\frac{d}{dt}\rho + \rho trM = 0, \tag{1.8}$$

with d/dt standing for the usual material derivative, $\partial_t + U\cdot\nabla$. It is the global forcing, $kR[\rho-c]$, which presents the main obstacle to studying the CT phenomenon of the multidimensional Euler–Poisson setting.

In this work we focus on the restricted Euler–Poisson (REP) system introduced in [19], which is obtained from (1.7) by restricting attention to the local isotropic trace, $\frac{k}{2}(\rho-c)I_{2\times 2}$, of the global coupling term $kR[\rho-c]$, namely,

$$\frac{d}{dt}M + M^2 = \frac{k}{2}(\rho-c)\cdot I_{2\times 2}, \tag{1.9}$$

$$\frac{d}{dt}\rho + \rho trM = 0. \tag{1.10}$$

We are concerned with the initial value REP problem (1.9), (1.10), subject to initial data

$$(M,\rho)(\cdot,0) = (M_0,\rho_0).$$

We note in passing that the REP system is to the full Euler–Poisson equations what the restricted Euler model is to the full Euler equations; consult [29, 4, 1, 2, 5, 19]. The existence of a critical threshold phenomenon associated with this 2D REP model with zero background, $c=0$, was first identified by us [19]. The current paper provides a precise description of the critical threshold for the 2D REP system (1.9), (1.10), with both zero and nonzero background charges. In particular, we use the so-called spectral dynamics lemma [19, Lemma 3.1] to obtain remarkable explicit formulae for the critical threshold surfaces summarized in the main Theorems 1.1 and 1.2 below.

To state our main results, we introduce two quantities with which we characterize the behavior of the velocity gradient tensor M. These are the trace, $d := trM$ (and we note that in case M coincides with ∇U, then d stands for the divergence, $d = u_x + v_y$), and the nonlinear quantity $\Gamma := (trM)^2 - 4detM$, which serves as an index for the *spectral gap*. Indeed, if $\lambda_i, i=1,2$, are the eigenvalues of M, then

$$\lambda_1 = \frac{1}{2}[d - \sqrt{\Gamma}], \qquad \lambda_2 = \frac{1}{2}[d+\sqrt{\Gamma}],$$

and hence Γ is nothing but the square of the spectral gap $\Gamma = (\lambda_2 - \lambda_1)^2$. We note that when M coincides with ∇U, then $\Gamma = (u_x - v_y)^2 + 4u_y v_x$, and the role of this spectral gap was first identified in the context of the 2D Eikonal equation in [19, Lemma 5.2].

We observe that if $\Gamma < 0$, then the spectral gap is purely imaginary. Otherwise, the spectral gap is real.

THEOREM 1.1 (2D REP with zero background). *Consider the 2D repulsive REP system* (1.9)–(1.10), *with $k > 0$ and with zero background $c = 0$. The solution of the 2D REP remains smooth for all time if and only if the initial data (ρ_0, M_0) lies in one of the following two regions, $(\rho_0, d_0, \Gamma_0) \in S_1 \cup S_2$:*

(i) $(\rho_0, d_0, \Gamma_0) \in S_1$,

$$S_1 := \left\{ (\rho, d, \Gamma) \;\middle|\; \Gamma \le 0 \quad \textit{and} \quad \left\{ \begin{array}{ll} d \ge 0 & \textit{if } \rho = 0, \\ d \textit{ arbitrary} & \textit{if } \rho > 0, \end{array} \right\} \right\}$$

(ii) $(\rho_0, d_0, \Gamma_0) \in S_2$,

$$S_2 := \left\{ (\rho, d, \Gamma) \;\middle|\; \rho > 0,\ \Gamma > 0, \quad \textit{and} \quad d \ge g(\rho, \Gamma) \right\},$$

where

$$g(\rho, \Gamma) := \operatorname{sgn}(\Gamma - 2k\rho) \sqrt{\Gamma - 2k\rho + 2k\rho \ln \left(\frac{2k\rho}{\Gamma} \right)}.$$

THEOREM 1.2 (2D REP with nonzero background). *Consider the 2D repulsive REP system* (1.9)–(1.10), *with $k > 0$ and with nonzero background $c > 0$. The solution of the 2D REP remains smooth for all time if and only if the initial data (ρ_0, M_0) lies in one of the following three regions, $(\rho_0, d_0, \Gamma_0) \in S_1 \cup S_2 \cup S_3$:*

(i) $(\rho_0, d_0, \Gamma_0) \in S_1$,

$$S_1 := \left\{ (\rho, d, \Gamma) \;\middle|\; \Gamma \le 0 \quad \textit{and} \quad \left\{ \begin{array}{ll} d \ge 0 & \textit{if } \rho = 0, \\ d \textit{ arbitrary} & \textit{if } \rho > 0, \end{array} \right\} \right\}$$

(ii) $(\rho_0, d_0, \Gamma_0) \in S_2$,

$$S_2 := \left\{ (\rho, d, \Gamma) \;\middle|\; 0 < \Gamma < \frac{k}{2c}\rho^2 \quad \textit{and} \quad \left\{ \begin{array}{lll} |d| \le g_1(\rho, \Gamma) & \textit{if} & \Gamma < 2k(\rho - c), \\ d \ge g_1(\rho, \Gamma) & \textit{if} & \Gamma \ge 2k(\rho - c), \end{array} \right\} \right\}$$

where

$$g_1(\rho, \Gamma) := \sqrt{\Gamma - 2k \left[c + \sqrt{\rho^2 - 2ck^{-1}\Gamma} + \rho \ln \left(\frac{\rho - \sqrt{\rho^2 - 2ck^{-1}\Gamma}}{2c} \right) \right]},$$

(iii) $(\rho_0, d_0, \Gamma_0) \in S_3$,

$$S_3 := \left\{ (\rho, d, \Gamma) \;\middle|\; \Gamma = \frac{k}{2c}\rho^2, \quad d = g_2(\rho, \Gamma), \quad \rho > 0 \right\},$$

where

$$g_2(\rho) = g_1(\rho, \Gamma)|_{\Gamma = \frac{k}{2c}\rho^2} := \sqrt{-2ck + \frac{k}{2c}\rho^2 + 2k\rho \ln \left(\frac{2c}{\rho} \right)}.$$

Several remarks are in order.

1. The above results show that the global smooth solution is ensured if the *initial* velocity gradient has complex eigenvalues, which applies, for example, for a class of initial configurations with sufficiently large vorticity $|u_{0y}-v_{0x}| \gg 1$. With other initial configurations, however, the finite-time breakdown of solutions may, and actually does, occur unless the initial divergence is above a critical threshold, expressed in terms of the initial density and initial spectral gap. Hence, global regularity depends on whether the initial configuration crosses an intrinsic $\mathcal{O}(1)$ critical threshold.

2. The critical threshold in the 1D Euler–Poisson equations depends on the relative size of the initial velocity slope and the initial density; consult [9]. In contrast to the 1D scenario, the critical threshold presented here depends on three initial quantities: density ρ_0, divergence $\nabla \cdot U_0$, and initial spectral gap $\Gamma_0 = (u_{0x} - v_{0y})^2 + 4u_{0y}v_{0x}$.

3. Theorem 1.1 tells us that the size of the initial subcritical range which gives rise to the regular solution is decreasing as the initial ratio Γ_0/ρ_0 is increasing. In particular, when this ratio is larger than $2k$, then the initial divergence must stay above a positive critical threshold to avoid the finite-time breakdown.

4. From Theorem 1.2 we see that the initial critical range which guarantees global regularity shrinks as the initial ratio Γ_0/ρ_0^2 is increasing in $(-\infty, \frac{k}{2c})$. Finite-time breakdown must occur when this ratio is larger than $\frac{k}{2c}$.

5. The limit $c \downarrow 0$ is a sort of *singular limit*, and hence one cannot recover Theorem 1.1 simply by passing to the limit $c \to 0$ in Theorem 1.2. □

It is well known that a finite-time breakdown is a generic phenomenon for nonlinear hyperbolic convection equations, which is realized by the formation of shock discontinuities. In the context of Euler–Poisson equations, however, there is a delicate balance between the forcing mechanism (governed by a Poisson equation) and the nonlinear focusing (governed by Newton's second law), which supports a critical threshold phenomenon.

In this paper we show how the persistence of the global features of the solutions for REP hinges on a delicate balance between the nonlinear convection and the localized forcing mechanism dictated by the Poisson equation. Here we use these restricted models to demonstrate the ubiquity of *critical thresholds* in the solutions of some of the equations of mathematical physics. This remarkable CT phenomenon has been found in other contexts, such as the scalar convolution model for nonlinear conservation laws [18], a nonlocal model in the nonlinear wave propagation [28], etc. Let us mention in particular the recent study [20], which shows, in the 2D case, how rotation enforces a CT phenomenon through which it prevents finite-time breakdown of nonlinear convection. Let us point out that the approach taken in this paper applies to the 3D case, leading to a closed 4×4 nonlinear system of ODEs governing the time-dynamics of the 3D REP. Identifying the CT phenomenon for such a system, however, is a formidable task which we hope to pursue in a future work.

In this paper we focus our attention on the restricted Euler–Poisson equations, "restricted" in the sense of using the same recipe for localized forcing as in the restricted Euler dynamics [29, 4, 1, 2, 5, 19]. We note in passing that the presence of global forcing in the full 2D Euler–Poisson equation, where $(\rho - c)I_{2\times 2}$ on the right-hand side of (1.9) is restored to the full $R[\rho - c]$ term, should allow for an additional stabilizing effect. We conjecture, therefore, that the full 2D Euler–Poisson equations admit a similar CT phenomenon, and in particular, that they admit global smooth solutions for subcritical initial data. As remarked earlier, the main obstacle in handling

this global case is the lack of an accurate description for the propagation of the Risez transform. Finally, one should not expect the current pressureless model to provide a faithful description of the model with pressure. The addition of a pressure term provides yet an additional mechanism for mixing between different particle paths.

We now conclude this section by outlining the rest of the paper. In section 2 we study the critical threshold for the REP with zero background. The key observation is that the spectral gap is conserved along the particle path. With this property we will be able to reduce the full dynamics on the 2D manifold parameterized by this initial spectral gap. In section 3 we discuss the critical threshold for the REP with nonzero background, where the CT arguments become considerably more involved. We treat the different cases which are indexed by the initial spectral gap.

2. 2D REP with zero background. In this section we prove the existence of the critical threshold of the 2D REP with zero background ($c = 0$)

$$\frac{d}{dt}M + M^2 = \frac{k}{2}\rho I_{2\times 2}, \tag{2.1}$$

$$\frac{d}{dt}\rho + \rho trM = 0. \tag{2.2}$$

This system with initial data (ρ_0, M_0) is well posed in the usual H^s Sobolev spaces for a short time. The global regularity follows from the standard boot-strap argument, once an a priori estimate on $\|M(\cdot)\|_{L^\infty}$ is obtained. First we show that, for the 2D REP (2.1)–(2.2), the velocity gradient tensor is completely controlled by the divergence d and the density ρ.

LEMMA 2.1. *Let M be the solution of the 2D REP; then the boundedness of M depends on the boundedness of trM and ρ; namely, there exists a constant, $Const = Const_T$, such that*

$$\|M(\cdot,t)\|_{L^\infty[0,T]} \le Const_T.\|(trM,\rho)\|_{L^\infty[0,T]}.$$

Proof. For the 2D case the velocity gradient tensor is completely governed by $p := M_{11} - M_{22}$, $q := M_{12} + M_{21}$, $\omega = M_{12} - M_{21}$, and $d = M_{11} + M_{22}$. From the M equation (2.1),

$$\frac{d}{dt}\begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix} + \begin{pmatrix} M_{11}^2 + M_{21}M_{12} & dM_{12} \\ dM_{21} & M_{21}M_{12} + M_{22}^2 \end{pmatrix} = \frac{k}{2}\rho I_{2\times 2},$$

one can obtain

$$\frac{d}{dt}p + pd = 0,$$
$$\frac{d}{dt}q + qd = 0,$$
$$\frac{d}{dt}\omega + \omega d = 0,$$

which, when combined with the mass equation

$$\frac{d}{dt}\rho + \rho d = 0,$$

gives

$$(p, q, \omega) = (p_0, q_0, \omega_0)\rho_0^{-1}\rho.$$

This shows that $|M_{ij}|_{L^\infty}$ are bounded in terms of $|d|_{L^\infty}$ and $|\rho|_{L^\infty}$ as asserted. ∎

This lemma tells us that to show the global regularity it suffices to control the divergence d and the density ρ. Let $\lambda_i, i = 1, 2$, be the eigenvalues of the velocity gradient tensor; then $d = \lambda_1 + \lambda_2$, and the continuity equation (1.10) reads

$$\frac{d}{dt}\rho + \rho(\lambda_1 + \lambda_2) = 0. \tag{2.3}$$

The spectral dynamics lemma [19, Lemma 3.1] tells us that the velocity gradient equation (1.9) yields

$$\frac{d}{dt}\lambda_1 + \lambda_1^2 = \frac{k}{2}\rho, \tag{2.4}$$

$$\frac{d}{dt}\lambda_2 + \lambda_2^2 = \frac{k}{2}\rho. \tag{2.5}$$

Following [19], we consider the difference of the last two equations, which gives for $\eta := \lambda_2 - \lambda_1$

$$\frac{d}{dt}\eta + \eta(\lambda_1 + \lambda_2) = 0.$$

This, combined with the mass equation (2.3) and $trM = \lambda_1 + \lambda_2$, yields

$$\frac{d}{dt}\left(\frac{\eta}{\rho}\right) = 0 \Rightarrow \quad \frac{\eta}{\rho} = \frac{\eta_0(\alpha)}{\rho_0(\alpha)}, \quad \alpha \in \mathbb{R}^2.$$

Set $\beta := \eta_0^2(\alpha)/\rho_0^2(\alpha)$ as a moving parameter with the initial position $\alpha \in \mathbb{R}^2$; one then obtains a closed system for ρ and d:

$$\rho' + \rho d = 0, \quad ' := \frac{d}{dt}, \tag{2.6}$$

$$d' + \frac{d^2 + \beta\rho^2}{2} = k\rho. \tag{2.7}$$

The first is the mass equation; the second is a restatement of summing (2.4), (2.5), $d' + (d^2 + \eta^2)/2 = k\rho$ with $\eta^2 = \beta\rho^2$.

We shall study the dynamics of (ρ, d) parameterized by β. It is easy to see that if the initial eigenvalues are complex, then the eigenvalues remain complex as time evolves. From

$$\beta = \frac{\Gamma_0}{\rho_0^2}, \qquad \Gamma_0 = (\lambda_2(0) - \lambda_1(0))^2,$$

we see that we need to distinguish between two cases, namely, $\beta < 0$, where the initial spectral gap is complex, and $\beta \geq 0$, where the initial spectral gap is real.

2.1. Complex spectral gap. We first study the case $\beta < 0$ when the initial eigenvalues are complex, i.e., $\mathrm{Im}(\lambda_i) \neq 0$.

LEMMA 2.2. *The solution of a 2D REP remains smooth for all time if eigenvalues are initially complex. Moreover, there is a global invariant given by*

$$\frac{d^2 - \beta\rho^2}{\rho} + 2k\ln\rho = Const. \tag{2.8}$$

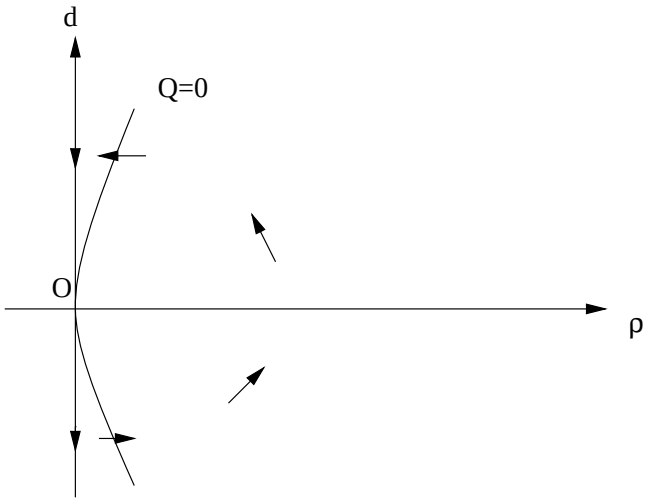


FIG. 2.1. *Zero level set $d' = 0$. Complex spectral gap.*

Proof. To obtain the desired global invariant we set $q := d^2$; then from (2.6)–(2.7) we deduce

$$\frac{dq}{d\rho} = 2d\frac{d'}{\rho'} = -2k + \beta\rho + \frac{q}{\rho}.$$

Integration gives

$$\frac{q}{\rho} - \beta\rho + 2k\ln\rho = Const,$$

which leads to (2.8). The boundedness of d follows at once, since for negative β's,

$$d^2 \le \max_{\rho>0}\left\{Const.\rho - 2k\ln\rho + \beta\rho^2\right\} =: C_1^2.$$

In particular, substitution of the lower bound $d \ge -C_1$ into the mass equation gives

$$\rho' \le C_1\rho,$$

which yields the desired upper bound for the density, $\rho(\cdot,t) \le \rho_0(\alpha)e^{C_1 t}$. □

Remark. More precise information about the large time behavior is available from phase plane analysis. According to (2.7), the zero level set $d' = 0$ is the hyperbola $Q := k\rho - (d^2 + \beta\rho^2)/2 = 0$, with a right branch passing the critical point $(0,0)$ and a left branch located in the left half-plane, $\rho < 0$; see Figure 2.1.

The trajectory on the plane $\rho < 0$ does not affect the solution behavior in the region $\rho > 0$ since $\rho = 0$ is an invariant set governed by

$$\rho \equiv 0, \qquad d' = -\frac{d^2}{2} \to d(t) = \frac{d_0}{1 + \frac{d_0}{2}t}.$$

Note that $(0,0)$ is the only critical point of the autonomous ODE system (2.6), (2.7) on the right half phase plane, and that the vector field in $\{(\rho,d),\ Q < 0, d \ge 0\}$ is

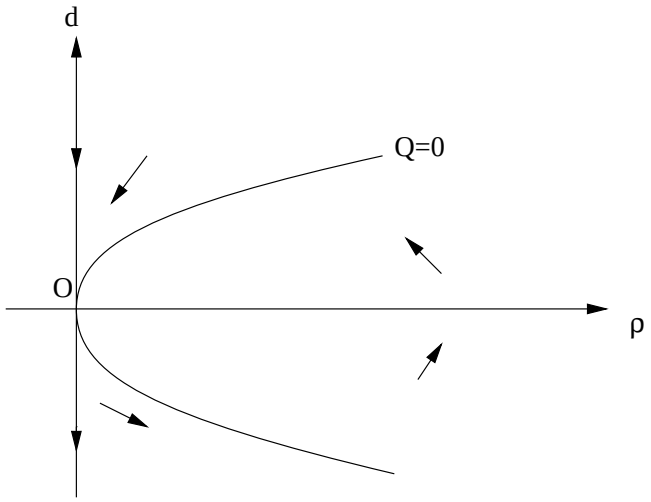


FIG. 2.2. *Zero level set $d' = 0$. Real spectral gap.*

converging to the critical point $(0, 0)$. It follows that for global smoothness it suffices to control the divergence d from below in the region $\{(\rho, d),\ Q < 0, d < 0, \rho > 0\}$ and to control the density from above in the region $\{(\rho, d), Q > 0\}$.

For the former case we have, recalling that $\beta < 0$,

$$\left(\frac{d}{\rho}\right)' = k + \frac{d^2}{2\rho} - \frac{\beta\rho}{2} \geq k,$$

and its integration along a particle path gives

$$d \geq \left(kt + \frac{d_0}{\rho_0}\right)\rho.$$

This shows that the divergence d is bounded from below, and, in particular, it becomes positive for large time since the density is positive. To the upper bound for ρ in the region $Q > 0$, where $d(t) \geq d_0(\alpha)$, we substitute this estimate into the mass equation, yielding

$$\rho' \leq -\rho d_0(\alpha).$$

This clearly gives the upper bound for the density $\rho \leq \rho_0(\alpha)e^{-d_0 t}$.

2.2. Real spectral gap. When $\beta \geq 0$, the initial spectral gap is real, and there are two cases to be considered, as follows.

Subcase 1. $\beta = 0$ when the eigenvalues are equal, i.e., $\lambda_1(0) = \lambda_2(0)$. In this case the zero level set $d' = 0$ becomes a parabola passing through the only critical point $(0, 0)$ (see Figure 2.2), and one can repeat arguments similar to our phase plane analysis in the previous case of distinct real roots. Note that the global invariant (2.8) becomes

$$\frac{d^2}{\rho} + 2k \ln \rho = Const.$$

Subcase 2. $\beta > 0$ when the eigenvalues are initially real.

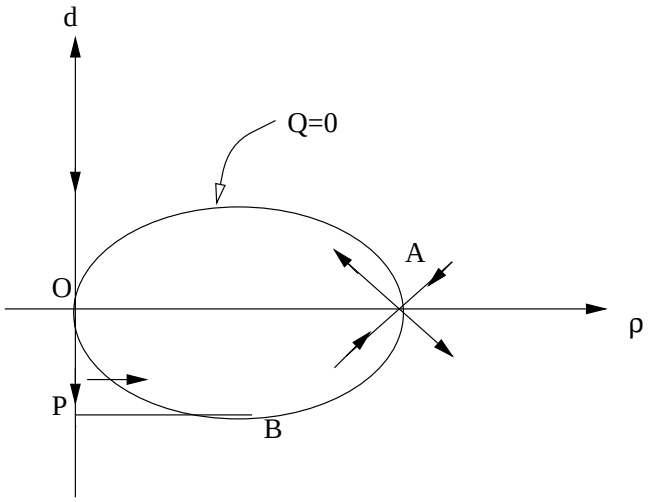


FIG. 2.3. *Critical points in the ρ-d plane. Real spectral gap.*

LEMMA 2.3. *If eigenvalues of ∇U_0 are real, then the solution of 2D REP remains smooth for all time if and only if*

$$\lambda_1(0) + \lambda_2(0) \geq g(\rho_0),$$

where

$$g(\rho) := \operatorname{sgn}\left(\rho - \frac{2k}{\beta}\right)\sqrt{\rho \times \left(F(\rho) - F\left(\frac{2k}{\beta}\right)\right)}, \quad F(\rho) = \beta\rho - 2k\ln\rho.$$

Proof. The system (2.6)–(2.7) has two critical points on the phase plane: $O(0,0)$ and $A(\frac{2k}{\beta}, 0)$; see Figure 2.3.

The coefficient matrix of the linearized system around (ρ^*, d^*) is

$$L(\rho^*, d^*) = \begin{pmatrix} -d^* & -\rho^* \\ k - \beta\rho^* & -d^* \end{pmatrix}.$$

A simple calculation gives the eigenvalues of L,

$$\lambda_\pm = -d^* \pm \sqrt{\rho^*(\beta\rho^* - k)}.$$

At $(0,0)$, we have $\lambda_1 = \lambda_2 = 0$, and hence $(0,0)$ is a nonhyperbolic critical point. Another critical point, $A(\frac{2k}{\beta}, 0)$, is a saddle since $\lambda_{1,2} = \pm\sqrt{\frac{2}{\beta}}k$. We shall use the above facts to construct the critical threshold via the phase plane analysis.

Assume that the seperatrix enters (leaves) A along the line $d = s(\rho - \frac{2k}{\beta})$. Upon substitution into the linearized system around A, i.e.,

$$\rho' = -\frac{2k}{\beta}d, \qquad d' = -k\left(\rho - \frac{2k}{\beta}\right),$$

one can obtain

$$s = \pm\sqrt{\frac{\beta}{2}}.$$

Thus, two seperatrices leave/enter A along the directions

$$\theta_1 = -\text{arctg}\sqrt{\frac{\beta}{2}} \quad \text{and} \quad \theta_2 = \text{arctg}\sqrt{\frac{\beta}{2}}.$$

In the phase plane the zero level set $d' = 0$ is an ellipse (see Figure 2.3),

$$d^2 + \beta\left(\rho - \frac{k}{\beta}\right)^2 = \frac{k^2}{\beta}.$$

Let $\gamma_s(A)$ be the portion of the stable manifold of the system coming into A from $\{d < 0\}$. In order to prove the existence of a critical threshold it suffices to show that $\gamma_s(A)$ can come only from O. Let B be the lowest point of the ellipse with coordinates $(\frac{k}{\beta}, -\frac{k}{\sqrt{\beta}})$, and let $\overline{PB}$ be a horizontal line intersecting with $\rho = 0$ at $(0, -\frac{k}{\sqrt{\beta}})$. According to the vector field inside the ellipse we see that the trajectory $\gamma_s(A)$ can come only from the area OPB by crossing the curve $\overline{OB}$. Note that the vector field on $\overline{PB}$ is going outside OPB and that $\rho = 0$ is invariant. Thus all trajectories in the area OPB originate from O. Therefore $\gamma_s(A)$ can originate only from O (as $t \to -\infty$) and becomes a portion of one unstable manifold of O. By symmetry we can show that the unstable manifold of the system issued from A entering $\{d > 0\}$ will end through the portion $\{d > 0\}$ at O.

Thus the critical curve $g : \mathbb{R}^+ \to \mathbb{R}$ is the one defined as

$$\{(\rho, d),\ d = g(\rho)\} = \gamma_s(A).$$

In order to have a precise formula for g we need to use the global invariant (2.8), i.e.,

$$\frac{d^2 - \beta\rho}{\rho} + 2k\ln\rho = Const.$$

Thus all trajectories can be expressed as

$$d^2 = \rho[C(\alpha) + F(\rho)],$$

with

$$C(\alpha) = \frac{d_0^2}{\rho_0} - F(\rho_0), \qquad F(\rho) := \beta\rho - 2k\ln\rho.$$

Note that $F(\rho)$ is a convex function and $\min_{\rho>0} F = F(\frac{2k}{\beta}) = 2k[1 - \ln(\frac{2k}{\beta})]$. Due to the symmetry, the homoclinic connection is possible when the trajectory passes $(\rho_0, 0)$ with $\rho_0 \le \frac{2k}{\beta}$ and converging to $(0,0)$ as $t \to \pm\infty$; i.e., the initial data must satisfy $0 < \rho_0 < \frac{2k}{\beta}$ and

$$C(\alpha) \le -F\left(\frac{2k}{\beta}\right), \quad \text{i.e.,} \quad \frac{d_0^2}{\rho_0} \le F(\rho_0) - F\left(\frac{2k}{\beta}\right).$$

The seperatrices passing through $(\frac{2k}{\beta}, 0)$ correspond to $C(\alpha) = -F(\frac{2k}{\beta})$. The stable manifold $\gamma_s(A)$ can be written as $d = g(\rho)$ for $0 \le \rho < \infty$, where

$$g(\rho) = \text{sgn}\left(\rho - \frac{2k}{\beta}\right)\sqrt{\rho\left(F(\rho) - F\left(\frac{2k}{\beta}\right)\right)}.$$

It remains to prove that the initial data satisfying $d_0 < g(\rho_0)$ always lead to finite-time breakdown.

First, in the region $\{(\rho, d),\ d < -\sqrt{\rho(F(\rho) - F(\frac{2k}{\beta}))}\}$, there must exist a finite time $T_1 > 0$ such that $\rho(T_1) > \frac{2k}{\beta}$ for $\rho_0 \le \frac{2k}{\beta}$ (take $T_1 = 0$ for $\rho_0 > \frac{2k}{\beta}$) since $\rho' > 0$. Therefore $\rho(t) \ge \rho(T_1)$ for $t \ge T_1$ and

$$\rho' = -\rho d \ge \rho\sqrt{\rho\left(\rho(T_1) - F\left(\frac{2k}{\beta}\right)\right)} \quad \text{for} \quad t \ge T_1.$$

Integration over $[T_1, t]$ gives

$$\sqrt{\rho(t)} \ge \frac{\sqrt{2\rho(T_1)}}{2 - (t - T_1)\sqrt{\rho(T_1)(F(\rho(T_1)) - F(\frac{2k}{\beta}))}}, \qquad t \ge T_1.$$

Thus the solution must become unbounded before the time

$$T_1 + \frac{2}{\sqrt{\rho(T_1)(F(\rho(T_1)) - F(\frac{2k}{\beta}))}}.$$

Second, we consider the trajectories in the region

$$\left\{(\rho, d), \quad \rho > \frac{2k}{\beta}, \quad |d| < \sqrt{\rho\left(F(\rho) - F\left(\frac{2k}{\beta}\right)\right)}\right\}.$$

Note that at finite time the trajectory must enter the subregion $\{(\rho, d),\ d < 0\}$ through the left point $(\rho^*, 0)$ identified as

$$d^2 = \rho[F(\rho) - F(\rho^*)], \qquad \rho \ge \rho^* > \frac{2k}{\beta}.$$

This, combined with the Riccati-type inequality

$$d' < \frac{-d^2}{2},$$

ensures the breakdown at finite time. This completes the confirmation of the curve $d = g(\rho)$ as a critical threshold. □

Proof of Theorem 1.1. It suffices to summarize the above cases, taking

$$\beta = \frac{\Gamma_0}{\rho_0^2}$$

into account. Clearly the cases $\beta < 0$ and $\beta = 0$ correspond to the set

$$\{(\rho_0, M_0), \quad \Gamma_0 \le 0\}.$$

For $\beta > 0$, i.e., $\Gamma_0 > 0$, we rewrite the critical threshold as

$$\begin{aligned} d_0 &= \operatorname{sgn}\left(\rho_0 - \frac{2k}{\beta}\right)\sqrt{\rho_0\left(F(\rho_0) - F\left(\frac{2k}{\beta}\right)\right)} \\ &= \operatorname{sgn}(\Gamma_0 - 2k\rho_0)\sqrt{\Gamma_0 - 2k\rho_0 + 2k\rho_0 \ln\left(\frac{2k\rho_0}{\Gamma_0}\right)}, \end{aligned}$$

where we have used the relation $F(\rho) = \beta\rho - 2k\ln\rho$ and

$$F(\rho_0) = \frac{\Gamma_0}{\rho_0} - 2k\ln\rho_0,$$
$$F\left(\frac{2k}{\beta}\right) = 2k - 2k\ln\left(\frac{2k\rho_0^2}{\Gamma_0}\right).$$

This completes the proof of Theorem 1.1. □

3. 2D REP with nonzero background. This section is devoted to the study of the REP with nonzero background $c > 0$, for which the velocity gradient tensor $M = \nabla U$ solves

$$\frac{d}{dt}M + M^2 = \frac{k}{2}[\rho - c], \tag{3.1}$$

$$\frac{d}{dt}\rho + \rho tr M = 0. \tag{3.2}$$

Again, using the spectral dynamics lemma presented in [19], the spectral dynamics of M is governed by

$$\lambda_1' + \lambda_1^2 = \frac{k}{2}(\rho - c), \quad ' := \frac{d}{dt},$$
$$\lambda_2' + \lambda_2^2 = \frac{k}{2}(\rho - c),$$
$$\rho' + \rho(\lambda_1 + \lambda_2) = 0.$$

As in the zero background case, the difference $\eta := \lambda_2 - \lambda_1$ is proportional to the density along the particle path in the sense that

$$\frac{\eta(t)}{\rho(t)} = \frac{\eta_0(\alpha)}{\rho_0(\alpha)}, \quad \alpha \in \mathbb{R}^2.$$

Further manipulation gives a closed system

$$\rho' = -\rho d, \tag{3.3}$$

$$d' = k(\rho - c) - \frac{d^2 + \beta\rho^2}{2} =: Q, \quad \beta := \frac{\eta_0^2}{\rho_0^2}. \tag{3.4}$$

Once again the dynamics of (3.3), (3.4) is influenced by the choice of β. We proceed to discuss the solution behavior of (3.3), (3.4) by distinguishing two cases:

(1) for $\beta < 0$, the spectral gap is complex;

(2) for $\beta \geq 0$, the spectral gap is real.

3.1. Complex spectral gap. We first discuss the case $\beta < 0$, which corresponds to the case in which the eigenvalues are initially complex.

LEMMA 3.1. *Assume that the eigenvalues are initially complex with* $\mathrm{Im}(\lambda_i(0)) \neq 0$. *Then the solution of* (3.3), (3.4) *remains smooth for all time. Moreover, there is a global invariant in time, given by*

$$V(\rho, d) = \rho^{-1}\left[d^2 - \beta\rho^2 + 2k\rho ln\left(\frac{\rho}{2c}\right) + 2ck\right].$$

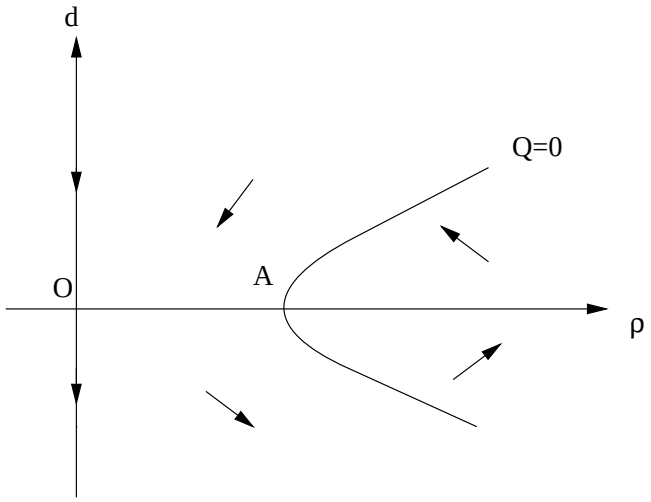


FIG. 3.1. *Zero level set $d' = 0$. Complex spectral gap with nonzero background.*

Proof. A straightforward computation yields $\dot{V} = 0$ along the 2D REP solutions, which implies that the curves $V = Const$ are invariants of the flow. As before, for negative β's we have

$$d^2 \le \max_{\rho>0} \left\{ Const.\rho - 2k\rho \ln\left(\frac{\rho}{2c}\right) - 2ck + \beta\rho^2 \right\} \le C_1^2,$$

and the bounds of d (and hence of ρ) follow. □

Remark. As before, more detailed information is available in this case by a phase plane analysis. If the eigenvalues are initially complex, then one has $\beta = \eta_0^2/\rho_0^2 < 0$. The zero level set $d' = Q = 0$ becomes a hyperbola; see Figure 3.1.

The intersection of its right branch with $d = 0$ is the rest point $A = (\rho^*, 0)$ of the system, where

$$\rho^* = \frac{k}{\beta} + \sqrt{\frac{k^2}{\beta^2} - \frac{2ck}{\beta}}.$$

The coefficient matrix of the linearization around $(\rho^*, 0)$ is

$$L(\rho^*, 0) = \begin{pmatrix} 0 & -\rho^* \\ k - \beta\rho^* & 0 \end{pmatrix}.$$

Its eigenvalues satisfy

$$\lambda^2 = \rho^*(\beta\rho^* - k) = -\rho^*\sqrt{k^2 - 2ck\beta} < 0.$$

Hence such a critical point is a nonhyperbolic equilibrium. The nonlinear effect plays essential roles in the solution behavior. In order to locate the possible critical threshold, we first study the solution around $(\rho^*, 0)$. Setting $n = \rho - \rho^*$, we then have

$$n' = -\rho^* d - nd, \tag{3.5}$$

$$d' = \sqrt{k^2 - 2ck\beta}\, n - \frac{d^2}{2} - \frac{\beta}{2} n^2. \tag{3.6}$$

It is easy to see that the flow governed by the linear part stays on the ellipse

$$\sqrt{k^2 - 2ck\beta}n^2 + \rho^* d^2 = Const.$$

In order to capture the dynamics of the nonlinear system in the neighborhood of the critical point $(n, d) = (0, 0)$, we employ the polar coordinates of the form

$$n = \frac{r\cos\theta}{(k^2 - 2ck\beta)^{1/4}},$$
$$d = \frac{-r\sin\theta}{\sqrt{\rho^*}}.$$

Careful calculation with these polar coordinates yields that (3.5)–(3.6) can be recast into the form

$$r' = R(r, \theta), \tag{3.7}$$
$$\theta' = -\sqrt{\rho^*}(k^2 - 2ck\beta)^{1/4} + \Theta(r, \theta), \tag{3.8}$$

where

$$R(r, \theta) = \frac{r^2\sin\theta}{2\sqrt{\rho^*}}\left[1 + \frac{k\cos^2\theta}{\sqrt{k^2 - 2ck\beta}}\right],$$
$$\Theta(r, \theta) = -\frac{r\cos\theta}{2\sqrt{\rho^*}\sqrt{k^2 - 2ck\beta}}\left[\sqrt{k^2 - 2ck\beta}\sin^2\theta - \beta\rho^*\cos^2\theta\right].$$

When r is sufficiently small, θ' is strictly negative. The pleasant implication of this is that the orbits of system (3.5), (3.6) spiral monotonically in θ around $(\rho^*, 0)$. But the even power of r^2 does not indicate the stability property of the critical point.

Observe that if $(n(t), d(t))$ is a solution, so is $(n(-t), -d(-t))$. Such symmetry implies that there is a center in the neighborhood of $(\rho^*, 0)$.

In order to clarify the global behavior of the flow around such a center, we appeal to the global invariant

$$V(\rho, d) = \rho^{-1}\left[d^2 - \beta\rho^2 + 2k\rho\ln\left(\frac{\rho}{2c}\right) + 2ck\right].$$

We claim that V is positive definite, which serves as a (majorization of) Lyapunov functional. To this end, we consider the function $H(\rho) := -\beta\rho^2 + 2k\rho\ln\left(\frac{\rho}{2c}\right) + 2ck$, which is convex and takes its minimum at $\rho_{\min}$, satisfying

$$\ln\left(\frac{\rho_{\min}}{2c}\right) = -1 + \frac{\beta}{k}\rho_{\min}.$$

Observe that, since $\beta < 0$, the function

$$h(\rho) := 1 - \frac{\beta}{k}\rho + \ln\left(\frac{\rho}{2c}\right)$$

is an increasing function in $\rho > 0$ and $h(\rho_{\min}) = 0$, which, when combined with the fact that $h(\rho^*) \geq 0$, verifies that

$$0 < \rho_{\min} \leq \rho^*, \qquad \rho^* := \beta^{-1}[k - \sqrt{k^2 - 2ck\beta}].$$

Indeed, for ρ^* we have

$$h(\rho^*) = 1 - \frac{\beta}{k}\rho^* + \ln\left(\frac{\rho^*}{2c}\right) = \sqrt{1-2ck^{-1}\beta} - \ln(1+\sqrt{1-2ck^{-1}\beta}) \geq 0.$$

Therefore $H(\rho)$ is nonnegative since

$$H(\rho_{\min}) = \beta\rho_{\min}^2 - 2k\rho_{\min} + 2ck = \beta(\rho_{\min} - \rho^*)(\rho_{\min} - \bar{\rho}^*) \geq 0,$$

where $\bar{\rho}^* = \beta^{-1}[k + \sqrt{k^2 - 2ck\beta}]$.

The invariant curves, $V = Const.$, represent, of course, the bounded periodic orbits containing $(\rho^*, 0)$.

3.2. Real spectral gap. We divide the region $\beta \in [0, \infty)$ into subregions depending on the number of critical points on the phase plane, and then study the solution behavior with β in each subregion. The solution behavior depends strongly on the number of critical points and their stability property.

Let (ρ^*, d^*) be a critical point of the system; then the coefficient matrix of the linearization around (ρ^*, d^*) reads

$$L(\rho^*, d^*) = \begin{pmatrix} -d^* & -\rho^* \\ k - \beta\rho^* & -d^* \end{pmatrix}.$$

Its eigenvalues are given by

$$\lambda = -d^* \pm \sqrt{\rho^*(\beta\rho^* - k)}. \tag{3.9}$$

We now discuss subcases distinguished by the number and type of critical points as β changes.

- $\beta = 0$. Here the zero level set $d' = Q = 0$ is a parabola, $d^2 = 2k(\rho - c)$, intersecting with $d = 0$ at $(\rho^*, d^*) = (c, 0)$. From (3.9) we see that at this point the eigenvalues of L are $\lambda = \pm\sqrt{ck}i$, a pure imaginary number, and the critical point $(c, 0)$ is nonhyperbolic. The stability property of this critical point has to be determined by taking into account the nonlinear effect.
- $0 < \beta < \frac{k}{2c}$. The zero level set $Q = 0$ is an ellipse, located on the right half-plane $\rho > 0$. There are two critical points $(\rho^*, d^*) = (\rho^*, 0)$ with

  $$\rho^* = \frac{k}{\beta} \pm \sqrt{\frac{k^2}{\beta^2} - \frac{2kc}{\beta}}.$$

  The associated eigenvalues of L are

  $$\lambda(\rho_1^*) = \pm\sqrt{\rho_1^*\sqrt{k^2 - 2ck\beta}}i, \qquad \lambda(\rho_2^*) = \pm\sqrt{\rho_2^*\sqrt{k^2 - 2ck\beta}}.$$

  Therefore $(\rho_1^*, 0)$ is a center of the linearized system, and $(\rho_2^*, 0)$ is a saddle; see Figure 3.2.

  Possible bifurcation as β changes from 0 to $\frac{k}{2c}$ may be responsible for the complicated solution structure in this regime.
- $\beta = \frac{k}{2c}$. The zero level set $Q = 0$, i.e.,

  $$d^2 + \beta\left(\rho - \frac{k}{\beta}\right)^2 = 0,$$

  degenerates to a single point $(\rho^*, d^*) = (\frac{k}{\beta}, 0)$, the only critical point with zero eigenvalues.

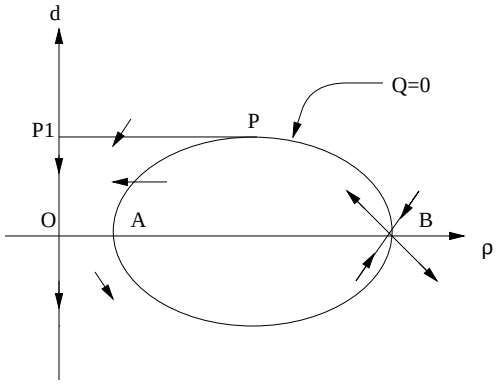


FIG. 3.2. *Critical points in the ρ-d plane. Real spectral gap with nonzero background.*

- $\beta > \frac{k}{2c}$. In this case

$$Q = -\frac{1}{2}\left[d^2 + \beta\left(\rho - \frac{k}{\beta}\right)^2 + 2kc - \frac{k^2}{\beta}\right] \leq \frac{k^2}{2\beta} - kc < 0.$$

There is no critical point at all in the finite phase plane.

The solution behavior distinguished by the above cases is given in the following lemmata.

LEMMA 3.2. *If $\lambda_1(0) = \lambda_2(0)$, then the solution of* (3.3), (3.4) *remains smooth for all time, indicated by the global invariant*

$$\frac{d^2 + 2ck}{\rho} + 2k\ln\rho = Const. \tag{3.10}$$

Proof. The assumption amounts to $\beta = 0$. As discussed above, $(c, 0)$ is the only critical point and the center of the linearized system. In order to find the global invariant we set $R(t) := k(\rho - c)^2 + cd^2$. Along the trajectory $\frac{dx}{dt} = U(x,t)$,

$$\frac{d}{dt}R(t) = 2k(\rho - c)\rho' + 2cdd' = -d[k(\rho - c)^2 + R(t)]. \tag{3.11}$$

From the mass equation it follows that

$$d = -\frac{\rho'}{\rho},$$

which, when inserted into the relation (3.11), gives

$$\frac{dR}{d\rho} = \frac{k(\rho - c)^2}{\rho} + \frac{R}{\rho}.$$

Integration gives

$$\frac{R}{\rho} + \frac{kc^2}{\rho} + 2ck\ln\rho - k\rho = Const,$$

which leads us to the global invariant as asserted in (3.10). This global invariant is compact and ensures that both divergence d and the density ρ remain bounded as time evolves. ∎

We leave the case $0 < \beta < \frac{k}{2c}$ for later and study the critical case $\beta = \frac{k}{2c}$.

LEMMA 3.3. *If $\lambda_2(0) - \lambda_1(0) = \sqrt{\frac{k}{2c}}\rho_0$, then the solution of* (3.3), (3.4) *always develops finite-time breakdown unless the initial data lies within the set*

$$\left\{(\rho, d) \in \mathbb{R}^+ \times \mathbb{R}^+, \quad \frac{d^2 + 2ck}{\rho} - \frac{k}{2c}\rho + 2k\ln\rho = 2k\ln(2c)\right\}.$$

Proof. The given assumption is equivalent to the case $\beta = \frac{k}{2c}$. In this case the divergence always decreases except at the critical point $(2c, 0)$ since

$$d' = -\frac{d^2}{2} - \frac{k(\rho - 2c)^2}{4c} \le 0.$$

In order to clarify the solution behavior, we proceed to obtain the global invariant. Setting $q := d^2$, one then has

$$\frac{dq}{d\rho} = \frac{2dd'}{\rho'} = \frac{q + \beta\rho^2 - 2k(\rho - c)}{\rho}.$$

Solving this equation, we obtain

$$\frac{q}{\rho} = \beta\rho - 2k\ln\rho - \frac{2ck}{\rho} + Const.$$

Therefore we come up with a global invariant

$$\frac{d^2 + 2ck}{\rho} - \beta\rho + 2k\ln\rho = Const. \tag{3.12}$$

The only trajectory converging to the critical point is realized by a half-trajectory converging to $(2c, 0)$ from the first quadrant. For all other trajectories not passing the critical point $(2c, 0)$, the rate d' is strictly negative. The divergence will become negative at finite time even if it is initially positive, which, when combined with the Riccati-type inequality $d' \le -d^2/2$, confirms the finite-time breakdown. ∎

We now look at the case $\beta > \frac{k}{2c}$.

LEMMA 3.4. *Assume that the eigenvalues are initially real and $|\lambda_2(0) - \lambda_1(0)| > \sqrt{\frac{k}{2c}}\rho_0(\alpha)$. Then the solution of* (3.3), (3.4) *always develops finite-time breakdown.*

Proof. The given assumption is nothing but the inequality $\beta > \frac{k}{2c}$. Note that there is no critical point in the finite phase plane; actually Q remains negative for all time. The solution must develop breakdown in finite time. In fact from

$$d' = -\frac{d^2}{2} - \frac{\beta}{2}\left(\rho - \frac{k}{\beta}\right)^2 - \frac{kc}{\beta}\left(\beta - \frac{k}{2c}\right), \tag{3.13}$$

we find that

$$d' \le -\delta \quad \text{with} \quad \delta := \frac{kc}{\beta}\left(\beta - \frac{k}{2c}\right) > 0.$$

This ensures that d must become negative beyond a finite time T_0, say, $T_0 > \max\{\frac{d_0}{\delta}, 0\}$. The $d-$ equation (3.13) also gives

$$d' \le -\frac{d^2}{2},$$

whose integration over $[T_0, t]$ leads to

$$d(t) \le \frac{d(T_0)}{1 - \frac{1}{2}d(T_0)(t - T_0)}.$$

Hence the solution must break down at a finite time before $T_0 - \frac{2}{d(T_0)}$. □

Finally we conclude this subsection by discussing the delicate case $0 < \beta < \frac{k}{2c}$. Set

$$G(\rho, \rho^*, \beta) := \beta(\rho - \rho^*) - 2k \ln\left(\frac{\rho}{\rho^*}\right) - \frac{2ck}{\rho} + \frac{2ck}{\rho^*},$$

with $\rho^* = \beta^{-1}[k \pm \sqrt{k^2 - 2ck\beta}]$ being the ρ-coordinate of the intersection point of the trajectory with the ρ-axis.

LEMMA 3.5. *Assume that the real eigenvalues satisfy* $0 < |\lambda_2(0) - \lambda_1(0)| < \sqrt{\frac{k}{2c}}\rho_0(\alpha)$. *Then for any* $\beta \in (0, \frac{k}{2c})$ *the solutions of* (3.3), (3.4) *remain smooth for all time if and only if*

$$|\lambda_1(0) + \lambda_2(0)| \le \sqrt{\rho_0 G(\rho_0, \rho_2^*, \beta_0)} \quad \textit{for} \quad \rho_0 \le \rho_2^*$$

and

$$\lambda_1(0) + \lambda_2(0) = \sqrt{\rho_0 G(\rho_0, \rho_2^*, \beta_0)} \quad \textit{for} \quad \rho_0 > \rho_2^*.$$

Proof. The assumption tells us that $\beta < \frac{k}{2c}$. In this case there are two critical points in the phase plane, $A = (\rho_1^*, 0)$ and $B = (\rho_2^*, 0)$; see Figure 3.2. B is a saddle whose two manifolds pass, enclosing the critical point A, which is a center of the linearized system. Let $W_s(B)$ denote the stable manifold coming from the region $\{(\rho, d), \rho < \rho_2^*, d < 0\}$, and $W_u(B)$ the unstable manifold entering into $\{(\rho, d), \rho < \rho_2^*, d > 0\}$. To prove the results stated in the theorem it suffices to show that for any $\beta \in (0, \frac{k}{2c})$ such that $W_u(B) \cap W_s(B)$ is not empty, there exists a saddle connection (homoclinic orbit).

This follows from the continuity argument supported by the following facts:

(1) Both $W_u(B)$ and $W_s(B)$ pass through the segment OA, with flow going downward since $d' < 0$ and $\rho' = 0$ on OA; see Figure 3.2

The level curve $d' = 0$ is an ellipse with upper vortex P located at $(\frac{k}{\beta}, 0)$. Let P_1 denote the intersection of the tangent line of the ellipse $Q = 0$ through P with the axis $\rho = 0$. The vector field inside the ellipse shows that $W_u(B)$ must escape the ellipse from the curve $\overline{PA}$. Note that the trajectories on $\overline{PP_1}$ and $\overline{AP}$ enter into the region $PAOP_1$, and the axis $\rho = 0$ is an invariant set. These facts ensure that $W_u(B)$ must enter the region $d < 0$ through OA. Similarly we can show that $W_s(B)$ for $\rho \le \rho_2^*$ must enter the region $d > 0$ through OA.

(2) As β increases in $(0, \frac{k}{2c})$, the point $W_u(B) \cap OA$ moves to the right, and the point $W_s(B) \cap OA$ moves to the left.

We prove the claim for $W_u(B) \cap OA$, and the case for $W_s(B) \cap OA$ follows similarly. The claim follows from the following two observations:

(i) The slope of the unstable manifold $W_u(B, \beta)$ at $(\rho_2^*, 0)$ is $\partial_\rho d|_{\rho=\rho_2^*} = \lambda_-(\rho_2^*, \beta)$, and the eigenvalue $\lambda_-(\rho_2^*, \beta)$ is increasing in β. Indeed,

$$\frac{d\lambda_-(\rho_2^*, \beta)}{d\beta} = -\frac{\beta^2}{k\lambda_-(\rho_2^*, \beta)} \left\{ k + \sqrt{k^2 - 2ck\beta} + \frac{ck\beta}{\sqrt{k^2 - 2ck\beta}} \right\} > 0.$$

(ii) $W_u(B,\beta_1)$ does not intersect with $W_u(B,\beta_2)$ for $\beta_1 \neq \beta_2$. As previously, we can find the global invariant of the system

$$d^2 = \rho\left[\beta\rho - 2k\ln\rho - \frac{2ck}{\rho} + Const\right],$$

from which the left branch of the unstable manifold of B can be explicitly expressed as

$$d = \sqrt{\rho\left[\beta(\rho-\rho_2^*) - 2k\ln\left(\frac{\rho}{\rho_2^*}\right) - \frac{2ck}{\rho} + \frac{2ck}{\rho_2^*}\right]}, \qquad 0<\rho<\rho_2^*.$$

A careful calculation gives

$$\frac{\partial d}{\partial \beta} = \frac{\rho}{2d}(\rho-\rho_2^*) < 0,$$

which ensures the claim (ii).

(3) Let $\rho_u(\beta)$ be the ρ-coordinate of the point $W_u(B,\beta)\cap OA$, and $\rho_s(\beta)$ be the ρ-coordinate of the point $W_s(B,\beta)\cap OA$. We claim

$$\lim_{\beta\to 0+}\rho_u(\beta) < \lim_{\beta\to\frac{k}{2c}-}\rho_s(\beta).$$

In fact, from the expression of seperatrices

$$d^2 = \rho G(\rho,\rho_2^*,\beta), \quad \rho<\rho^*,$$

we find that the ρ-coordinates of points $W_{u/s}(B)\cap OA$ satisfy

$$G(\rho,\rho_2^*,\beta)\equiv 0.$$

Note that

$$\frac{\partial G}{\partial\rho} = \frac{\beta}{\rho^2}(\rho-\rho_1^*)(\rho-\rho_2^*), \qquad \frac{\partial G}{\partial\beta} = \rho-\rho_2^*.$$

Thus we have for $0<\rho<\rho_1^*$

$$\frac{\partial\rho}{\partial\beta} = -\frac{\frac{\partial G}{\partial\beta}}{\frac{\partial G}{\partial\rho}} = \frac{\rho^2}{\beta}(\rho_1^*-\rho) > 0.$$

This confirms the above assertion.

Combining the above observations, we conclude that there exists a $\beta_0\in(0,\frac{k}{2c})$ for which a saddle connection exists. It remains to show that, as β changes in the region $(0,\frac{k}{2c})$, the above saddle connection is preserved. Observe that if $(\rho(t),d(t))$ is a solution, so is $(\rho(-t),-d(-t))$. Such symmetry prevents the occurrence of the possible bifurcation when β changes.

Using the nonlinear terms in the equation and the vector field, we can show for the initial data outside the closed curve—saddle connection—that the solution always develops finite-time breakdown; details are omitted. □

Proof of Theorem 1.2. Summarizing the results stated in the above lemmata, we see that the case $\beta<0$ and $\beta=0$ corresponds to the set

$$S_1 = \left\{(\rho_0,d_0,\Gamma_0)\;\middle|\;\Gamma_0\le 0 \quad\text{and}\quad \left\{\begin{array}{ll} d_0\ge 0 & \text{if } \rho_0=0,\\ d_0 \text{ arbitrary} & \text{if } \rho_0>0,\end{array}\right\}\right\}$$

since $\Gamma_0 = \beta\rho_0^2$. The case $0 < \beta < \frac{k}{2c}$ corresponds to $0 < \Gamma_0 < \frac{k}{2c}\rho_0^2$, and the divergence is required to satisfy the critical threshold condition

$$|d_0| \le \sqrt{\rho_0 G(\rho_0, \rho_2^*, \beta)}, \quad 0 < \rho_0 < \rho_2^*,$$

and $d_0 = \sqrt{\rho_0 G(\rho_0, \rho_2^*, \beta)}$ for $\rho_0 \ge \rho_2^*$. Using $\Gamma_0 = \beta\rho_0^2$ and

$$\rho_2^* = \beta^{-1}[k + \sqrt{k^2 - 2ck\beta}] = \frac{2ck}{k^2 - \sqrt{k^2 - 2ck\beta}} = \frac{2c\rho_0}{\rho_0 - \sqrt{\rho_0^2 - \frac{2c}{k}\Gamma_0}},$$

one has

$$\begin{aligned}
\rho_0 G(\rho_0, \rho_2^*, \beta) &= \rho_0 \left[\beta(\rho_0 - \rho_2^*) - 2k \ln\left(\frac{\rho_0}{\rho_2^*}\right) - \frac{2ck}{\rho_0} + \frac{2ck}{\rho_2^*}\right] \\
&= \Gamma_0 \left(1 - \frac{2c}{\rho_0 - \sqrt{\rho_0^2 - \frac{2c\Gamma_0}{k}}}\right) - 2k\rho_0 \ln\left(\frac{\rho_0 - \sqrt{\rho_0^2 - \frac{2c\Gamma_0}{k}}}{2c}\right) \\
&\quad - 2ck + k\left(\rho_0 - \sqrt{\rho_0^2 - \frac{2c\Gamma_0}{k}}\right) \\
&= \Gamma_0 - 2ck - 2k\sqrt{\rho_0^2 - \frac{2c\Gamma_0}{k}} - 2k\rho_0 \ln\left(\frac{\rho_0 - \sqrt{\rho_0^2 - \frac{2c\Gamma_0}{k}}}{2c}\right),
\end{aligned}$$

which leads to the critical threshold described by the set S_2. The set S_3 can be determined in a similar manner. □

REFERENCES

[1] O. Boratav and R. Pelz, *Locally isotropic pressure Hessian in a high-symmetry flow*, Phys. Fluids, 7 (1995), pp. 895–897.

[2] O. Boratav and R. Pelz, *On the local topology evolution of a high-symmetry flow*, Phys. Fluids, 7 (1995), pp. 1712–1731.

[3] U. Brauer, A. Rendal, and O. Reula, *The cosmic no-hair theorem and the non-linear stability of homogeneous Newtonian cosmological models*, Class. Quantum Grav., 11 (1994), pp. 2283–2296.

[4] B.J. Cantwell, *Exact solution of a restricted Euler equation for the velocity gradient tensor*, Phys. Fluids A, 4 (1992), pp. 782–793.

[5] M. Chertkov, A. Pumir, and B. Shraiman, *Lagrangian tetrad dynamics and phenomenology of turbulence*, Phys. Fluids A, 11 (1999), pp. 2394–2410.

[6] G.-Q. Chen and D. Wang, *Convergence of shock capturing scheme for the compressible Euler-Poisson equations*, Comm. Math. Phys., 179 (1996), pp. 333–364.

[7] J. Dolbeault and G. Rein, *Time-dependent rescaling and Lyapunov functionals for the Vlasov-Poisson and Euler-Poisson systems, and for related models of kinetic equations, fluid dynamics and quantum physics*, Math. Models Methods Appl. Sci., 11 (2001), pp. 407–432.

[8] S. Engelberg, *Formation of singularities in the Euler-Poisson equations*, Phys. D, 98 (1996), pp. 67–74.

[9] S. Engelberg, H. Liu, and E. Tadmor, *Critical thresholds in Euler-Poisson equations*, Indiana Univ. Math. J., 50 (2001), pp. 109–157.

[10] P. Gamblin, *Solution régulière à temps petit pour l'équation d'Euler-Poisson*, Comm. Partial Differential Equations, 18 (1993), pp. 731–745.

[11] I. Gasser, C.-K Lin, and P.A. Markowich, *A review of dispersive limits of (non)linear Schrödinger-type equations*, Taiwanese J. Math., 4 (2000), pp. 501–529.

[12] Y. Guo, *Smooth irrotational flows in the large to the Euler-Poisson system in* $\mathbb{R}^{3+1}$, Comm. Math. Phys., 195 (1998), pp. 249–265.
[13] D. Holm, S.F. Johnson, and K.E. Lonngren, *Expansion of a cold ion cloud*, Appl. Phys. Lett., 38 (1981), pp. 519–521.
[14] S.-Y. Ha and M. Slemrod, *Global existence of plasma ion-sheath and their dynamics*, Comm. Math. Phys., 238 (2003), pp. 149–186.
[15] J.D. Jackson, *Classical Electrodynamics*, 2nd ed., Wiley, New York, 1975.
[16] S. Junca and M. Rascle, *Relaxation of the isothermal Euler-Poisson system to the drift-diffusion equations*, Quart. Appl. Math., 58 (2000), pp. 511–521.
[17] T. Luo, R. Natalini, and Z. Xin, *Large time behavior of the solutions to a hydrodynamic model for semiconductors*, SIAM J. Appl. Math., 59 (1999), pp. 810–830.
[18] H. Liu and E. Tadmor, *Critical thresholds in a convolution model for nonlinear conservation laws*, SIAM J. Math. Anal., 33 (2001), pp. 930–945.
[19] H. Liu and E. Tadmor, *Spectral dynamics of velocity gradient field in restricted flows*, Comm. Math. Phys., 228 (2002), pp. 435–466.
[20] H. Liu and E. Tadmor, *Rotation prevents finite time breakdown*, Phys. D, to appear.
[21] T. Makino, *On a local existence theorem for the evolution of gaseous stars*, in Patterns and Waves, T. Nishida, M. Mimura, and H. Fujii, eds., North-Holland/Kinokuniya, 1986, pp. 459–479.
[22] P.A. Markowich, *A non-isentropic Euler-Poisson model for a collisionless plasma*, Math. Methods Appl. Sci., 16 (1993), pp. 409–442.
[23] T. Makino, and B. Perthame, *Sur les solutions a symetrie spherique de l'equation d'Euler-Poisson pour l'etoiles gazeuses*, Japan J. Appl. Math., 7 (1990), pp. 165–170.
[24] P. Marcati and R. Natalini, *Weak solutions to a hydrodynamic model for semiconductors and relaxation to the drift-diffusion equation*, Arch. Ration. Mech. Anal., 129 (1995), pp. 129–145.
[25] P.A. Markowich, C. Ringhofer, and C. Schmeiser, *Semiconductor Equations*, Springer-Verlag, Berlin, Heidelberg, New York, 1990.
[26] T. Makino and S. Ukai, *Sur l'existence des solutions locales de l'équation d'Euler-Poisson pour l'évolution d'étoiles gazeuses*, J. Math. Kyoto Univ., 27 (1987), pp. 387–399.
[27] B. Perthame, *Nonexistence of global solutions to the Euler-Poisson equations for repulsive forces*, Japan J. Appl. Math., 7 (1990), pp. 363–367.
[28] J.B. Thoo and J.K. Hunter, *Nonlinear wave propagation in a one-dimensional random medium*, Comm. Math. Sci., to appear.
[29] P. Vieillefosse, *Local interaction between vorticity and shear in a perfect incompressible flow*, J. Phys. (Paris), 43 (1982), pp. 837–842.
[30] D. Wang, *Global solutions and relaxation limits of Euler-Poisson equations*, Z. Angew. Math. Phys., 52 (2001), pp. 620–630.
[31] D. Wang, *Global solutions to the equations of viscous gas flows*, Proc. Roy. Soc. Edinburgh Sect. A, 131 (2001), pp. 437–449.
[32] D. Wang *Global solutions to the Euler-Poisson equations of two-carrier types in one-dimension*, Z. Angew. Math. Phys., 48 (1997), pp. 680–693.
[33] D. Wang and G.-Q. Chen, *Formation of singularities in compressible Euler-Poisson fluids with heat diffusion and damping relaxation*, J. Differential Equations, 144 (1998), pp. 44–65.

SIAM J. APPL. MATH.
Vol. 63, No. 6, pp. 1911–1934

ON EFFECTIVE STOPPING TIME SELECTION FOR VISCO-PLASTIC NONLINEAR BV DIFFUSION FILTERS USED IN IMAGE DENOISING*

I. A. FRIGAARD†, G. NGWA‡, AND O. SCHERZER§

Abstract. We consider denoising applications using nonlinear diffusion filters of BV type. Using the multiple timescales method, an equation is derived that approximates the time evolution of the image noise. Analysis of the corresponding variational inequality leads to an estimate of the timescale over which the noise decays to its local mean, given in terms of the filter parameters. We present a number of computed examples that demonstrate the validity of our stopping time estimate.

Key words. visco-plastic fluids, nonlinear diffusion filtering, stability, variational methods, image processing

AMS subject classifications. 76A05, 35B35

DOI. 10.1137/S0036139902400465

1. Introduction. In a diffusion filtering experiment, one attempts to recover an underlying image $U_I(\boldsymbol{x})$ from a noisy image $u_0(\boldsymbol{x}) : \ \boldsymbol{x} \in \Omega \subset \mathbb{R}^2$. This is achieved by taking $u_0(\boldsymbol{x})$ as initial data for a diffusion equation,

$$\frac{\partial u}{\partial t} = \nabla \cdot [D(\boldsymbol{x}, u, \nabla u)\nabla u] : \qquad u(\boldsymbol{x}, 0) = u_0(\boldsymbol{x}), \tag{1.1}$$

and integrating (1.1) over a certain time interval, say $t \in [0, T]$. The final solution $u(\boldsymbol{x}, T)$ is taken to be a reasonable approximation to $U_I(\boldsymbol{x})$. Choice of the diffusivity $D(\boldsymbol{x}, u, \nabla u)$ and selection of T control the effectiveness of the filtering process. In this paper we are interested in diffusion filters D, of the form:

$$D(\boldsymbol{x}, u, \nabla u) = D(\boldsymbol{x}, |\nabla u|) \equiv \mu + \frac{\tau_0}{|\nabla u|}, \tag{1.2}$$

with $\mu > 0$ and $\tau_0 > 0$.

Formally, since $D \to \infty$, where $|\nabla u| \to 0$, in place of (1.1) we generally consider

$$\frac{\partial u}{\partial t} = \nabla \cdot \boldsymbol{\tau}, \tag{1.3}$$

*Received by the editors January 2, 2002; accepted for publication (in revised form) February 20, 2003; published electronically August 15, 2003. Part of this work was carried out during a study visit partly supported by the Universität Innsbruck.

http://www.siam.org/journals/siap/63-6/40046.html

†Department of Mathematics and Department of Mechanical Engineering, University of British Columbia, 2324 Main Mall, Vancouver, BC V6T 1Z4, Canada (frigaard@math.ubc.ca). The work of this author was supported partly by the British Columbia Advanced Systems Institute through the award of a research fellowship.

‡Department of Mathematics, University of Buea, P.O. Box 63, Buea, Cameroon (ubuea@uycdc.uninet.cm). The work of this author was supported by the Abdus Salam International Centre for Theoretical Physics in Trieste, Italy, via a regular associateship.

§Department of Computer Science, Universität Innsbruck, Technikerstraße 25, A-6020 Innsbruck, Austria (Otmar.Scherzer@uibk.ac.at). The work of this author was supported by the Austrian Science Foundation (FWF) project Y-123-INF.

where $\boldsymbol{\tau} = (\tau_x, \tau_y)$:

$$|(\tau_x, \tau_y)| > \tau_0(\boldsymbol{x}) \Longleftrightarrow \begin{cases} \tau_x &= \left[\mu + \dfrac{\tau_0(\boldsymbol{x})}{|\nabla u|}\right] \dfrac{\partial u}{\partial x}, \\ \tau_y &= \left[\mu + \dfrac{\tau_0}{|\nabla u|}\right] \dfrac{\partial u}{\partial y}, \end{cases} \tag{1.4}$$

$$|(\tau_x, \tau_y)| \leq \tau_0 \Longleftrightarrow |\nabla u| = 0. \tag{1.5}$$

We remark here that (1.3)–(1.5) also model an inelastic visco-plastic fluid (a Bingham fluid [17]), of unit density, flowing axially along a cylindrical duct with cross-sectional area Ω. In this analogy, u is the fluid velocity and D represents the effective viscosity of the fluid. The parameters μ and τ_0 are referred to as the plastic viscosity and the yield stress. Various industrial muds, slurries, and pastes are represented fairly well by such rheological models [20], as well as certain porous media flows [15] and the flows of particular types of lava [14]. There is a long and well-established literature concerning the flow of such materials; see, for example, [13, 17, 33, 36, 34, 35, 39, 40, 43, 47, 50, 59, 64, 67, 72, 75, 77, 78] and many others.

The constitutive laws (1.4), (1.5) distinguish between $|\nabla u| = 0$ and $|\nabla u| > 0$. These laws could alternatively be written as

$$\tau \in \partial h(\nabla u),$$

where $\partial h(\mathbf{t})$ is the subgradient of

$$h(\mathbf{t}) := \frac{\mu}{2}|\mathbf{t}|^2 + \tau_0|\mathbf{t}|.$$

However, (1.4) and (1.5) are more physically instructive, since they immediately reveal the key characteristic of Bingham fluids, i.e., that a certain yield stress has to be exceeded in order for a flow to be initiated. The formulation (1.4), (1.5) is the classical formulation commonly used in fluid mechanics applications.

The aim of our study is to develop a *physically* motivated rationale for choosing the stopping time T when using a diffusion filter of type (1.2). We note that for computational ease we will often use a regularized form of (1.3)–(1.5). In this case, we solve (1.1) on Ω, replacing D by D_β:

$$D_\beta \equiv \mu + \frac{\tau_0}{(|\nabla u|^2 + \beta^2)^{1/2}}, \qquad \beta \ll 1. \tag{1.6}$$

Nonlinear diffusion filtering, using either (1.3)–(1.5) or (1.6), has been considered previously by Alvarez and colleagues [2, 3, 4], Catté et al. [21], Perona and Malik [63], and Weickert [76], to name but a few.

In the nonlinear diffusion framework, natural relations exist between biased diffusion and regularization theory via the Euler equation for the regularization functional. The regularization parameter and the diffusion time are analogous if one regards regularization as time-discrete diffusion filtering with a single implicit time step [68, 49, 69, 71]. A popular specific energy functional that arises from unconstrained total variation denoising [1, 22, 25] is

$$\frac{1}{2}\int_\Omega (u(\boldsymbol{x}) - u_0(\boldsymbol{x}))^2 \, d\boldsymbol{x} \; + \; \alpha \int_\Omega |\nabla u(\boldsymbol{x})| \, d\boldsymbol{x}. \tag{1.7}$$

Constrained total variation minimization also leads to a nonlinear diffusion process with a bias term using a time-dependent penalization parameter [68]. Total variation denoising in the continuous and discrete setting has been considered by many authors recently. We give a list which is not at all complete: [11, 12, 18, 22, 23, 24, 26, 27, 28, 31, 32, 37, 38, 44, 45, 46, 48, 51, 52, 53, 54, 55, 58, 60, 61, 62, 63, 66, 70, 71, 73, 74]. The minimizer of (1.7) approximates the solution of the partial differential equation

$$\frac{\partial u}{\partial t} = \nabla \cdot \left(\frac{\nabla u}{|\nabla u|} \right), \tag{1.8}$$

which has been integrated implicitly over one time step, $\Delta t = \alpha$, from the initial condition $u_0(\boldsymbol{x})$. An iterative regularization, consisting of minimizing the functionals

$$\frac{1}{2} \int_\Omega (u(\boldsymbol{x}) - u_{n-1}(\boldsymbol{x}))^2 \, d\boldsymbol{x} \; + \; \alpha \int_\Omega |\nabla u(\boldsymbol{x})| \, d\boldsymbol{x}$$

over $u(\boldsymbol{x})$, and denoting successive minimizers by u_n, will approximate the solution of (1.1) at discrete time points $t_n = n\alpha$, $n = 1, 2, \ldots$. The approximation via implicit time steps is justified by *nonlinear* semigroup theory (see, e.g., [19]). Actually, nonlinear semigroup theory defines the solution of the partial differential equation via implicit time steps. Remarkable properties of the total variation flow equation and bounded variation regularization have been derived recently [7, 9, 8, 10, 16], including analytically calculated solutions. A numerical comparison of denoising with iterative regularization and total variation flow has been given in [65, 66].

Total variation flow and bounded variation regularization are capable of reconstructing blocky data. See [30] for the continuous regularization setting. In [57] it is shown in a discrete setting that bounded variation regularization favors piecewise constant minimizers. For work on this topic in the PDE framework we refer to [7, 9, 8, 10, 16]. It is considered a drawback of the low-shear viscosity regularization (1.6) that it does not recover blocky structures accurately. However, as we have shown in the different context of multiphase Bingham fluid flow [34, 35], this filtering technique is capable of preserving horizontal fluid flow regions, which in the image processing context are data regions of constant grey level intensity. This, to a certain extent, shows that Bingham filtering techniques are capable of preserving essential image details. Since the solution of the Bingham fluid flow equation is in the Sobolev-space H^1 with respect to the space variable, the effect of staircasing, which is sometimes considered a drawback of total variation regularization and total variation flow (see [16, 57]), is limited. Of course, whether or not blocky structures are of interest depends entirely on the image and the application being considered.

The main scope of the paper is to utilize the similarity of (1.1) and Bingham mud flow to derive an asymptotic estimate for the optimal stopping time in problems such as (1.1). The basic idea is to split the solution of (1.1) into low and high frequency components. The high frequency components decay rapidly, and a time can be determined at which they have essentially disappeared. The separation of low and high frequency components as presented below is generally applicable to any sort of diffusion filtering (such as total variation flow) as long as an estimate for the time of disappearance of the high frequency components is available. For the Bingham mud flow counterpart, optimal stopping criteria can be given in terms of the *physical* parameters of the filter. In order to derive the optimal stopping time we adapt results derived by Bristeau for a Bingham fluid flow, described in Glowinski, Lions, and Trémolières [40]. These results are limited to the case $\mu > 0$. However, for

$\mu = 0$, the estimates are still formally applicable and are sharp in predicting the time of disappearance of a highly oscillating piecewise constant function, as we find by a comparison with recent results in [16].

2. Decay in nonlinear diffusion filtering. A key idea behind any form of diffusion filtering is that the decay of high frequency components in the initial data (i.e., the noise) will occur more rapidly than of the underlying image. In a digital image the spatial frequency of the noise is typically the pixel scale, say ϵ, where $\epsilon \ll 1$. We can generally expect that the basic image will be denoised by a diffusion filter only if U_I varies over length-scales larger than ϵ, i.e., unless additional information on U_I is available. For example, if $\Omega = (0,1) \times (0,1)$ and a linear diffusion filter $\tau_0 = 0$ is used, the nth Fourier mode decays like $\mathrm{e}^{-(\mu n \pi)^2 t}$. Frequencies above $n \sim \epsilon^{-1}$ are not present, and hence selection of $\mu \sim \epsilon$ ensures that linear diffusion of the noise occurs on an $O(1)$ timescale. If $T \ll 1$, we expect that the linear diffusion filter will have little effect on the initial image. Including $\tau_0 > 0$ in the filter results in a nonlinear problem, in which the decay of both the noise and underlying image is accelerated.

Suppose that (1.1) is solved numerically, imposing homogeneous Neumann conditions on $\partial\Omega$. Without loss of generality, we may assume that $u_0(\boldsymbol{x})$ has zero mean, and we may thus expect that $\|u\|(t) \to 0$ as $t \to \infty$. (Here and later, $\|\cdot\|$ will denote an L^2 norm over the domain of interest.) Thus, as $t \to \infty$, we have that $\|u_n\| \to \|U_I\|$, where

$$u(\boldsymbol{x}, t) = U_I(\boldsymbol{x}) + u_n(\boldsymbol{x}, t), \tag{2.1}$$

i.e., $u_n(\boldsymbol{x}, t)$ is the noise. Our interest here is to find an (asymptotic) expression for T, in terms of the *physical* parameters of the problem, ϵ, μ, τ_0, and $u_0(\boldsymbol{x})$, that will give an optimal recovered image. Presumably an *optimal* image will be recovered at a time T that is close to the minimum ratio of $\|u_n\|$ to $\|U_I\|$.

An illustrative one-dimensional example is given in Figure 1. An initial image (Figure 1(a)) is perturbed by white noise (Figure 1(b)). Equation (1.1) with Neumann conditions is integrated using a fully implicit finite difference scheme. In Figure 1(c) we show $\|u\|(t)$ and $\|u_n\|(t)$. We see that $\|u_n\|(t)$ exhibits an initial rapid decay followed by a slow increase, as the entire image $u(t)$ decays to its mean. Figures 1(d)–(f) show $u(\boldsymbol{x}, t)$ for two choices of t, close to the minimum of $\|u_n\|(t)$, and for a third choice of t, as $\|u\|(t)$ has begun to decay significantly. It is apparent from Figures 1(d)–(f) that during the initial rapid decay of $\|u_n\|(t)$, the noise in the image is averaged over a local length-scale, and that on a longer timescale this averaging length-scale increases to the entire image.

2.1. Decay estimates for $u(x, t)$. The eventual decay of $\|u\|(t)$ in Figure 1(c) would be expected for a diffusive process with nondegenerate diffusivity, but what is peculiar to the problem with $\tau_0 > 0$ is that $\|u\|(t) \to 0$ in a finite time. This decay is a result of having a finite yield stress: physically there comes a time when the stress everywhere decays below the yield stress, and (see (1.5)) there will be zero *rate of strain* everywhere; i.e., $|\nabla u| = 0$.

From[1] Theorem 3.2 in Appendix 6 of [40], it is well known that if u is the (weak)

[1] Theorem 3.2 in Appendix 6 of [40] is slightly more general and includes a pressure gradient term on the right-hand side of (1.3); i.e., in [40], set $f = 0$, $\beta = b_D$, and $g = \tau_0$ to apply this result.

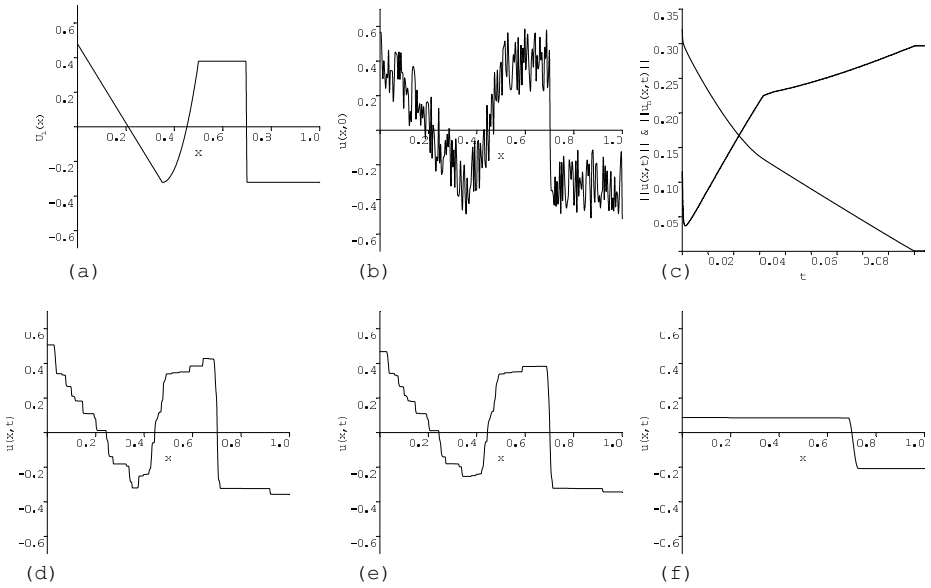


FIG. 1. *One-dimensional example: discretization/pixel-scale* $\epsilon = \Delta x = 1/256$, $\mu = 1/256$, $\tau = 1.0$. (a) $U_I(\boldsymbol{x})$; (b) *initial data,* $u_0(\boldsymbol{x})$, *image plus white noise with maximum amplitude* ± 0.2; (c) $\|u\|(t)$ *and* $\|u_n\|(t)$ *(dashed line);* (d) $u(\boldsymbol{x}, t)$*:* $t = 1.078 \times 10^{-3}$; (e) $u(\boldsymbol{x}, t)$*:* $t = 2.121 \times 10^{-3}$; (f) $u(\boldsymbol{x}, t)$*:* $t = 3.481 \times 10^{-2}$.

solution of the Dirichlet problem corresponding to (1.3)–(1.5), then

$$u(t) = 0 : \qquad t \geq t_0 = \frac{1}{\lambda_0 \mu} \log \left(1 + \lambda_0 \mu \frac{\|u_0\|}{b_D \tau_0} \right), \tag{2.2}$$

where $\lambda_0 > 0$ is the smallest eigenvalue of $-\Delta$ in $V = H_0^1(\Omega)$ and

$$b_D = \inf_{v \in V,\ v \neq 0} \frac{\int_\Omega |\nabla v| \, \mathrm{d}\Omega}{\|v\|} > 0. \tag{2.3}$$

Thus, for $\tau_0 > 0$, decay of $\|u\|(t)$ is no longer exponential, and we expect that $\|u\|(t) \to 0$ in a finite time t_0. For the Neumann problem, we set

$$\hat{V} = \left\{ v \in C^1(\bar{\Omega}) : \ \int_\Omega v \, \mathrm{d}\Omega = 0, \ \ \frac{\partial v}{\partial n} = 0 \text{ on } \partial\Omega \right\}$$

and let V be the closure of $\hat{V}$ with respect to the $H^1(\Omega)$ norm. We then proceed essentially as in [40]; the exact details are not particularly interesting. We have that if u is the weak solution of the Neumann problem corresponding to (1.3)–(1.5), then

$$u = 0 : \qquad t \geq t_0 = \frac{1}{\lambda_0 \mu} \log \left(1 + \lambda_0 \mu \frac{\|u_0\|}{b_N \tau_0} \right), \tag{2.4}$$

where $\lambda_0 > 0$ is the smallest eigenvalue of $-\Delta$ in V and

$$b_N = \inf_{v \in V,\ v \neq 0} \frac{\int_\Omega |\nabla v| \, \mathrm{d}\Omega}{\|v\|} > 0. \tag{2.5}$$

For the example in Figure 1 and associated parameters, we compute $t_0 \approx 0.154$, which

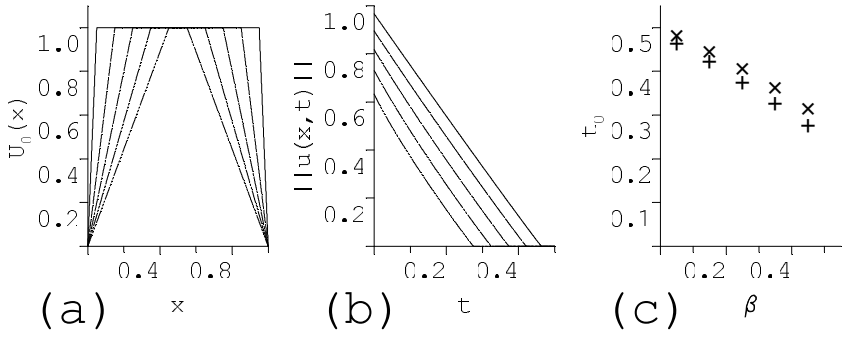


FIG. 2. *Decay of* $\|u\|(t)$ *for different initial conditions* $u(x,0) = u_\beta(x)$, $\Delta x = 0.002$, $\mu = 0.002$, $\tau = 1.0$*:* (a) $u_\beta(x)$, $\beta = 0.05,\ 0.15,\ 0.25,\ 0.35,\ 0.45$*;* (b) $\|u\|(t)$, $\beta = 0.05,\ 0.15,\ 0.25,\ 0.35,\ 0.45$*;* (c) t_0 $(\times)$ *from* (2.2) *compared with computed values* $(+)$.

clearly overestimates the decay shown in Figure 1(c). However, estimates such as (2.2) are not always this poor and will depend on the initial data. In particular, as $u(x,0)$ approaches a minimizer of (2.3), the estimate can be very sharp. As an illustration, we have solved a one-dimensional Dirichlet problem (1.1), for various initial conditions:

$$u(x,0) = u_\beta(x) = \left\{ \begin{array}{ll} \dfrac{x}{\beta}, & 0 \le x < \beta, \\ 1, & \beta \le x < 1-\beta, \\ \dfrac{1-x}{\beta}, & 1-\beta \le x < 1. \end{array} \right\} \tag{2.6}$$

Figure 2(b) shows the decay of $\|u\|(t)$ for the range of initial conditions shown in Figure 2(a). Figure 2(c) shows the difference between the computed decay time and t_0, given by (2.2), as β is varied.

2.2. Local decay estimates for $u_n(x,t)$. Assuming the separation of spatial length-scales between $u_n(\boldsymbol{x},t)$ and $U_I(\boldsymbol{x})$, what we really wish to know is the time interval taken for $u_n(\boldsymbol{x},t)$ to decay to a local mean value. Further, we would like to understand how this decay timescale can be affected by the image $U_I(\boldsymbol{x})$. The estimate we develop is based loosely on the method of multiple scales; e.g., see [42]. Suppose that $\Omega = (0,1)\times(0,1)$ and that, in the neighborhood of an arbitrary $\boldsymbol{x}_0 \in \Omega$, $u_0(\boldsymbol{x})$ is of the form

$$u_0(\boldsymbol{x}) = U_I(\boldsymbol{x}) + u_{n,0}(\tilde{\boldsymbol{x}}): \qquad \tilde{\boldsymbol{x}} = \frac{\boldsymbol{x}-\boldsymbol{x}_0}{\epsilon}, \tag{2.7}$$

where $\epsilon \ll 1$ is the pixel scale, i.e., we assume that the initial image has multiple scales. Consider the forward integration of (1.1) in the neighborhood of $\boldsymbol{x}_0$. Since we expect the high frequency components to decay faster than the low frequency components, we seek a solution of (1.1) of the form

$$u(\boldsymbol{x},t) = U(\boldsymbol{x},t) + u_n(\tilde{\boldsymbol{x}},\tilde{t}), \tag{2.8}$$

where $\tilde{t} = t/\delta$ for some $\delta \ll 1$. The differential operators in (1.1) become

$$\frac{\partial}{\partial t} \to \frac{1}{\delta}\frac{\partial}{\partial \tilde{t}} + \frac{\partial}{\partial t}, \tag{2.9}$$

$$\nabla \to \frac{1}{\epsilon}\tilde{\nabla} + \nabla, \tag{2.10}$$

i.e., with $\tilde{\nabla}$ containing spatial derivatives with respect to the components of $\tilde{\boldsymbol{x}}$. We now make a formal expansion of the right-hand side of (1.1), in terms of ϵ and δ:

$$\begin{aligned}\left[\mu + \frac{\tau_0}{|\nabla u|}\right]\nabla u \to{}& \frac{\mu}{\epsilon}\tilde{\nabla}u_n + \mu\nabla U + \frac{\tau_0}{|\tilde{\nabla}u_n|}\tilde{\nabla}u_n \\ &+ \epsilon\frac{\tau_0}{|\tilde{\nabla}u_n|}\left[\nabla U - \frac{(\nabla U\cdot\tilde{\nabla}u_n)}{|\tilde{\nabla}u_n|^2}\tilde{\nabla}u_n\right] \\ &+ \epsilon^2\frac{\tau_0}{|\tilde{\nabla}u_n|}\left[\frac{3}{2}\frac{(\nabla U\cdot\tilde{\nabla}u_n)^2}{|\tilde{\nabla}u_n|^4} - \frac{1}{2}\frac{|\nabla U|^2}{|\tilde{\nabla}u_n|}\right]\tilde{\nabla}u_n \\ &- \epsilon^2\frac{\tau_0}{|\tilde{\nabla}u_n|}\frac{(\nabla U\cdot\tilde{\nabla}u_n)}{|\tilde{\nabla}u_n|^2}\nabla U + O(\epsilon^3).\end{aligned} \tag{2.11}$$

Substituting into (1.1), we obtain

$$\begin{aligned}\frac{1}{\delta}\frac{\partial u_n}{\partial \tilde{t}} + \frac{\partial U}{\partial t} = {}& \frac{1}{\epsilon^2}\mu\tilde{\nabla}^2 u_n + \frac{1}{\epsilon}\tilde{\nabla}\cdot\left[\frac{\tau_0}{|\tilde{\nabla}u_n|}\tilde{\nabla}u_n\right] + \nabla\cdot(\mu\nabla U) \\ &+ \tilde{\nabla}\cdot\left[\frac{\tau_0}{|\tilde{\nabla}u_n|}\left(\nabla U - \frac{(\nabla U\cdot\tilde{\nabla}u_n)}{|\tilde{\nabla}u_n|^2}\tilde{\nabla}u_n\right)\right] + O(\epsilon).\end{aligned} \tag{2.12}$$

We balance the leading order terms, by setting $\delta = \epsilon^2$, and neglect terms of order ϵ^3:

$$\begin{aligned}\frac{\partial u_n}{\partial \tilde{t}} = {}& \tilde{\nabla}\cdot\left[\left(\mu + \frac{\epsilon\tau_0}{|\tilde{\nabla}u_n|}\right)\tilde{\nabla}u_n\right] + \epsilon^2[\nabla\cdot(\mu\nabla U) - U_t] \\ &+ \epsilon^2\tilde{\nabla}\cdot\left[\frac{\tau_0}{|\tilde{\nabla}u_n|}\left(\nabla U - \frac{(\nabla U\cdot\tilde{\nabla}u_n)}{|\tilde{\nabla}u_n|^2}\tilde{\nabla}u_n\right)\right].\end{aligned} \tag{2.13}$$

Rather than proceeding asymptotically, we simply consider ϵ as a small finite parameter (known from the resolution of the image), which we retain to order ϵ^2 in order to understand the leading order effects of the underlying image U_I.

It is, however, apparent that (2.13) is incomplete, since the limiting behavior as $|\tilde{\nabla}u_n| \to 0$ has been ignored in our formal expansions. In order to have a regular solution to (2.13) (and indeed for our expansions to be valid), it is necessary that $|\tilde{\nabla}u_n| \to 0$ only if $|\nabla U| = 0$. Thus, in place of (2.13) we consider

$$\frac{\partial u_n}{\partial \tilde{t}} = \tilde{\nabla}\cdot[\tilde{\boldsymbol{\tau}} + \epsilon^2\tilde{\boldsymbol{m}}] + \epsilon^2 f, \tag{2.14}$$

where

$$f(\boldsymbol{x}, t) = [\nabla\cdot(\mu\nabla U) - U_t] \tag{2.15}$$

and where we define $\tilde{\boldsymbol{\tau}} = (\tilde{\tau}_x, \tilde{\tau}_y)$:

$$|\tilde{\nabla} u_n| = 0 \Longleftrightarrow |\nabla U| = 0 \quad \text{and} \quad |(\tilde{\tau}_x, \tilde{\tau}_y)| \leq \tilde{\tau}_0, \tag{2.16}$$

$$|\tilde{\nabla} u_n| > 0 \Longleftrightarrow \begin{cases} \tilde{\boldsymbol{\tau}} &= \left[\mu + \dfrac{\epsilon \tau_0}{|\tilde{\nabla} u_n|}\right] \tilde{\nabla} u_n, \\ \tilde{\boldsymbol{m}} &= \dfrac{\tau_0}{|\tilde{\nabla} u_n|} \left[\nabla U - \dfrac{(\nabla U \cdot \tilde{\nabla} u_n)}{|\tilde{\nabla} u_n|^2} \tilde{\nabla} u_n\right]. \end{cases} \tag{2.17}$$

2.2.1. Variational formulation. We return in section 2.2.3 to a physical interpretation of (2.14)–(2.17), but to derive our stopping estimate it is somewhat easier to work with a variational formulation.

We consider $\boldsymbol{x}_0$ to be the corner of a pixel and consider solution of (2.14)–(2.17) in a local domain surrounding $\boldsymbol{x}_0$: $\tilde{\boldsymbol{x}} \in \tilde{\Omega}_k = (-k, k) \times (-k, k)$, where k is numerically of order 1, with initial conditions

$$u_n(\tilde{\boldsymbol{x}}, 0) = u_{n,0}(\tilde{\boldsymbol{x}}). \tag{2.18}$$

It is also necessary to impose boundary conditions on $\partial\tilde{\Omega}_k$, and for this we choose homogeneous Neumann conditions. Equations (2.14)–(2.17) are still diffusive (see section 2.2.3), and the effect of this choice of boundary conditions is that u_n is expected to decay to its mean value over $\tilde{\Omega}_k$. Subtracting any constant value from u_n leaves (2.14)–(2.17) unchanged, and hence we consider only functions of zero mean.

A classical solution to (2.14)–(2.18) will also satisfy the following variational inequality:

$$\begin{aligned} \epsilon^2 f \langle [v - u_n] \rangle \leq & \left\langle \frac{\partial u_n}{\partial \tilde{t}} [v - u_n] \right\rangle + \mu a(u_n, v - u_n) + \epsilon \tau_0 [j(v) - j(u_n)] \\ & + \epsilon^2 \langle \tilde{\boldsymbol{m}} \cdot \nabla (v - u_n) \rangle \\ & \forall v \in V, \qquad u_n(\tilde{\boldsymbol{x}}, 0) = u_{n,0}(\tilde{\boldsymbol{x}}), \end{aligned} \tag{2.19}$$

where

$$\langle v \rangle = \int_{\tilde{\Omega}_k} v \, \mathrm{d}\tilde{x}, \tag{2.20}$$

$$a(u_n, v) = \left\langle \tilde{\nabla} u_n \cdot \tilde{\nabla} v \right\rangle, \tag{2.21}$$

$$j(v) = \left\langle |\tilde{\nabla} v| \right\rangle, \tag{2.22}$$

$$V = \left\{ v \in H^1(\tilde{\Omega}_k) : \int_{\tilde{\Omega}_k} v \, \mathrm{d}\Omega = 0, \quad |\tilde{\nabla} v| = 0 \Rightarrow |\nabla U| = 0 \right\}. \tag{2.23}$$

Note that $\frac{\partial v}{\partial n} = 0$ on $\partial\tilde{\Omega}_k$ is inherent in the above variational formulation. In deriving (2.19), note that f does not vary over the length-scale of $\tilde{\Omega}$. We assume that there exists a solution u_n to (2.19), and that $u_n \in V$, with sufficient regularity for what follows. Existence and uniqueness of solutions to (2.19) and the steady version of (2.19) are interesting problems in their own right. We consider this further in Appendix A. By choosing $v = 0$ and $v = 2u_n$, we see that u_n satisfies

$$\|u_n\| \frac{\mathrm{d}}{\mathrm{d}\tilde{t}} \|u_n\| = \left\langle \frac{\partial u_n}{\partial \tilde{t}} u_n \right\rangle = -\mu a(u_n, u_n) - \epsilon \tau_0 j(u_n) + \epsilon^2 f \langle u_n \rangle. \tag{2.24}$$

Following [39, 40], we bound as follows:

$$\|u_n\| \frac{\mathrm{d}}{\mathrm{d}\tilde{t}} \|u_n\| \leq -\mu\lambda_0 \|u_n\|^2 - b_{N,k}\epsilon\tau_0 \|u_n\| + \epsilon^2|f| \|u_n\|,$$

with λ_0 the least positive eigenvalue of $-\Delta$ over V on $\tilde{\Omega}_k$. Assuming that $\|u_n\| \neq 0$,

$$\frac{\mathrm{d}}{\mathrm{d}\tilde{t}} \|u_n\| \leq -\mu\lambda_0 \|u_n\| - b_{N,k}\epsilon\tau_0 + \epsilon^2|f|. \tag{2.25}$$

Note that $\lambda_0 = \pi^2/k^2$, and by a simple mapping,

$$b_{N,k} \equiv \inf_{v\in V,\ v\neq 0} \frac{\int_{\tilde{\Omega}_k} |\tilde{\nabla} v|\ \mathrm{d}\tilde{\boldsymbol{x}}}{\|v\|_{L^2(\tilde{\Omega}_k)}} = b_N = 2. \tag{2.26}$$

Consider also the steady version of (2.19):

$$\begin{aligned} \mu a(u_n^*, v - u_n^*) + \epsilon\tau_0[j(v) - j(u_n^*)] \geq \epsilon^2 f \langle [v - u_n^*] \rangle - \epsilon^2 \langle \tilde{\boldsymbol{m}} \cdot \nabla(v - u_n^*) \rangle \\ \forall v \in V, \quad u_n^* \in V. \end{aligned} \tag{2.27}$$

Proceeding as for the transient case, we find the bound:

$$0 \leq -\mu\lambda_0 \|u_n^*\|^2 - b_{N,k}\epsilon\tau_0 \|u_n^*\| + \epsilon^2|f| \|u_n^*\|. \tag{2.28}$$

Therefore $u_n^* = 0$ is the unique steady solution (and also a transient solution) if

$$\tau_0 > \frac{\epsilon|f|}{b_{N,k}} \tag{2.29}$$

is satisfied. Returning now to the transient problem and assuming that (2.29) is satisfied, we proceed as in [40, 39]. Applying Gronwall's lemma, we get that

$$\begin{aligned} \|u_n\|(\tilde{t}) &\leq \left[\|u_n\|(0) + \frac{b_{N,k}\epsilon\tau_0 - \epsilon^2|f|}{\lambda_0\mu}\right] \mathrm{e}^{-\lambda_0\mu\tilde{t}} - \frac{b_{N,k}\epsilon\tau_0 - \epsilon^2|f|}{\lambda_0\mu} \\ &\leq 0 \qquad \text{if } \ \lambda_0\mu\tilde{t} \geq \ln\left[1 + \frac{\lambda_0\mu \|u_n\|(0)}{b_{N,k}\epsilon\tau_0 - \epsilon^2|f|}\right]. \end{aligned} \tag{2.30}$$

Returning to the integration of (1.1) over the time variable $t = \epsilon^2\tilde{t}$, we expect that u_n will have decayed to its mean value over $\tilde{\Omega}_k$, after integrating for

$$t \geq t_0 = \frac{\epsilon^2}{\mu\lambda_0} \ln\left[1 + \frac{\lambda_0\mu \|u_n\|(0)}{b_{N,k}\epsilon\tau_0 - \epsilon^2|f|}\right]. \tag{2.31}$$

2.2.2. Relation to total variation filtering. In the total variation denoising case (i.e., $\mu = 0$ and $\tau_0 = 1$) it follows purely formally from (2.25) that

$$\frac{\mathrm{d}}{\mathrm{d}\tilde{t}} \|u_n\| \leq -b_{N,k}\epsilon + \epsilon^2|f|.$$

Thus, by integrating with respect to $\tilde{t}$, it follows that

$$\|u_n\|(\tilde{t}) \leq \|u_n\|(0) + \tilde{t}(-b_{N,k}\epsilon + \epsilon^2|f|).$$

Thus, returning to the integration variable t, we expect $\|u_n\|(\tilde{t}) = 0$ for

$$t \geq t_0 := \frac{\epsilon \|u_n\|(0)}{b_{N,k} - \epsilon|f|} = \frac{\epsilon \|u_n\|(0)}{2 - \epsilon|f|}. \tag{2.32}$$

This above estimate also follows from (2.31) as the leading order term of a Taylor series expansion as $\mu \to 0$. Since we have derived (2.31) in H^1, the estimate (2.32) remains purely formal for total variation denoising.

To evaluate the quality of (2.32) we can rely on a variety of results. In [16] the minimizer u_α of (1.7) has been calculated analytically for given piecewise constant input data in $\mathbb{R}^2$. As a consequence of these calculations a minimal regularization parameter α_* can be calculated such that $u_{\alpha_*} = 0$. For the sake of simplicity of presentation we assume that u_0 is just noise, in which case we have

$$u_0(x) = \sum_{i=1}^{n_p} u_{0,i}\chi_{\Omega_i} = \sum_{i=1}^{n_p} \frac{\epsilon}{4} u_{0,i} \frac{4\epsilon}{\epsilon^2}\chi_{\Omega_i}, \tag{2.33}$$

where Ω_i is a square (pixel) with length ϵ, n_p is the number of pixels (index i), χ_{Ω_i} is an indicator function on Ω_i, and $|u_{0,i}| \leq \rho$ is the noise amplitude. We have introduced (2.33) to be in accordance with [16], and the factor $\frac{4\epsilon}{\epsilon^2}$ is the ratio of the perimeter to the Lebesgue measure of Ω_i. Thus, for a regularization parameter α satisfying $\alpha \geq \epsilon\rho/4$ we have $u_\alpha = 0$. For $\Omega = (0,1) \times (0,1)$, setting $U = 0$, $f = 0$, $\tau_0 = 1$, and $\mu = 0$, it follows from (2.32) that $u_n(t) = 0$ for

$$t \geq t_0 := \frac{\rho\epsilon}{2}.$$

This shows that the regularization parameter of BV-regularization and our estimated stopping time for total variation flow (2.32) differ by a factor of 2. Note, however, that there is a discrepancy between the models: we concentrate on a diffusion framework on a bounded domain, while in [16] a regularization framework on $\mathbb{R}^2$ is considered. Moreover, the estimates in [16] are for piecewise constant functions, while here the estimates are valid for all noise functions. The factor of 2 arises by our defining $b_{N,k}$ over the set of functions of zero mean. If relaxed, the two estimates will coincide. Equivalently, the estimate for $b_{N,k}$ is not optimal in the situation that the total variation flow solution is piecewise constant. Since the total variation flow solution is piecewise constant over time, we could use $b_N = 4$ in this situation, which would yield an estimate analogous to [16].

Finally, we mention that the topic of solvability and exact solution of the total variation flow equation has been addressed in a series of papers [7, 9, 16].

2.2.3. Physical interpretation of (2.14)–(2.17). The underlying image U_I affects u_n at $O(\epsilon^2)$ in two ways in (2.14). The interpretation of f is quite straightforward: as a fluid flow problem, f is an imposed pressure gradient or body force, whereas as a heat/diffusion problem, f is a heat/concentration source term. The second way in which U affects u_n at $O(\epsilon^2)$ in (2.14) is through $\tilde{\boldsymbol{m}}$. We show that $\tilde{\boldsymbol{m}}$ has the effect of introducing an anisotropy into the diffusivity, aligned with the

underlying image. Observe that for $|\tilde{\nabla} u_n| \neq 0$

$$\begin{aligned}\frac{|\tilde{\nabla} u_n|^3}{\tau_0}\tilde{\boldsymbol{m}} &= \nabla U |\tilde{\nabla} u_n|^2 - (\nabla U \cdot \tilde{\nabla} u_n)\tilde{\nabla} u_n \\ &= \begin{pmatrix} -\dfrac{\partial u_n}{\partial \tilde{x}_2}\dfrac{\partial U}{\partial x_2} & \dfrac{\partial u_n}{\partial \tilde{x}_2}\dfrac{\partial U}{\partial x_1} \\ \dfrac{\partial u_n}{\partial \tilde{x}_1}\dfrac{\partial U}{\partial x_2} & -\dfrac{\partial u_n}{\partial \tilde{x}_1}\dfrac{\partial U}{\partial x_1} \end{pmatrix} \begin{pmatrix} \dfrac{\partial u_n}{\partial \tilde{x}_1} \\ \dfrac{\partial u_n}{\partial \tilde{x}_2} \end{pmatrix}.\end{aligned}$$

Together with (2.17), this shows that (2.14) can be rewritten as

$$\frac{\partial u_n}{\partial \tilde{t}} = \tilde{\nabla} \cdot \left[A(\tilde{\nabla} u_n)\tilde{\nabla} u_n \right] + \epsilon^2 f,$$

with

$$A(\tilde{\nabla} u_n) = \left(\mu + \frac{\epsilon\tau_0}{|\tilde{\nabla} u_n|}\right) I + \frac{\epsilon^2\tau_0}{|\tilde{\nabla} u_n|^3} \begin{pmatrix} -\dfrac{\partial u_n}{\partial \tilde{x}_2}\dfrac{\partial U}{\partial x_2} & \dfrac{\partial u_n}{\partial \tilde{x}_2}\dfrac{\partial U}{\partial x_1} \\ \dfrac{\partial u_n}{\partial \tilde{x}_1}\dfrac{\partial U}{\partial x_2} & -\dfrac{\partial u_n}{\partial \tilde{x}_1}\dfrac{\partial U}{\partial x_1} \end{pmatrix}.$$

This shows that (2.14) is an inhomogeneous anisotropic heat equation.

LEMMA 2.1. *Let $|\tilde{\nabla} u_n| \neq 0$. The two eigenvalues of $A(\tilde{\nabla} u_n)$ are given by*

$$\begin{aligned}\lambda_1 &= \left(\mu + \frac{\epsilon\tau_0}{|\tilde{\nabla} u_n|}\right), \\ \lambda_2 &= \left(\mu + \frac{\epsilon\tau_0}{|\tilde{\nabla} u_n|}\right) - \frac{\epsilon^2\tau_0}{|\tilde{\nabla} u_n|^3}\left(\frac{\partial u_n}{\partial \tilde{x}_1}\frac{\partial U}{\partial x_1} + \frac{\partial u_n}{\partial \tilde{x}_2}\frac{\partial U}{\partial x_2}\right).\end{aligned}$$

If $|\nabla U| \neq 0$, then the (right) eigenvectors are given by

$$v_1 = \frac{1}{|\nabla U|}\left(\frac{\partial U}{\partial x_1}, \frac{\partial U}{\partial x_2}\right)^t, \qquad v_2 = \frac{1}{|\tilde{\nabla} u_n|}\left(-\frac{\partial u_n}{\partial \tilde{x}_2}, \frac{\partial u_n}{\partial \tilde{x}_1}\right)^t.$$

Proof.

1. Let us define

$$a = \left(\mu + \frac{\epsilon\tau_0}{|\tilde{\nabla} u_n|}\right) \quad \text{and} \quad c = \frac{\epsilon^2\tau_0}{|\tilde{\nabla} u_n|^3}.$$

Then

$$A(\tilde{\nabla} u_n) = \begin{pmatrix} a - c\dfrac{\partial u_n}{\partial \tilde{x}_2}\dfrac{\partial U}{\partial x_2} & c\dfrac{\partial u_n}{\partial \tilde{x}_2}\dfrac{\partial U}{\partial x_1} \\ c\dfrac{\partial u_n}{\partial \tilde{x}_1}\dfrac{\partial U}{\partial x_2} & a - c\dfrac{\partial u_n}{\partial \tilde{x}_1}\dfrac{\partial U}{\partial x_1} \end{pmatrix}.$$

Now, if λ is an eigenvalue of $A(\tilde{\nabla} u_n)$, we have $\det(\lambda - A(\tilde{\nabla} u_n)) = 0$, which gives the assertion after simple algebra.

2. The first eigenvector is easily calculated from the identity:

$$A(\tilde{\nabla} u_n) = \lambda_1 I + R,$$

with

$$R = \frac{\epsilon^2 \tau_0}{|\tilde{\nabla} u_n|^3} \begin{pmatrix} -\dfrac{\partial u_n}{\partial \tilde{x}_2}\dfrac{\partial U}{\partial x_2} & \dfrac{\partial u_n}{\partial \tilde{x}_2}\dfrac{\partial U}{\partial x_1} \\ \dfrac{\partial u_n}{\partial \tilde{x}_1}\dfrac{\partial U}{\partial x_2} & -\dfrac{\partial u_n}{\partial \tilde{x}_1}\dfrac{\partial U}{\partial x_1} \end{pmatrix}.$$

This shows that the eigenvector according to the first eigenvalue λ_1 must be in the nullspace of R. If $|\nabla U| \neq 0$, ∇U spans the nullspace. The second eigenvector can be calculated by using simple linear algebra.

This shows that the anisotropic diffusion directions are determined from the diffusion directions of the diffused data originating from noise free data.

3. Numerical results. Application of (2.31) and the preceding estimates is not immediate, since it is necessary to estimate $|f|$. Provided that $t_0 \ll 1$, we can consider $|f| = |\nabla(\mu \nabla U) - U_t|$ to be pseudosteady with respect to $\tilde{t}$. To estimate f we might consider deriving the problem satisfied by U in the multiple timescales method. However, due to the nonlinearity, the problem for U does not decouple, and it would be necessary to compute various long-time averages of u_n. We have not derived this problem and instead take a more direct approach. We assume an approximate balance,

$$|U_t| \sim |\nabla(\mu \nabla U)|,$$

for the long-time U-problem. Provided that $\epsilon |\nabla(\mu \nabla U)| \ll 1$, any uniform value $\tau_0 \sim 1$ should dominate the underlying effects of the image. This balance will be met when U_I changes rapidly over a length-scale $\sim [\mu\epsilon]^{1/2} \ll 1$. Consequently, we suggest selecting $\mu \sim \epsilon$ and effectively neglecting the $\epsilon^2 f$ term in the denominator (2.31). Our estimate for the decay timescale of u_n to its local mean over the scale $k\epsilon$ is

$$t \sim T = \frac{\epsilon^2 k^2}{\mu \pi^2} \ln \left[1 + \frac{\pi^2 \mu \left\| u_n \right\| (0)}{2k^2 \epsilon \tau_0} \right], \tag{3.1}$$

which we expect to produce reasonable results, provided that spatial variations in U_I occur over length-scales $\gg [\mu\epsilon]^{1/2}$, i.e., the pixel scale for our choice of μ. We proceed with a number of numerical examples.

In the following examples, we have integrated (1.1) using the regularized form (1.6) of D, with $\beta = 10^{-3}$. A fully implicit fractional steps method is used. Since the derivation of (3.1) is purely formal and mostly heuristic, our first objective is to demonstrate that (3.1) produces a sensible estimate of the stopping time to be used with filters of the form (1.2).

3.1. Example 1. As our first example, we take the image shown in Figure 3, which has been normalized so as to have zero mean and so that $\|U_I\| = 1$. This image is perturbed on each pixel with random white noise of maximum amplitude $\Delta u_{n,\max}$. We fix a base case with parameters $\Delta x = \epsilon = 0.01$, $\mu = 0.01$, $\tau_0 = 1.0$, $\Delta u_{n,\max} = 0.5$, and vary each of μ, τ_0, and $\Delta u_{n,\max}$ in turn. Plotted in Figures 3(b)–(f) is the relative noise level $\|u_n\|(t)/\|u_n\|(0)$ against the time t, normalized with the stopping time estimate (3.1), i.e., t/T. Figures 3(b)–(d) show this decay for varying $\Delta u_{n,\max}$, τ_0, and μ, respectively. Here we have fixed $k = 1$. In Figures 3(e)–(f) we show the variations in the decay of $\|u_n\|(t)/\|u_n\|(0)$ with τ_0 and μ, for $k = 2$; i.e., we wait for the local average to develop over a larger length-scale.

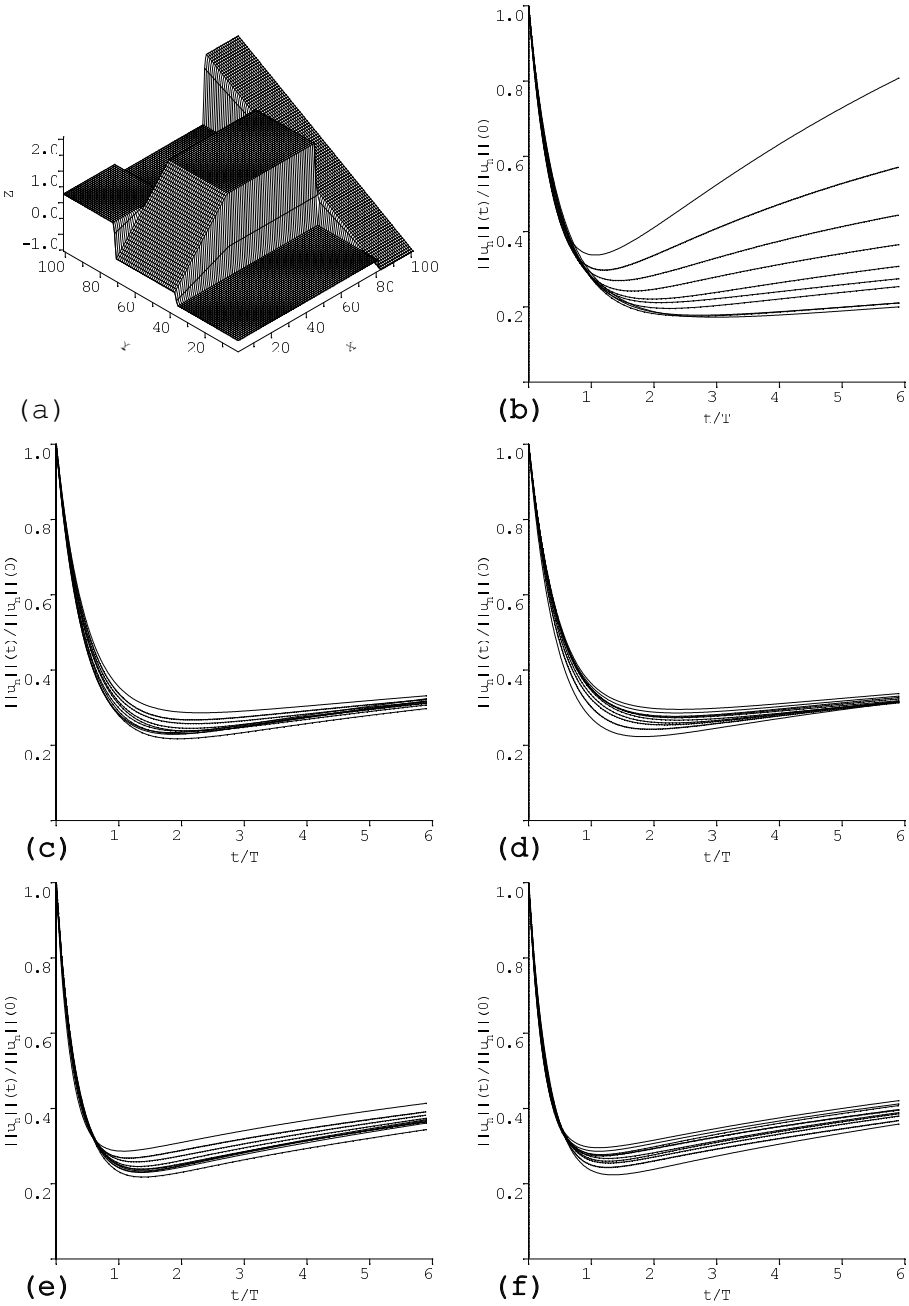


FIG. 3. *Decay of relative noise:* $\|u_n\|(t)/\|u_n\|(0)$ *with* t/T. *Base parameters:* $\Delta x = \epsilon = 0.01$, $\mu = 0.01$, $\tau_0 = 1.0$, $\Delta u_{n,\max} = 0.5$. (a) $U_I(\boldsymbol{x})$; (b) *variations with* $\Delta u_{n,\max} = 0.1,\ 0.2, \ldots,\ 0.9,\ 1.0$, $k = 1$; (c) *variations with* $\tau_0 = 0.1,\ 0.2, \ldots,\ 0.9,\ 1.0$, *initial perturbation of* $U_I(\boldsymbol{x})$ *with white noise of maximum amplitude* $\Delta u_{n,\max} = 0.5$, $k = 1$; (d) *variations with* $\mu = 0.01,\ 0.02, \ldots,\ 0.09,\ 0.1$, $k = 1$; (e) *variations with* $\tau_0 = 0.1,\ 0.2, \ldots,\ 0.9,\ 1.0$, *initial perturbation of* $U_I(\boldsymbol{x})$ *with white noise of maximum amplitude* $\Delta u_{n,\max} = 0.5$, $k = 2$; (f) *variations with* $\mu = 0.01,\ 0.02, \ldots,\ 0.09,\ 0.1$, $k = 2$.

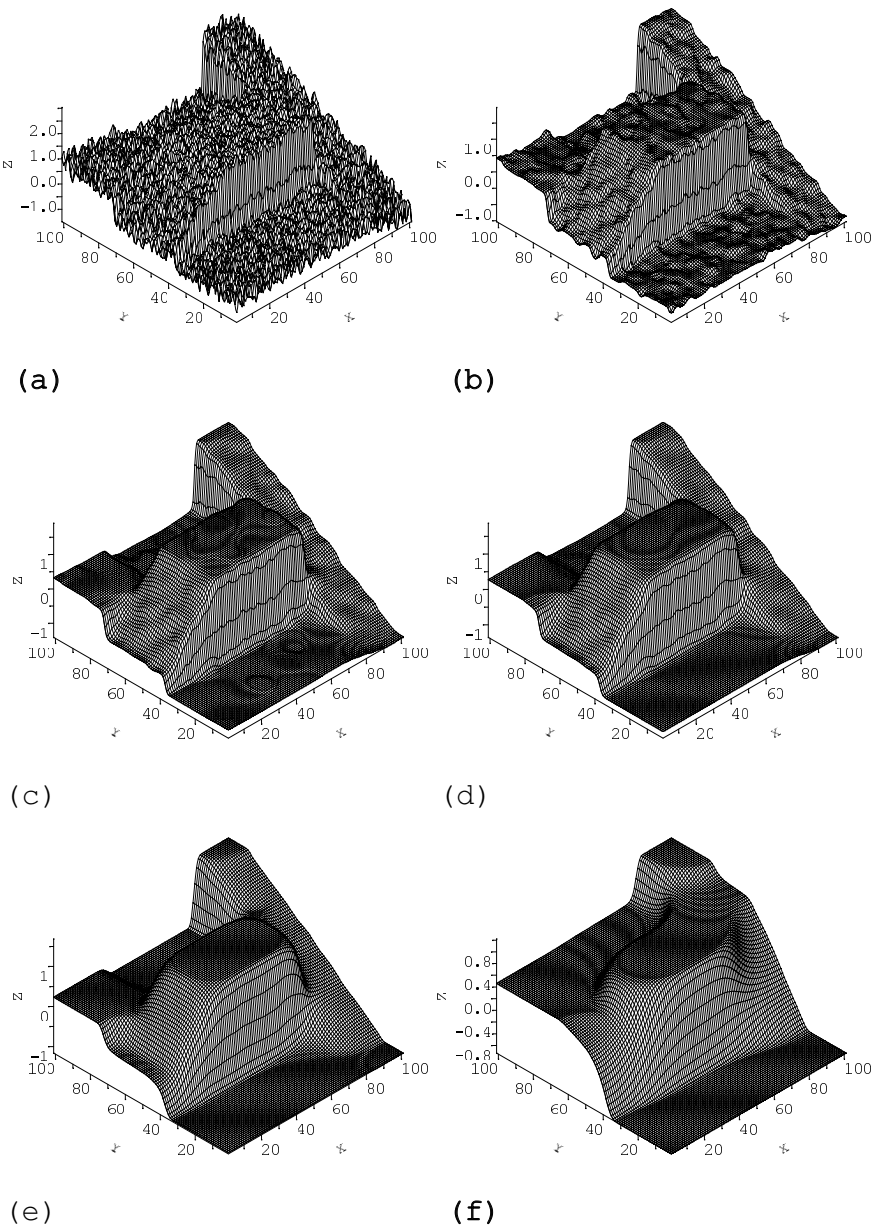


FIG. 4. *Evolution of the image* $u(\boldsymbol{x},t)$ *of Figure* 3(a)*, perturbed with white noise of maximum amplitude* $\pm\Delta u_{n,\max} = 0.5$. (a) $t/T = 0$, (b) $t/T = 1$, (c) $t/T = 2$ *(approximately the minimum of* $\|u_n\|(t)/\|u_n\|(0)$*)*, (d) $t/T = 3$, (e) $t/T = 10$, (f) $t/T = 50$.

The first observation from Figure 3 is that the minimum in $\|u_n\|(t)/\|u_n\|(0)$ is reached for $t/T \sim O(1)$, implying that (3.1) produces a sensible estimate of an appropriate stopping time. For $k = 1$, the minimum is fairly shallow and can be made slightly more extreme by increasing k. Note, however, that we generally do not wish to average the noise over too large a length-scale since the image may then deteriorate. As an example of the evolution of the image, we show in Figure 4 the solution $u(t)$ at different times $t/T = 0,\ 1,\ 2,\ 3,\ 10,\ 50$ for the base case $\Delta x = \epsilon = 0.01$, $\mu = 0.01$, $\tau_0 = 1.0$, $\Delta u_{n,\max} = 0.5$, $k = 2$. The more acceptable recovered images are close to the minimum of $\|u_n\|(t)/\|u_n\|(0)$ (Figures 4(b)–(d)). At larger times, important features of U_I begin to be lost (Figure 4(e)), and for $t \gg T$ the recovered image is very poor (Figure 4(f)).

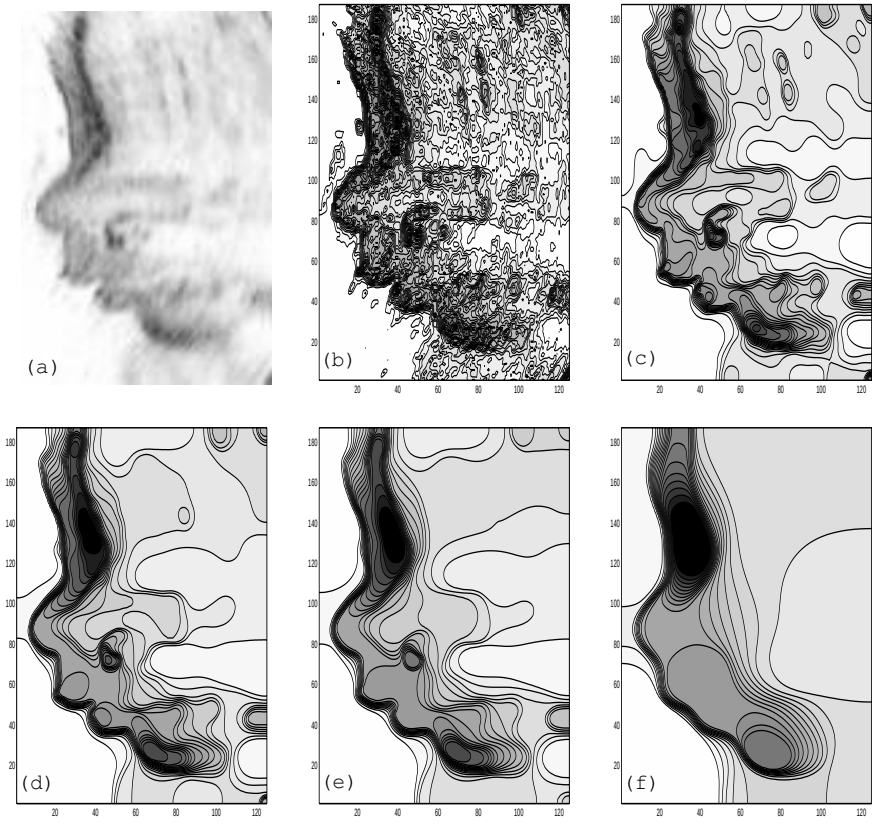


FIG. 5. *Results of applying the diffusion filter to an MR-image;* $\Delta x = \epsilon = 1/124$, $\mu = \epsilon$, $\tau_0 = 1.0$, $k = 2$. (a) *MR-image* $u(\boldsymbol{x}, 0)$, (b) *contour plot of* $u(\boldsymbol{x}, 0)$, (c) $u(\boldsymbol{x}, T)$, (d) $u(\boldsymbol{x}, 2T)$, (e) $u(\boldsymbol{x}, 3T)$, (f) $u(\boldsymbol{x}, 10T)$.

3.2. Example 2. As a second example, we consider an MR-imaged head; see Figure 5(a). Here the initial image is noisy and we can estimate $\|u_n\|(0) \approx 0.2$. For this image, $\Delta x = \epsilon = 1/124$, and we set $\mu = \epsilon$, $\tau_0 = 1.0$, and $k = 2$. Contours of the initial image are shown in Figure 5(b). In Figures 5(c)–(f) the results of applying the diffusion filter with these parameters to the solution $u(\boldsymbol{x}, t)$ at times $t = T,\ 2T,\ 3T,\ 10T$ are shown. On integrating further, the image is effectively

constant by $t \approx 60T$. The contours of the mouth and nose are particularly well-defined for $t \sim T$, but deteriorate rapidly for $t \gg T$.

3.3. Example 3. As a final example, we consider the more complex geometrical image $U_I(\boldsymbol{x})$ in Figure 6(a). This type of image is quite challenging for a diffusion filter due to the fine features, e.g., the edges around the windows, the chair legs, the shadows on the computer screen, etc. In Figure 6(b) we perturb the image with random white noise of maximum amplitude $\pm\Delta u_{n,\max} = 0.1$ and then integrate using $\mu = \epsilon$, $\tau_0 = 1.0$, $k = 2$; here $\Delta x = \epsilon = 1/256$. The recovered image $u(\boldsymbol{x}, t)$ is shown at times $t = T,\ 2T,\ 3T,\ 10T$ in Figures 6(c)–(f). Again for $t = 10T$ many of the fine-scale features of the image are lost, whereas Figures 6(c)–(e) give a reasonable recovered image.

For an image such as $U_I(\boldsymbol{x})$ in Figure 6(a), it is evidently hard to recover fine-scale features when the initial noise level is high, since fine-scale features of the image and noise become indistinguishable. However, the stopping estimate still produces reasonable results, mathematically speaking; it is simply that they are not particularly good aesthetically! In Figure 7 we repeat the numerical experiment of Figure 6, perturbing $U_I(\boldsymbol{x})$ with random white noise of maximum amplitude $\pm\Delta u_{n,\max} = 1.0$, (see Figure 7(a)), and integrating using $\mu = \epsilon$, $\tau_0 = 1.0$, $k = 2$. The recovered image $u(\boldsymbol{x}, t)$ is shown at times $t = T,\ 2T,\ 3T,\ 10T,\ 50T$ in Figures 7(b)–(f). The recovered image is predictably much poorer than in Figure 6. For $t = 50T$ the image has diffused away. Figures 7(c)–(d) are quite reasonable, considering the initial image. (Note that the initial image $u(\boldsymbol{x}, 0)$ is actually much more noisy than it appears in Figure 7(a), since the grey-scale chosen represents only intensities in $[0, 1]$; hence the *salt-and-pepper* look.)

4. Discussion. We have used an analogy between visco-plastic fluid flow and image selective smoothing to determine quantitatively accurate asymptotic stopping time estimates for Bingham diffusion models. The Bingham model is a physical fluid model that, applied to filtering data, is capable of removing the noise locally in a finite time, which, for instance, the Gaussian filtering method ($\tau \to 0$) does not do. The Bingham model is also capable of preserving blocky structure in the original data in the sense that horizontal level lines are preserved as horizontal, although corners are blurred.

There are a number of potential extensions of our work. First, it is important to note that the multiple scales estimate leaves room for the selection of locally varying filter parameters, $\tau_0(\boldsymbol{x})$ and $\mu(\boldsymbol{x})$, which can be chosen within this rational stopping time framework that we have established. This can be used to produce local enhancement of the restored image, particularly in cases where fine-scale features of the image are obscured by noise (e.g., our final example in Figure 7). We have already made some progress in the development of local filtering techniques, which will be reported elsewhere. Second, there are obvious extensions to multidimensional image restoration tasks (e.g., color filtering) wherever Bingham-type filters are used. Third, we mention that the Bingham model is only one of many visco-plastic fluid models. It is of some academic and practical interest to further explore the frontier between non-Newtonian fluid mechanics and image restoration. Currently there are concerns in the image processing community about the use of bounded variation regularization models and their tendency to generate blocky structures (see, e.g., [30, 57]). Recent work of Gousseau and Morel [41] seem to point towards the use of power law models in image denoising, which again have counterparts in fluid mechanics. Filtering methods can be developed for visco-plastic shear-thinning models such as Herschel–Bulkley

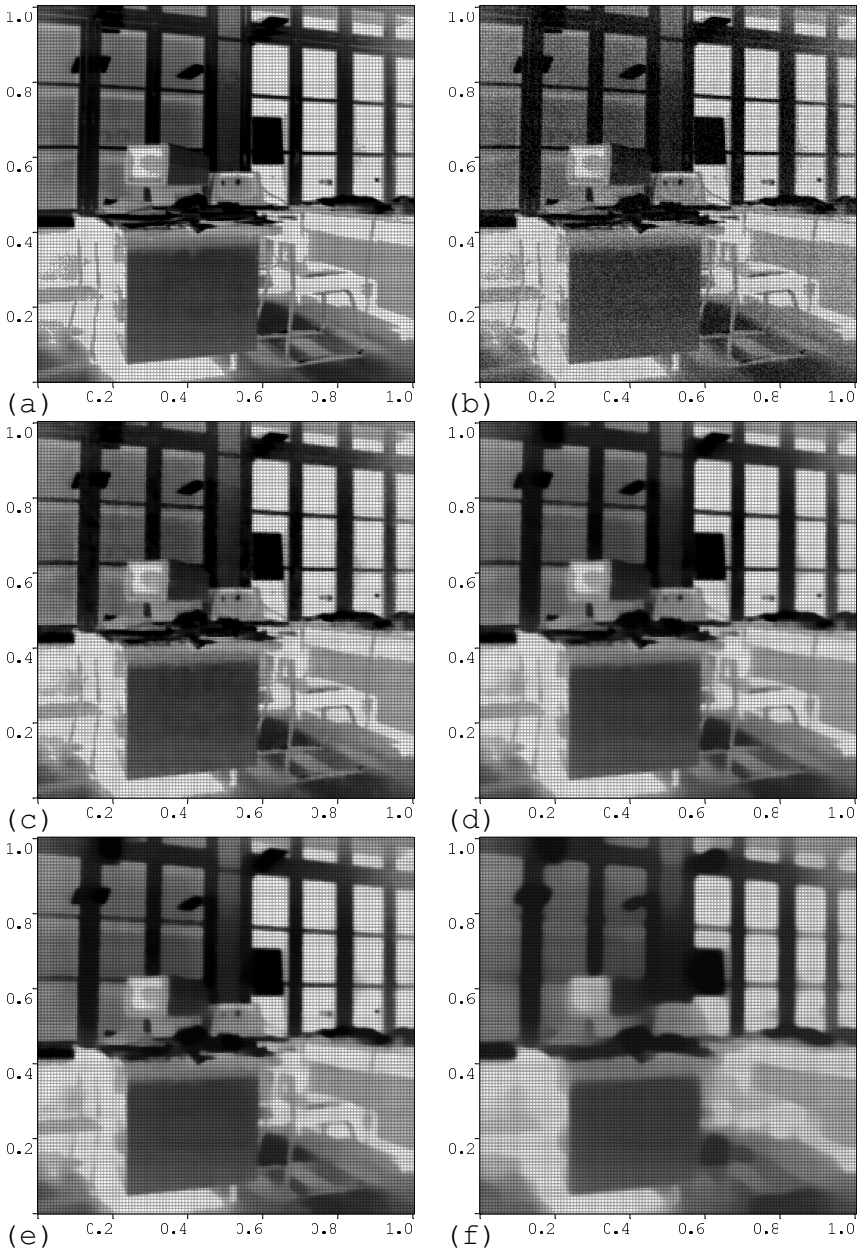


FIG. 6. *Denoising of a more complex geometrical image:* $\Delta x = \epsilon = 1/256$, $\mu = \epsilon$, $\tau_0 = 1.0$, $k = 2$. (a) $U_I(\boldsymbol{x})$, (b) $u(\boldsymbol{x}, 0)$ *after perturbing* $U_I(\boldsymbol{x})$ *with random white noise of maximum amplitude* $\pm\Delta u_{n,\max} = 0.1$, (c) $u(\boldsymbol{x}, T)$, (d) $u(\boldsymbol{x}, 2T)$, (e) $u(\boldsymbol{x}, 3T)$, (f) $u(\boldsymbol{x}, 10T)$.

and Casson fluid models too. The inclusion of visco-elastic (memory) effects is also of extreme practical interest. Our work is progressing in this direction.

Appendix. Weak solutions of (2.14)–(2.17). The differential equation (2.14) has to be understood in a generalized setting. We interpret (2.14) following the concept of nonlinear semigroup theory (see, e.g., [19]) and consider instead a semi-infinite

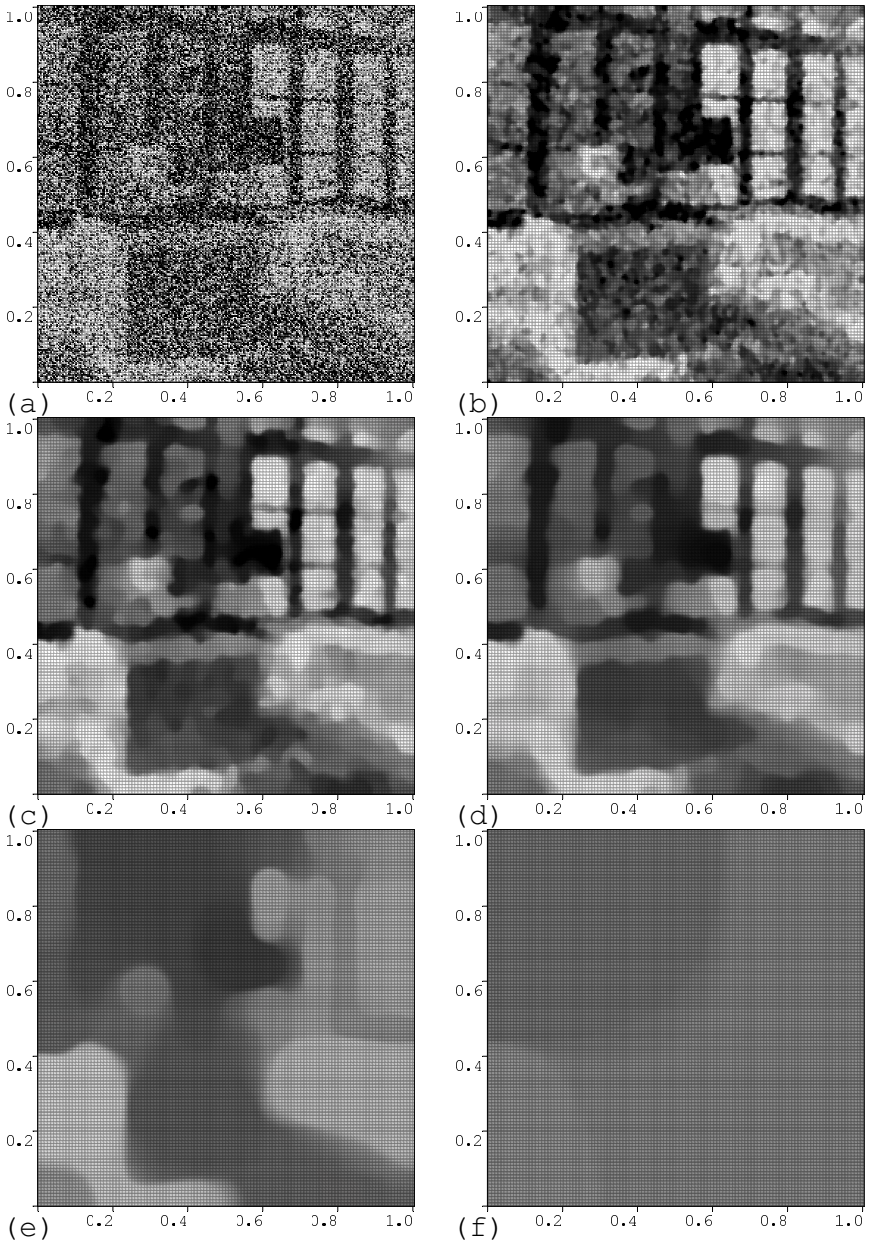


FIG. 7. *Denoising of the image of Figure* 6(a)*:* $\Delta x = \epsilon = 1/256$*,* $\mu = \epsilon$*,* $\tau_0 = 1.0$*,* $k = 2$*.* (a) $u(\boldsymbol{x}, 0)$ *after perturbing* $U_I(\boldsymbol{x})$ *with random white noise of maximum amplitude* $\pm\Delta u_{n,\max} = 1.0$*, (note that the initial image is actually worse than it appears since the grey-scale chosen only represents intensities in* $[0, 1]$*);* (b) $u(\boldsymbol{x}, T)$*;* (c) $u(\boldsymbol{x}, 2T)$*;* (d) $u(\boldsymbol{x}, 3T)$*;* (e) $u(\boldsymbol{x}, 10T)$*;* (f) $u(\boldsymbol{x}, 50T)$*.*

implicit time-discretized approximation $\hat{u}_n$ for $u_n(\tilde{t})$ at the discrete time samples

$$\tilde{t} = \tilde{t}_i = \tilde{t}_{i-1} + \Delta\tilde{t} \quad (i = 1, \dots), \quad \tilde{t}_0 = 0,$$

via

$$\frac{\hat{u}_n(\tilde{t}+\Delta\tilde{t})-\hat{u}_n(\tilde{t})}{\Delta\tilde{t}} = \tilde{\nabla}\cdot[\tilde{\boldsymbol{\tau}}+\epsilon^2\tilde{\boldsymbol{m}}](\tilde{t}+\Delta\tilde{t})+\epsilon^2 f. \tag{A.1}$$

Note here that in our multiple-timescales approximation, partial derivatives of U, and hence also of f, are considered constant in time and space for the rapid timescale and short length-scale evolution associated with $\hat{u}_n$. The approximation at time $\tilde{t}+\Delta\tilde{t}$ is understood as the minimizer of the functional

$$\begin{aligned}\mathcal{F}(\hat{u}) := \quad \Delta\tilde{t}\Bigg\{ &\frac{\mu}{2}a(\hat{u},\hat{u})+\epsilon\tau_0 j(\hat{u})+\epsilon^2\tau_0\left\langle\frac{\nabla U\cdot\tilde{\nabla}\hat{u}}{|\tilde{\nabla}\hat{u}|}\right\rangle-\epsilon^2\langle f\hat{u}\rangle\Bigg\}\\ &-\tfrac{1}{2}\|\hat{u}_n(\tilde{t})-\hat{u}\|^2\end{aligned}$$

over

$$H^1_{\mathtt{mean}} := \left\{\hat{u}\in H^1 : \int\hat{u}=0\right\}.$$

We consider the variational problem of minimizing $\mathcal{F}$ over the space $H^1_{\mathtt{mean}}$. In contrast to classical solutions for (2.14)–(2.18), we do not have to impose Neumann boundary conditions, since they are inherent in the variational formulation. Note that the space $H^1_{\mathtt{mean}}$ is a Hilbert space with scalar product

$$\langle\nabla u\cdot\nabla v\rangle.$$

Since the functional $\mathcal{F}$ is nonconvex with respect to $\hat{u}$ in general we do not have existence of a minimizer (see, e.g., [29]), and generalized solution concepts have to be employed. There are at least three possible solution concepts:

1. Convexification (see, e.g., [29]). This would be the most attractive concept for numerical purposes: it consists of calculating the convex envelope of the function

$$\mathcal{F}^c(\hat{u}_1,\hat{u}_2)=\Delta\tilde{t}\left\{\frac{\mu}{2}(\hat{u}_1^2+\hat{u}_2^2)+\epsilon\tau_0\sqrt{\hat{u}_1^2+\hat{u}_2^2}+\epsilon^2\tau_0\frac{\nabla U\cdot\hat{u}}{\sqrt{\hat{u}_1^2+\hat{u}_2^2}}\right\}$$

with respect to the variable $(\hat{u}_1,\hat{u}_2)\in\mathbb{R}^2$. Once the convex envelope is calculated, the resulting minimization problem can be solved by solving the first order optimality condition, which is a degenerated elliptic partial differential equation. The difficulty associated with calculation of the convex envelope is the dependence of $\mathcal{F}^c$ on ∇U. Thus the convex envelope is only tractable if we find an analytical expression for the convex envelope in terms of two free parameters (U_1,U_2) representing ∇U.

2. Γ-limits (see, e.g., [5]). That is, we reconsider the functional $\mathcal{F}$:

$$\mathcal{F}^\Gamma(\hat{u}) := \liminf_{\{\{\hat{v}_l,\xi_l\}_{l\in\mathbb{N}}:\hat{v}_l\to u\in H^1_{\mathtt{mean}}(\tilde{\Omega}_k)\text{ and }\xi_l\to 0^+\}}\mathcal{F}_{\xi_l}(\hat{v}_l),$$

with

$$\begin{aligned}\mathcal{F}_\xi(\hat{v}_l) := \quad \Delta\tilde{t}\Bigg\{ &\frac{\mu}{2}a(\hat{v}_l,\hat{v}_l)+\epsilon\tau_0 j(\hat{v}_l)-\epsilon^2\langle f\hat{v}_l\rangle\\ &+\epsilon^2\tau_0\left\langle\frac{\nabla U\cdot\tilde{\nabla}\hat{v}_l}{\sqrt{|\tilde{\nabla}\hat{v}_l|^2+\xi^2}}\right\rangle\Bigg\}+\tfrac{1}{2}\|\hat{u}_n(\tilde{t})-v_e\|^2\end{aligned}$$

As we show below, the modified functional has a minimizer. The modified functional is a reasonable model since, if the original functional $\mathcal{F}$ has a minimizer, it also minimizes $\mathcal{F}^\Gamma$. We remark that for convex functionals the modified functional is identical to the original. We employ Γ^--limits below.

3. Relaxation (see, e.g., [5, 29]). We call a relaxation method a Γ-limit, without additional parameter dependency, i.e., without dependency of the positive sequence $\{\xi_l\}_{l\in\mathbb{N}}$.

In order to simplify the considerations, we abbreviate the functional $\mathcal{F}$ and note that it is of the general form

$$\mathcal{F}(\hat{u}) := \nu_1 a(\hat{u},\hat{u}) + \nu_2 j(\hat{u}) + \nu_3 \left\langle \frac{\nabla U\cdot\tilde{\nabla}\hat{u}}{|\tilde{\nabla}\hat{u}|}\right\rangle - \left\langle \hat{f}\hat{u}\right\rangle + \nu_4\|\hat{u}_n(\tilde{t})-\hat{u}\|^2,$$

with $0<\nu:=(\nu_1,\nu_2,\nu_3,\nu_4)\le\overline{\nu}$.

Accordingly we set

$$\mathcal{F}_\xi(\hat{u}) := \nu_1 a(\hat{u},\hat{u}) + \nu_2 j(\hat{u}) + \nu_3 \left\langle \frac{\nabla U\cdot\tilde{\nabla}\hat{u}}{\sqrt{|\tilde{\nabla}\hat{u}|^2+\xi^2}}\right\rangle - \left\langle \hat{f}\hat{u}\right\rangle + \nu_4\|\hat{u}_n(\tilde{t})-\hat{u}\|^2$$

and define the Γ-limit as above.

LEMMA A.1. *Let $\hat{f}\in(H^1_{\mathtt{mean}})^*$, the dual of $H^1_{\mathtt{mean}}$, and $U\in H^1$. Moreover, let $\tilde{\Omega}_k$ be bounded with Lipschitz boundary. Then the function values $\mathcal{F}^\Gamma(\hat{u})$ are well defined for each $\hat{u}\in H^1_{\mathtt{mean}}$.*

Proof. Under the above assumptions, we have for each $\tilde{u}\in H^1_{\mathtt{mean}}$ and $\xi>0$ that

$$\begin{aligned}\mathcal{F}_\xi(\tilde{u}) &\ge -\overline{\nu}\|\nabla U\|_{L^2} + \nu_1\|\tilde{\nabla}\tilde{u}\|^2_{L^2} - \|\hat{f}\|_{(H^1_{\mathtt{mean}})^*}\|\tilde{\nabla}\tilde{u}\|_{L^2}\\ &\ge -\overline{\nu}\|\nabla U\|_{L^2} - \frac{\|\hat{f}\|^2_{(H^1_{\mathtt{mean}})^*}}{4\nu_1}\,.\end{aligned} \tag{A.2}$$

This shows that for any $\tilde{u}\in H^1_{\mathtt{mean}}$, $\mathcal{F}_\xi(\tilde{u})$ is uniformly bounded from below with respect to ξ. Moreover, $\mathcal{F}_\xi(\tilde{u})$ is uniformly bounded from above since

$$\mathcal{F}_\xi(\tilde{u}) \le \overline{\nu}\|\nabla U\|_{L^2} + C\|\tilde{u}\|^2_{H^1_{\mathtt{mean}}} + \|\hat{f}\|_{(H^1_{\mathtt{mean}})^*}\|\tilde{u}\|_{H^1_{\mathtt{mean}}} + \overline{\nu}\|\tilde{u}\|_{H^1_{\mathtt{mean}}}\sqrt{\mathrm{meas}(\tilde{\Omega}_k)} + C, \tag{A.3}$$

with appropriate $C>0$. Thus for any pair of sequences $\{\hat{v}_l\}_{l\in\mathbb{N}}$ converging weakly to $\hat{u}$, and $\{\xi_l\}_{l\in\mathbb{N}}$ converging to $0+$, the sequence $\{\mathcal{F}_{\xi_l}(\hat{v}_l)\}_{l\in\mathbb{N}}$ is uniformly bounded, and thus its infimum is finite. Taking the infimum over all limits of possible sequences $\{\mathcal{F}_{\xi_l}(\hat{v}_l)\}$ gives the assertion.

THEOREM A.2. *Under the same assumptions as in Lemma* A.1, *$\mathcal{F}^\Gamma$ attains a minimum in $H^1_{\mathtt{mean}}$.*

Proof. From (A.2) it follows that the infimum of $\mathcal{F}^\Gamma$ over $H^1_{\mathtt{mean}}$ is finite. Suppose that $\mathcal{F}^\Gamma$ does not attain a minimum; then there exists a sequence $\{\rho_l\}_{l\in\mathbb{N}}$ satisfying

$$\mathcal{F}^\Gamma(\rho_l)\to\inf\mathcal{F}^\Gamma.$$

By definition, for any ρ_l and any $\delta_l>0$ there exists a pair $(\hat{\rho}_l,\xi_l)$ satisfying

$$|\mathcal{F}^\Gamma(\rho_l)-\mathcal{F}_{\xi_l}(\hat{\rho}_l)|\le\delta_l.$$

Let $\delta_l \to 0$; then

$$\mathcal{F}_{\xi_l}(\hat{\rho}_l) \to \inf \mathcal{F}^\Gamma \,. \tag{A.4}$$

In particular, we have

$$\mathcal{F}_{\xi_l}(\hat{\rho}_l) \le \overline{C} < \infty \quad \text{ for all } l \in \mathbb{N} \,.$$

Thus from the definition of $\mathcal{F}_{\xi_l}$ it follows that

$$\begin{aligned} \overline{\nu}\|\hat{\rho}_l\|^2_{H^1_{\mathtt{mean}}} &\le \left(\overline{C} + \overline{\nu}\|\nabla U\|_{L^2}\right) + \left(\|\hat{f}\|_{(H^1_{\mathtt{mean}})^*} + \overline{\nu}\sqrt{\mathrm{meas}(\tilde{\Omega}_k)}\right)\|\hat{\rho}_l\|_{H^1_{\mathtt{mean}}} \\ &=: C_1 + C_2\|\hat{\rho}_l\|_{H^1_{\mathtt{mean}}}, \end{aligned}$$

which shows that $\{\hat{\rho}_l\}_{l\in\mathbb{N}}$ is uniformly bounded in $H^1_{\mathtt{mean}}$, and therefore it has a weakly convergent subsequence, which again, for simplicity of notation, is denoted by $\{\hat{\rho}_l\}_{l\in\mathbb{N}}$. Let us denote by $\hat{u}$ the weak limit; then by definition of $\mathcal{F}^\Gamma$ we have that

$$\mathcal{F}^\Gamma(\hat{u}) \le \liminf_{l\in\mathbb{N}} \mathcal{F}_{\xi_l}(\hat{\rho}_l),$$

which together with (A.4) shows that

$$\mathcal{F}^\Gamma(\hat{u}) = \inf \mathcal{F}^\Gamma.$$

Acknowledgments. The authors are grateful to two referees for their careful reading of the manuscript, their helpful advice, and bringing to our attention the references [8, 10, 12, 57]. Finally we thank Prof. F. Santosa for encouraging us to include subsection 2.2.2.

REFERENCES

[1] R. Acar and C.R. Vogel, *Analysis of bounded variation penalty methods for ill-posed problems,* Inverse Problems, 10 (1994), pp. 1217–1229.
[2] L. Alvarez, F. Guichard, P.-L. Lions, and J.-M. Morel, *Axioms and fundamental equations of image processing,* Arch. Ration. Mech. Anal., 123 (1993), pp. 199–257.
[3] L. Alvarez, P.-L. Lions, and J.-M. Morel, *Image selective smoothing and edge detection by nonlinear diffusion.* II, SIAM J. Numer. Anal., 29 (1992), pp. 845–866.
[4] L. Alvarez and J.-M. Morel, *Formalization and computational aspects of image analysis,* Acta Numer. 1994, Cambridge University Press, London, 1994, pp. 1–59.
[5] L. Ambrosio, *Geometric evolution problems, distance function and viscosity solutions,* in Calculus of Variations and Partial Differential Equations, Springer-Verlag, New York, 1999, pp. 5–94.
[6] L. Ambrosio and N. Dancer, *Calculus of Variations and Partial Differential Equations*, Springer-Verlag, New York, 1999.
[7] F. Andreu, C. Ballester, V. Caselles, and J.M. Mazón, *Minimizing total variation flow,* C. R. Acad. Sci. Paris Sér. I Math., 331 (2000), pp. 867–872.
[8] F. Andreu, C. Ballester, V. Caselles, and J.M. Mazón, *The Dirichlet problem for the total variation flow,* J. Funct. Anal., 180 (2001), pp. 347–403.
[9] F. Andreu, C. Ballester, V. Caselles, and J.M. Mazón, *Minimizing total variation flow,* Differential Integral Equations, 14 (2001), pp. 321–360.
[10] F. Andreu, V. Caselles, J.I. Díaz, and J.M. Mazón, *Some qualitative properties for the total variation flow,* J. Funct. Anal., 188 (2002), pp. 516–547.
[11] G. Aubert, R. Deriche, and P. Kornprobst, *Computing optical flow via variational techniques,* SIAM J. Appl. Math., 60 (1999), pp. 156–182.
[12] G. Aubert and L. Vese, *A variational method in image recovery,* SIAM J. Numer. Anal., 34 (1997), pp. 1948–1979.

[13] N.J. Balmforth and R.V. Craster, *A consistent thin-layer theory for Bingham fluids,* J. Non-Newtonian Fluid Mech., 84 (1999), pp. 65–81.
[14] N.J. Balmforth and R.V. Craster, *Dynamics of cooling domes of visco-plastic fluid,* J. Fluid Mech., 422 (2000), pp. 225–248.
[15] G.I. Barenblatt, V.M. Entov, and V.M. Ryzhik, *Theory of Fluid Flows Through Natural Rocks*, Kluwer Academic Publishers, Nowell, MA, 1990.
[16] G. Bellettini, V. Caselles, and M. Novaga, *The Total Variation Flow in* $\mathbb{R}^N$, J. Differential Equations, 184 (2002), pp. 475–525.
[17] E.C. Bingham, *Fluidity and Plasticity*, McGraw–Hill, New York, 1922.
[18] M.J. Black and P. Anandan, *The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields,* Comput. Vision Image Understanding, 63 (1996), pp. 75–104.
[19] H. Brezis, *Operateurs Maximaux Monotones et Semi-Groupes de Contractions dans les Espaces de Hilbert*, North–Holland, Amsterdam, 1973.
[20] R. Byron-Bird, G.C. Dai, and B.J. Yarusso, The rheology and flow of viscoplastic materials, Rev. Chem. Engrg., 1 (1983), pp. 2–70.
[21] F. Catté, P.-L. Lions, J.-M. Morel, and T. Coll, *Image selective smoothing and edge detection by nonlinear diffusion,* SIAM J. Numer. Anal., 29 (1992), pp. 182–193.
[22] A. Chambolle and P.L. Lions, *Image recovery via total variation minimization and related problems,* Numer. Math., 76 (1997), pp. 167–188.
[23] T.F. Chan, G.H. Golub, and P. Mulet, *A nonlinear primal-dual method for total variation-based image restoration,* SIAM J. Sci. Comput., 20 (1999), pp. 1964–1977.
[24] T.F. Chan and P. Mulet, *On the convergence of the lagged diffusivity fixed point method in total variation image restoration,* SIAM J. Numer. Anal., 36 (1999), pp. 354–367.
[25] T.F. Chan, G.H. Golub, and P. Mulet, *A nonlinear primal-dual method for total variation-based image restoration,* in Proceedings of the 12th International Conference on Analysis and Optimization of Systems, Images, Wavelets and PDEs (ICAOS '96), Berlin, 1996, Lecture Notes in Control and Inform. Sci. 219, Springer-Verlag, London, 1996, pp. 241–252.
[26] G. Chavent and K. Kunisch, *Regularization of linear least squares problems by total bounded variation,* ESAIM Control Optim. Calc. Var., 2 (1997), pp. 359–376.
[27] M. Chipot, R. March, M. Rosati, and G. Vergara Caffarelli, *Analysis of a nonconvex problem related to signal selective smoothing,* Math. Models Methods Appl. Sci., 7 (1997), pp. 313–328.
[28] A. Cohen, R. DeVore, P. Petrushev, and H. Xu, *Nonlinear approximation and the space* $BV(\mathbb{R}^2)$, Amer. J. Math., 121 (1999), pp. 587–628.
[29] B. Dacorogna, *Direct Methods in the Calculus of Variations*, Springer-Verlag, Berlin, 1989.
[30] D.C. Dobson and F. Santosa, *Recovery of blocky images from noisy and blurred data,* SIAM J. Appl. Math., 56 (1996), pp. 1181–1198.
[31] D.C. Dobson and O. Scherzer, *Analysis of regularized total variation penalty methods for denoising,* Inverse Problems, 12 (1996), pp. 601–617.
[32] D.C. Dobson and C.R. Vogel, *Convergence of an iterative method for total variation denoising,* SIAM J. Numer. Anal., 34 (1997), pp. 1779–1791.
[33] G. Duvaut and J.L. Lions, *Inequalities in Mechanics and Physics*, Springer-Verlag, New York, 1976.
[34] I.A. Frigaard and O. Scherzer, *Uniaxial exchange flows of two Bingham fluids in a cylindrical duct,* IMA J. Appl. Math., 61 (1998), pp. 237–266.
[35] I.A. Frigaard and O. Scherzer, *The effects of yield stress variation on uniaxial exchange flows of two Bingham fluids in a pipe,* SIAM J. Appl. Math., 60 (2000), pp. 1950–1976.
[36] I.A. Frigaard, S.D. Howison, and I.J. Sobey, *On the stability of Poiseuille flow of a Bingham fluid,* J. Fluid Mech., 263 (1994), pp. 133–150.
[37] D. Geman and C. Yang, *Nonlinear image recovery with half-quadratic regularization,* IEEE Trans. Image Processing, 4 (1995), pp. 932–945.
[38] S. Geman and D. Geman, *Stochastic relaxation, Gibbs distributions, and the Bayseian restoration of images,* IEEE Trans. Pattern Analysis and Machine Intelligence, 6 (1984), pp. 721–741.
[39] R. Glowinski, *Numerical Methods for Nonlinear Variational Problems*, Springer-Verlag, Berlin, New York, 1984.
[40] R. Glowinski, J.-L. Lions, and R. Trémolières, *Numerical Analysis of Variational Inequalities*, North–Holland, Amsterdam, 1981.
[41] Y. Gousseau and J.-M. Morel, *Are natural images of bounded variation?,* SIAM J. Math. Anal., 33 (2001), pp. 634–648.
[42] E.J. Hinch, *Perturbation Methods*, Cambridge University Press, Cambridge, UK, 1991.

[43] R.R. Huilgol and M.P. Panizza, On the determination of the plug flow region in Bingham fluids through the application of variational inequalities, J. Non-Newtonian Fluid Mech., 58 (1995), pp. 207–217.
[44] K. Ito and K. Kunisch, *Lagrangian formulation of nonsmooth convex optimization in Hilbert spaces,* in Control of Partial Differential Equations and Applications, Proceedings of the 17th IFIP TC7 Conference on System Modelling and Optimization, E. Casas, ed., Lecture Notes in Pure and Appl. Math. 174, Dekker, New York, 1996, pp. 107–117.
[45] K. Ito and K. Kunisch, *Augmented Lagrangian methods for nonsmooth, convex optimization in Hilbert spaces,* Nonlinear Anal., 41A (2000), pp. 591–616.
[46] S. Kichenassamy, *The Perona–Malik paradox,* SIAM J. Appl. Math., 57 (1997), pp. 1328–1342.
[47] K.F. Liu and C.C. Mei, *Slow spreading of a sheet of Bingham fluid on an inclined plane*, J. Fluid Mech., 207 (1989), pp. 505–529.
[48] S. Mallat, *A Wavelet Tour of Signal Processing,* 2nd ed., Academic Press, San Diego, CA, 1999.
[49] J.M. Morel and S. Solimini, *Variational Methods in Image Segmentation*, Birkhäuser Boston, Cambridge, MA, 1995.
[50] P.P. Mossolov and V.P. Miasnikov, *Variational methods in the theory of the fluidity of a viscous plastic medium,* J. Mech. Appl. Math., 29 (1965), pp. 468–492.
[51] N.H. Nagel, *On the estimation of optical flow: Relations between new approaches and some new results,* Artificial Intelligence, 33 (1987), pp. 299–324.
[52] N.H. Nagel and W. Enkelmann, *An investigation of smoothness constraints for the estimation of displacement vector fields from image sequences,* IEEE Trans. Pattern Anal. Machine Intelligence, 12 (1990), pp. 629–629.
[53] M.Z. Nashed and O. Scherzer, *Stable approximations of nondifferentiable optimization problems with variational inequalities,* Contemp. Math., 204 (1997), pp. 155–170.
[54] M.Z. Nashed and O. Scherzer, *Least squares and bounded variation regularization with nondifferentiable functional,* Numer. Funct. Anal. Optim., 19 (1998), pp. 873–901.
[55] M. Nielsen, L. Florack, and R. Deriche, *Regularization and scale space,* INRIA preprint series, 2352 (1994), pp. 1–39.
[56] M. Nielsen, P. Johansen, O.F. Olsen, and J. Weickert, eds., Scale-Space Theories in Computer Vision, Lecture Notes in Comput. Sci. 1682, Springer-Verlag, Berlin, 1999.
[57] M. Nikolova, *Local strong homogeneity of a regularized estimator,* SIAM J. Appl. Math., 61 (2000), pp. 633–658.
[58] N. Nordström, *Biased anisotropic diffusion—A unified regularization and diffusion approach to edge detection,* Image Vision Comput., 8 (1990), pp. 318–327.
[59] J.G. Oldroyd, *Two-dimensional plastic flow of a Bingham solid,* Proc. Camb. Phil. Soc., 43 (1947), pp. 383–395.
[60] M.E. Oman and C. Vogel, *Fast numerical methods for total variation minimization in image reconstruction,* in Advanced Signal Processing Algorithms, SPIE Proceedings Vol. 256, SPIE, Bellingham, WA, 1995.
[61] M.E. Oman and C. Vogel, *Fast total variation-based image reconstruction,* in Proceedings of the 1995 ASME Design Engineering Conferences, Vol. 3, ASME, New York, 1995, pp. 1009–1015.
[62] S. Osher and L.I. Rudin, Feature-oriented image enhancement using shock filters, SIAM J. Numer. Anal., 27 (1990), pp. 919–940.
[63] P. Perona and J. Malik, *Scale space and edge detection using anisotropic diffusion,* IEEE Trans. Pattern Anal. Machine Intelligence, 12 (1990), pp. 629–639.
[64] W. Prager, *On slow visco-plastic flow,* in Studies in Mathematics and Mechanics Presented to Richard von Mises, Academic Press, New York, 1954, pp. 208–216.
[65] E. Radmoser, O. Scherzer, and J. Weickert, *Scale-space properties of regularization methods,* in Scale-Space Theories in Computer Vision, M. Nielsen, P. Johansen, O.F. Olsen, and J. Weickert, eds., Lecture Notes in Comput. Sci. 1682, Springer-Verlag, Berlin, 1999, pp. 211–222.
[66] E. Radmoser, O. Scherzer, and J. Weickert, *Scale-space properties of nonstationary iterative regularization methods,* J. Visual Commun. Image Representation, 11 (2000), pp. 96–114.
[67] A.B. Ross, S.K. Wilson, and B.R. Duffy, *Thin-film flow of a viscoplastic material round a large horizontal stationary or rotating cylinder,* J. Fluid Mech., 430 (2001), pp. 309–333.
[68] L.I. Rudin, St. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms,* Phys. D, 60 (1992), pp. 259–268.
[69] O. Scherzer, *Stable evaluation of differential operators and linear and nonlinear multi-scale filtering,* Electronic J. Differential Equations, 15 (1997), pp. 1–12.

[70] O. Scherzer and J. Weickert, *Relations between regularization and diffusion filtering,* J. Math. Imag. Vision, 12 (2000), pp. 43–63.
[71] D. Strong and T.F. Chan, *Exact Solutions to the Total Variation Regularization Problem,* Technical report CAM 96-41, University of California, Los Angeles, 1996.
[72] P. Szabo and O. Hassager, *Flow of viscoplastic fluids in eccentric annular geometries,* J. Non-Newtonian Fluid Mech., 45 (1992), pp. 149–169.
[73] C. Vogel, *A multigrid method for total variation-based image denoising,* in Computation and Control IV, Progr. Systems Control Theory 20, K. Bowers and J. Lund, eds., Birkhäuser, Boston, Cambridge, MA, 1995.
[74] C.R. Vogel and M.E. Oman, *Iterative methods for total variation denoising,* SIAM J. Sci. Comput., 17 (1996), pp. 227–238.
[75] I.C. Walton and S.H. Bittleston, *The axial flow of a Bingham plastic in a narrow eccentric annulus*, J. Fluid Mech., 222 (1991), pp. 39–60.
[76] J. Weickert, *Anisotropic Diffusion in Image Processing*, Teubner, Stuttgart, 1998.
[77] S.D.R. Wilson, *Squeezing flow of a Bingham material,* J. Non-Newtonian Fluid Mech., 47 (1993), pp. 211–219.
[78] S.D.R. Wilson and A.J. Taylor, *The channel entry problem for a yield-stress fluid,* J. Non-Newtonian Fluid Mech., 65 (1996), pp. 165–176.

SIAM J. APPL. MATH.
Vol. 63, No. 6, pp. 1935–1953

SLOWLY COUPLED OSCILLATORS: PHASE DYNAMICS AND SYNCHRONIZATION*

EUGENE M. IZHIKEVICH† AND FRANK C. HOPPENSTEADT‡

Abstract. In this paper we extend the results of Frankel and Kiemel [*SIAM J. Appl. Math*, 53 (1993), pp. 1436–1446] to a network of slowly coupled oscillators. First, we use Malkin's theorem to derive a canonical phase model that describes synchronization properties of a slowly coupled network. Then, we illustrate the result using slowly coupled oscillators (1) near Andronov–Hopf bifurcations, (2) near saddle-node on invariant circle bifurcations, and (3) near relaxation oscillations. We compare and contrast synchronization properties of slowly and weakly coupled oscillators.

Key words. phase model, Andronov–Hopf, saddle-node on invariant circle, Class 1 excitability, relaxation oscillators, Malkin theorem, MATLAB

AMS subject classifications. 92B20, 92C20, 82C32, 58Fxx, 34Cxx, 34Dxx

DOI. 10.1137/S0036139902400945

1. Slowly coupled networks.

In this paper we study synchronization dynamics of a network of $n \geq 2$ coupled oscillators of the form

$$\dot{x}_i = f_i(x_i, s_1, \ldots, s_n), \tag{1.1}$$

$$\dot{s}_i = \varepsilon g_i(x_i, s_i), \tag{1.2}$$

where $x_i \in \mathbb{R}^m$ describes the state of the ith oscillator and $s_i \in \mathbb{R}$ describes how the ith oscillator affects the other oscillators for $i = 1, \ldots, n$. The parameter $\varepsilon \ll 1$ is small reflecting the assumption that the connection variables s_i are "slow." We analyze this system in this section and present several examples in section 2.

Proceeding as in Frankel and Kiemel (1993) we "freeze" the vector of slow variables $s = (s_1, \ldots, s_n) \in \mathbb{R}^n$ and assume that each oscillator described by (1.1) has a $T_i(s)$-periodic solution $x_i = x_i(t, s)$. Substituting this in (1.2) results in the system

$$\dot{s}_i = \varepsilon g_i(x_i(t, s), s_i), \qquad i = 1, \ldots, n,$$

which we can average and obtain a slow system

$$\dot{s}_i = \varepsilon \bar{g}_i(s), \qquad i = 1, \ldots, n, \tag{1.3}$$

where

$$\bar{g}_i(s) = \frac{1}{T_i(s)} \int_0^{T_i(s)} g_i(x_i(t, s), s_i)\, dt$$

is the average of g_i. In this paper we make the following two assumptions:

*Received by the editors January 14, 2002; accepted for publication (in revised form) February 28, 2003; published electronically September 4, 2003. This research was supported in part by NSF grant DMS-0109001.

http://www.siam.org/journals/siap/63-6/40094.html

†The Neurosciences Institute, 10640 John Jay Hopkins Drive, San Diego, CA 92121 (Eugene. Izhikevich@nsi.edu, http://www.izhikevich.com). The research of this author was carried out as part of the theoretical neurobiology program at The Neurosciences Institute, which is supported by the Neurosciences Research Foundation.

‡Systems Science and Engineering Research Center, Arizona State University, Tempe, AZ 85287-7606 (fchoppen@asu.edu.).

A1. The system (1.3) has an exponentially stable equilibrium $\bar{s} = (\bar{s}_1, \dots, \bar{s}_n)$.
A2. Each equation (1.1) has an exponentially stable limit cycle attractor $\gamma_i(t) \subset \mathbb{R}^m$ with period $T > 0$ when $s = \bar{s}$.

THEOREM 1.1 (phase model for slowly coupled oscillators). *Consider the slowly coupled system* (1.1), (1.2) *satisfying assumptions* A1 *and* A2 *above. Let* $\tau = \varepsilon t$ *be slow time. Let* $u_i(\tau)$ *be the rescaled deviation of the slow variable* s_i *from the asymptotic value* $\bar{s}_i$*, and let* $\varphi_i(\tau)$ *be the phase deviation of the* i*th oscillator from the natural oscillation* $\gamma_i(t)$*. Then, the phase dynamics and synchronization properties of the slowly coupled system* (1.1), (1.2) *are described by the canonical phase model*

$$\varphi_i' = \sum_{j=1}^{n} \{a_{ij} u_j + H_{ij}(\varphi_j - \varphi_i)\}, \tag{1.4}$$

$$u_i' = \sum_{j=1}^{n} \{b_{ij} u_j + K_{ij}(\varphi_j - \varphi_i)\}, \tag{1.5}$$

where $' = d/d\tau$*, and*

$$\begin{aligned}
a_{ij} &= \frac{1}{T}\int_0^T Q_i(t)^\top \frac{\partial f_i}{\partial s_j}(\gamma_i(t), \bar{s})\, dt,\\
b_{ij} &= \frac{1}{T}\int_0^T \left[P_i(t)^\top \frac{\partial f_i}{\partial s_j}(\gamma_i(t), \bar{s}) + \frac{\partial g_i}{\partial s_j}(\gamma_i(t), \bar{s}_i)\right] dt,\\
H_{ij}(\chi) &= \frac{1}{T}\int_0^T Q_i(t)^\top \frac{\partial f_i}{\partial s_j}(\gamma_i(t), \bar{s}) \int_0^{t+\chi} g_j(\gamma_j(\bar{t}), \bar{s})\, d\bar{t}\, dt,\\
K_{ij}(\chi) &= \frac{1}{T}\int_0^T \left(P_i(t)^\top \frac{\partial f_i}{\partial s_j}(\gamma_i(t), \bar{s}) + \frac{\partial g_i}{\partial s_j}(\gamma_i(t), \bar{s}_i)\right) \int_0^{t+\chi} g_j(\gamma_j(\bar{t}), \bar{s})\, d\bar{t}\, dt,
\end{aligned}$$

where $Q_i(t), P_i(t) \subset \mathbb{R}^m$ *are the unique nontrivial* T*-periodic solutions of the linear adjoint systems*

$$\dot{Q}_i = -\left\{\frac{\partial f_i}{\partial x_i}(\gamma_i(t), \bar{s})\right\}^\top Q_i \quad \textit{and} \quad \dot{P}_i = -\left\{\frac{\partial f_i}{\partial x_i}(\gamma_i(t), \bar{s})\right\}^\top P_i - \left\{\frac{\partial g_i}{\partial x_i}(\gamma_i(t), \bar{s})\right\}^\top \tag{1.6}$$

satisfying the normalization conditions

$$Q_i(t)^\top f_i(\gamma_i(t), \bar{s}) = 1 \qquad \textit{and} \qquad P_i(t)^\top f_i(\gamma_i(t), \bar{s}) = -g_i(\gamma_i(t), \bar{s}) \tag{1.7}$$

for some (and hence all) $t \geq 0$*.*

Remark 1.2. The same result holds when $g_i(x_i, s_i)$ also depend on $s_1, \dots, s_n$.

Remark 1.3. The same result holds for the *slowly and weakly* coupled system

$$\begin{aligned}
\dot{x}_i &= f_i(x_i, s_1, \dots, s_n) + \varepsilon \sum_{j=1}^{n} r_{ij}(x_i, x_j),\\
\dot{s}_i &= \varepsilon g_i(x_i, s_i),
\end{aligned}$$

provided that the term

$$\frac{1}{T}\int_0^T Q_i(t)^\top r_{ij}(\gamma_i(t), \gamma_j(t+\chi))\, dt$$

is added to the function $H_{ij}(\chi)$ and the term

$$\frac{1}{T}\int_0^T P_i(t)^\top r_{ij}(\gamma_i(t),\gamma_j(t+\chi))\,dt$$

is added to the function $K_{ij}(\chi)$.

Remark 1.4. This result not only extends the result of Frankel and Kiemel (1993) to a network of $n \geq 2$ oscillators, but also presents a precise description of all the parameters and functions in the canonical model (1.4), (1.5), which can easily be determined numerically; see Appendix B.

Proof. This result is a corollary to Malkin's theorem, which we restate in Appendix A. Consider the slowly coupled system (1.1), (1.2) in an ε-neighborhood of $\bar{s}$. Let

$$s_i = \bar{s}_i + \varepsilon w_i,$$

so that we can rewrite (1.1), (1.2) in the form (A.1),

$$\dot{x}_i = f_i(x_i,\bar{s}) + \varepsilon\sum_{j=1}^n h_{ij}(x_i,\bar{s})w_j, \tag{1.8}$$

$$\dot{w}_i = g_i(x_i,\bar{s}_i) + \varepsilon p_i(x_i,\bar{s}_i)w_i \tag{1.9}$$

plus higher-order terms in ε, where

$$h_{ij} = \frac{\partial f_i}{\partial s_j} \qquad \text{and} \qquad p_i = \frac{\partial g_i}{\partial s_i}\ .$$

In the rest of the proof we omit $\bar{s}$ for the sake of clarity of notation.

Since (1.8), (1.9) has a "weakly connected" form, it suffices to show that all the conditions of Malkin's theorem are satisfied for each individual oscillator.

Each unperturbed (uncoupled, $\varepsilon = 0$) system

$$\dot{x}_i = f_i(x_i),$$
$$\dot{w}_i = g_i(x_i)$$

has a 2-parameter family of T-periodic solutions

$$x_i(t) = \gamma_i(t+\varphi_i) \qquad \text{and} \qquad w_i(t) = u_i + \int_0^{t+\varphi_i} g_i(\gamma_i(\bar{t}))\,d\bar{t},$$

where φ_i and u_i are independent parameters, $i = 1,\ldots,n$. ($w_i(t)$ is periodic because s_i is at equilibrium and the average of g_i over the limit cycle is assumed to be zero.) Let $Q_i(t)$ and $P_i(t)$ be the unique nontrivial solutions to the adjoint system (1.6), which exist because $\gamma_i(t)$ is a normally hyperbolic attractor (Hoppensteadt and Izhikevich 1997). One can verify that

$$R_i(t) = \begin{pmatrix} Q_i(t) \\ 0 \end{pmatrix} \qquad \text{and} \qquad R_i(t) = \begin{pmatrix} P_i(t) \\ 1 \end{pmatrix}$$

are two independent nontrivial solutions to the adjoint system (A.2),

$$\dot{R}_i = -\begin{pmatrix} \frac{\partial f_i}{\partial x_i}(\gamma_i(t)) & 0 \\ \frac{\partial g_i}{\partial x_i}(\gamma_i(t)) & 0 \end{pmatrix}^\top R_i\ .$$

Equation (A.4) results in

$$\varphi_i' = \frac{1}{T}\int_0^T Q_i(t+\varphi_i)^\top \left\{ \sum_{j=1}^n h_{ij}(\gamma_i(t+\varphi_i)) \left(u_j + \int_0^{t+\varphi_j} g_j(\gamma_j(\bar{t}))\, d\bar{t} \right) \right\} dt$$

$$= \frac{1}{T}\int_0^T Q_i(t)^\top \left\{ \sum_{j=1}^n h_{ij}(\gamma_i(t)) \left(u_j + \int_0^{t+\varphi_j-\varphi_i} g_j(\gamma_j(\bar{t}))\, d\bar{t} \right) \right\} dt$$

and

$$u_i' = \frac{1}{T}\int_0^T P_i(t+\varphi_i)^\top \left\{ \sum_{j=1}^n h_{ij}(\gamma_i(t+\varphi_i)) \left(u_j + \int_0^{t+\varphi_j} g_j(\gamma_j(\bar{t}))\, d\bar{t} \right) \right\}$$

$$+ 1 \cdot \left\{ p_i(\gamma_i(t+\varphi_i)) \left(u_i + \int_0^{t+\varphi_i} g_i(\gamma_i(\bar{t}))\, d\bar{t} \right) \right\} dt$$

$$= \frac{1}{T}\int_0^T P_i(t)^\top \left\{ \sum_{j=1}^n h_{ij}(\gamma_i(t)) \left(u_j + \int_0^{t+\varphi_j-\varphi_i} g_j(\gamma_j(\bar{t}))\, d\bar{t} \right) \right\}$$

$$+ 1 \cdot \left\{ p_i(\gamma_i(t)) \left(u_i + \int_0^{t+\varphi_j-\varphi_i} g_i(\gamma_i(\bar{t}))\, d\bar{t} \right) \right\} dt,$$

which can be written in the form (1.4), (1.5). □

2. Examples. The major challenge in applying Theorem 1.1 is solving the linear adjoint system (1.6). In general, this could be done numerically, as we show in Appendix B. However, there are three important cases when (1.6) can be solved analytically:

- Each oscillator is near an Andronov–Hopf bifurcation.
- Each oscillator is near a saddle-node on invariant circle bifurcation.
- Each oscillator has two time scales (relaxation oscillator).

We consider all three cases below, but first we start with two simple examples.

2.1. Phase oscillators. The system of slowly coupled phase oscillators

$$\dot{\vartheta}_i = 1 + \sum_{j=1}^n c_{ij} s_j, \tag{2.1}$$

$$\dot{s}_i = \varepsilon(\cos\vartheta_i - b s_i) \tag{2.2}$$

illustrates the major steps in Theorem 1.1. When all s_i are not very large, the averaged slow system (1.3) has the form

$$\dot{s}_i = -\varepsilon b s_i.$$

If $b > 0$, $\bar{s} = 0 \in \mathbb{R}^n$ is an exponentially stable equilibrium, and assumption A1 is satisfied. Assumption A2 is also satisfied, since at $\bar{s} = 0$ all phase oscillators, described by $\dot{\vartheta}_i = 1$, have equal period $T = 2\pi$. Let $s_i = \varepsilon w_i$; then system (1.8), (1.9) has the form

$$\dot{\vartheta}_i = 1 + \varepsilon \sum_{j=1}^n c_{ij} w_j,$$

$$\dot{w}_i = \cos\vartheta_i - \varepsilon b w_i.$$

When $\varepsilon = 0$, this system has a family of solutions

$$\vartheta_i(t) = t + \varphi_i \qquad \text{and} \qquad w_i(t) = u_i + \sin(t + \varphi_i).$$

Since $\partial f_i/\partial \vartheta_i = 0$, the adjoint system (1.6) has solutions $Q_i(t) = 1$ and $P_i(t) = -\cos t$, which satisfy the normalization condition (1.7). It is easy to check that

$$\begin{aligned}
a_{ij} &= \frac{1}{T}\int_0^T 1 \cdot c_{ij}\, dt = c_{ij},\\
b_{ij} &= \frac{1}{T}\int_0^T [-\cos t \cdot c_{ij} + 0]\, dt = 0, \qquad i \neq j,\\
b_{ii} &= \frac{1}{T}\int_0^T [-\cos t \cdot c_{ii} - b]\, dt = -b,\\
H_{ij}(\chi) &= \frac{1}{T}\int_0^T \left[1 \cdot c_{ij} \cdot \int_0^{t+\chi} \cos\bar{t}\, d\bar{t}\right] dt = 0,\\
K_{ij}(\chi) &= \frac{1}{T}\int_0^T \left[-\cos t \cdot c_{ij} \int_0^{t+\chi} \cos\bar{t}\, d\bar{t}\right] dt = -\frac{c_{ij}}{2}\sin\chi, \qquad i \neq j,
\end{aligned}$$

and $K_{ii}(0) = 0$, so that the canonical phase model (1.4), (1.5) has the form

$$\begin{aligned}
\varphi_i' &= \sum_{j=1}^n c_{ij} u_j,\\
u_i' &= -bu_i - \frac{1}{2}\sum_{j=1}^n c_{ij}\sin(\varphi_j - \varphi_i).
\end{aligned}$$

It is a simple exercise to check that the same canonical model can be obtained via standard averaging of (2.1), (2.2).

2.2. Frankel and Kiemel's example. As an illustration, Frankel and Kiemel (1993) considered the six-dimensional system

$$\begin{aligned}
\dot{\vartheta}_i &= 1 + s_j(\alpha + \beta r_i \cos\vartheta_i + \gamma r_i^2 \cos^2\vartheta_i),\\
\dot{r}_i &= r_i - r_i^3 + \eta s_j r_i^2 \cos\vartheta_i,\\
\dot{s}_i &= \varepsilon(r_i \cos\vartheta_i - \mu s_i)
\end{aligned}$$

having fast variables in polar coordinates $\mathbb{S}^1 \times \mathbb{R}$ and parameters $\alpha, \beta, \gamma, \eta, \mu \in \mathbb{R}$, and $i, j \in \{1, 2\}$, $i \neq j$. They used a completely different approach to show how the model can be reduced to the planar system

$$\chi' = (-\alpha - \gamma/2)u - \beta\sin\chi, \tag{2.3}$$

$$u' = (\beta/2 - \eta/5 - \mu)u + (\alpha + 3\gamma/4)\sin\chi, \tag{2.4}$$

where

$$\chi = \varphi_2 - \varphi_1 \qquad \text{and} \qquad u = u_2 - u_1 \tag{2.5}$$

and u_i and φ_i have the same meaning as in this paper.

Let us verify Frankel and Kiemel's result using Theorem 1.1. Each uncoupled oscillator has a 2π-periodic solution $(t,1)\in\mathbb{S}^1\times\mathbb{R}$ when $\bar{s}=0$. It is easy to check by differentiating that the adjoint linear systems (1.6),

$$\dot{Q}=-\left\{\begin{array}{cc}0&0\\0&-2\end{array}\right\}^{\top}Q \qquad\text{and}\qquad \dot{P}=-\left\{\begin{array}{cc}0&0\\0&-2\end{array}\right\}^{\top}P-\left\{\begin{array}{cc}-\sin t, & \cos t\end{array}\right\}^{\top},$$

have solutions

$$Q(t)=\left(\begin{array}{c}1\\0\end{array}\right) \qquad\text{and}\qquad P(t)=\left(\begin{array}{c}-\cos t\\ \frac{2}{5}\cos t-\frac{1}{5}\sin t\end{array}\right)$$

satisfying the normalization conditions (1.7). Therefore, the parameters in the canonical model (1.4), (1.5) are (here $i\neq j$)

$$\begin{aligned}
a_{ij}&=\frac{1}{2\pi}\int_0^{2\pi}\left(\begin{array}{c}1\\0\end{array}\right)^{\top}\left(\begin{array}{c}\alpha+\beta\cos t+\gamma\cos^2 t\\ \eta\cos t\end{array}\right)dt=\alpha+\frac{\gamma}{2},\\
b_{ij}&=\frac{1}{2\pi}\int_0^{2\pi}\left(\begin{array}{c}-\cos t\\ \frac{2}{5}\cos t-\frac{1}{5}\sin t\end{array}\right)^{\top}\left(\begin{array}{c}\alpha+\beta\cos t+\gamma\cos^2 t\\ \eta\cos t\end{array}\right)dt=-\frac{\beta}{2}+\frac{\eta}{5},\\
H_{ij}(\chi)&=\frac{1}{2\pi}\int_0^{2\pi}\left(\begin{array}{c}1\\0\end{array}\right)^{\top}\left(\begin{array}{c}\alpha+\beta\cos t+\gamma\cos^2 t\\ \eta\cos t\end{array}\right)\int_0^{t+\chi}\cos\bar{t}\,d\bar{t}\,dt=\frac{\beta}{2}\sin\chi,\\
K_{ij}(\chi)&=\frac{1}{2\pi}\int_0^{2\pi}\left(\begin{array}{c}-\cos t\\ \frac{2}{5}\cos t-\frac{1}{5}\sin t\end{array}\right)^{\top}\left(\begin{array}{c}\alpha+\beta\cos t+\gamma\cos^2 t\\ \eta\cos t\end{array}\right)\int_0^{t+\chi}\cos\bar{t}\,d\bar{t}\,dt\\
&=\left(-\frac{\alpha}{2}-\frac{3}{8}\gamma\right)\sin\chi,
\end{aligned}$$

and

$$a_{ii}=0, \qquad b_{ii}=-\mu, \qquad H_{ii}(0)=0, \qquad\text{and}\qquad K_{ii}(0)=-1/2\ .$$

Using the difference variables (2.5) we arrive exactly at the same model (2.3), (2.4) that Frankel and Kiemel did.

2.3. Andronov–Hopf bifurcation. Next, we derive the canonical phase model for a network of slowly coupled oscillators near an Andronov–Hopf bifurcation. Without loss of generality we may assume that $\bar{s}=0$ and that each oscillator has already been converted into the topological normal form (by a continuous near-identity change of variables; see Hoppensteadt and Izhikevich (1997)). We use complex coordinates for convenience and consider the system

$$\begin{aligned}
&\dot{z}_i=(\mu+\mathrm{i})z_i-z_i|z_i|^2+q_i(z_i,\bar{z}_i,s_1,\ldots,s_n), && z_i\in\mathbb{C},\\
&\dot{s}_i=\varepsilon g_i(z_i,\bar{z}_i,s_i), && s_i\in\mathbb{R},
\end{aligned}$$

where $\mathrm{i}=\sqrt{-1}$ has a different font from the subscript i, $0<\varepsilon\ll\mu\ll 1$ (we assume μ is sufficiently small so that higher-order terms in μ may be neglected) and q_i and g_i are arbitrary smooth functions satisfying

$$q_i(z_i,\bar{z}_i,0,\ldots,0)=0 \qquad\text{and}\qquad g_i(0,0,0)=0.$$

In this case each unperturbed ($\varepsilon=0$) oscillator

$$\dot{z}_i=(\mu+\mathrm{i})z_i-z_i|z_i|^2$$

has a small amplitude limit cycle attractor

$$\gamma_i(t) = \sqrt{\mu}\, e^{\mathrm{i}t} \subset \mathbb{C}$$

with period $T = 2\pi$. We do not need to solve the adjoint linear system (1.6),

$$\dot{Q}_i = -\{\mathrm{i} + \mathcal{O}(\mu)\}^* Q_i \qquad \text{and} \qquad \dot{P}_i = -\{\mathrm{i} + \mathcal{O}(\mu)\}^* P_i - \{\partial g_i/\partial z_i\}^*,$$

since we can find the solutions

$$Q_i(t) = \mathrm{i}e^{\mathrm{i}t}/\sqrt{\mu} \qquad \text{and} \qquad P_i(t) = -\mathrm{i}e^{\mathrm{i}t}\left\{e^{\mathrm{i}t}\partial g_i/\partial z_i + \text{c.c.}\right\}$$

directly from the normalization condition (1.7),

$$Q_i(t)^*\left\{(\mu + \mathrm{i})\gamma_i(t) - \gamma_i(t)|\gamma_i(t)|^2\right\} = Q_i(t)^*\left\{\mathrm{i}\sqrt{\mu}e^{\mathrm{i}t}\right\} = 1$$

and

$$P_i(t)^*\left\{\mathrm{i}\sqrt{\mu}e^{\mathrm{i}t}\right\} = -g_i(\sqrt{\mu}e^{\mathrm{i}t}, 0) = -\left\{\sqrt{\mu}e^{\mathrm{i}t}\partial g_i/\partial z_i + \text{c.c.}\right\},$$

where $Q_i(t), P_i(t) \in \mathbb{C}$, * denotes transposition and complex conjugation, and c.c. means complex-conjugate. Now we can apply Theorem 1.1 to obtain the canonical phase model

$$\begin{aligned}
\varphi_i' &= \sum_{j=1}^{n} \{a_{ij}u_j + c_{ij}\sin(\varphi_j + \psi_{ij} - \varphi_i)\}, \\
u_i' &= \sum_{j=1}^{n} b_{ij}u_j,
\end{aligned}$$

where

$$\begin{aligned}
a_{ij} &= \operatorname{Im}\frac{\partial^2 q_i}{\partial s_j \partial z_i}, \\
b_{ij} &= -\operatorname{Im}\frac{\partial q_i}{\partial s_j}\frac{\partial g_i}{\partial z_i}, \qquad i \neq j, \\
b_{ii} &= -\operatorname{Im}\frac{\partial q_i}{\partial s_i}\frac{\partial g_i}{\partial z_i} + \frac{\partial g_i}{\partial s_i},
\end{aligned}$$

and all derivatives are evaluated at the origin $z = 0$ and $s = 0$. Notice that $P_i = \mathcal{O}(1)$ and $g_i = \mathcal{O}(\sqrt{\mu})$, hence $K_{ij} = \mathcal{O}(\sqrt{\mu})$ is infinitesimal. Let us show that

$$H_{ij}(\chi) = c_{ij}\sin(\psi_{ij} + \chi).$$

First,

$$\begin{aligned}
\int_0^{t+\chi} g_j(\sqrt{\mu}e^{\mathrm{i}\bar{t}}, \sqrt{\mu}e^{-\mathrm{i}\bar{t}}, 0)\, d\bar{t} &= \int_0^{t+\chi} \frac{\partial g_j}{\partial z_j}\sqrt{\mu}e^{\mathrm{i}\bar{t}}d\bar{t} + \text{c.c.} \\
&= \mathrm{i}\sqrt{\mu}\frac{\partial g_j}{\partial z_j}\left(1 - e^{\mathrm{i}(t+\chi)}\right) + \text{c.c.}
\end{aligned}$$

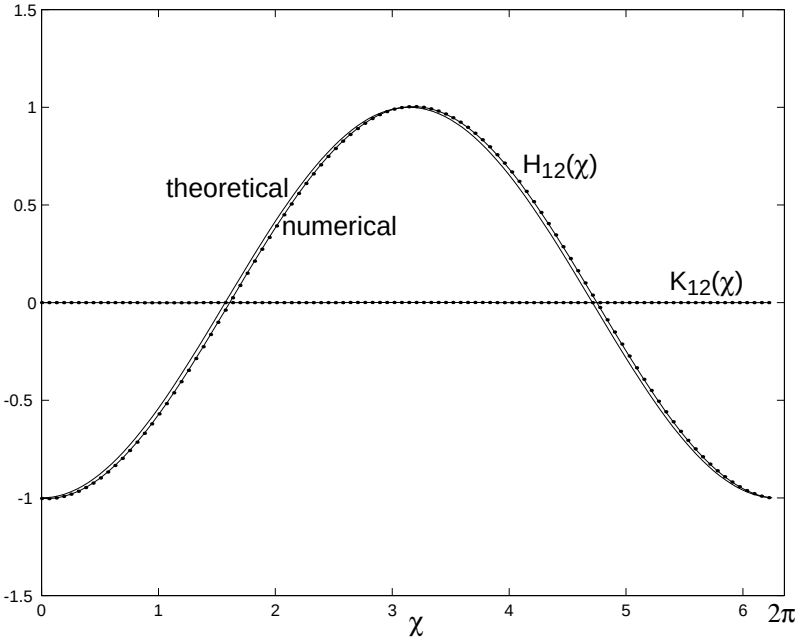


FIG. 2.1. *Numerical (see Appendix* B*) and analytical form of the connection functions for slowly coupled oscillators near an Andronov–Hopf bifurcation. Parameters:* $q_1(z_1, \bar{z}_1, s_1, s_2) = s_2$ *and* $g_1(z_1, \bar{z}_1, s_1) = z_1 + \bar{z}_1 - s_1$, $\mu = 0.01$.

Next,

$$\begin{aligned} H_{ij}(\chi) &= \operatorname{Re} \frac{1}{2\pi} \int_0^{2\pi} \left(\frac{-\mathrm{i} e^{-\mathrm{i}t}}{\sqrt{\mu}} \right) \left(\frac{\partial q_i}{\partial s_j} + \mathcal{O}(\sqrt{\mu}) \right) \left(\mathrm{i}\sqrt{\mu} \frac{\partial g_j}{\partial z_j} \left(1 - e^{\mathrm{i}(t+\chi)} \right) + \text{c.c.} \right) dt \\ &= \operatorname{Re} \frac{1}{2\pi} \int_0^{2\pi} \left(-\frac{\partial q_i}{\partial s_j} \frac{\partial g_j}{\partial z_j} e^{\mathrm{i}\chi} + \text{terms involving } e^{\mathrm{i}t} \right) dt \\ &= -\operatorname{Re} \frac{\partial q_i}{\partial s_j} \frac{\partial g_j}{\partial z_j} e^{\mathrm{i}\chi} = c_{ij} \sin(\psi_{ij} + \chi), \end{aligned}$$

where

$$c_{ij} = \left| \frac{\partial q_i}{\partial s_j} \frac{\partial g_j}{\partial z_j} \right| \qquad \text{and} \qquad \psi_{ij} = \operatorname{Arg} \frac{\partial q_i}{\partial s_j} \frac{\partial g_j}{\partial z_j} - \frac{\pi}{2}.$$

A typical example of the connection function $H_{ij}(\chi)$ is depicted in Figure 2.1. (Because theoretical and numerical curves were obtained using essentially the same formulae, this figure illustrates only the accuracy of the numerical method, and it does not validate the theory.)

Since the connection function $K_{ij}(\chi) = 0$ for all χ, the variables u_i do not depend on the phase variables φ_i. If the matrix (b_{ij}) is stable, all $u_i(t) \to 0$ as $t \to \infty$, and the locking properties of the *slowly* connected network are described by Kuramoto's (1984) system

$$\varphi_i' = \sum_{j=1}^{n} c_{ij} \sin(\varphi_j + \psi_{ij} - \varphi_i),$$

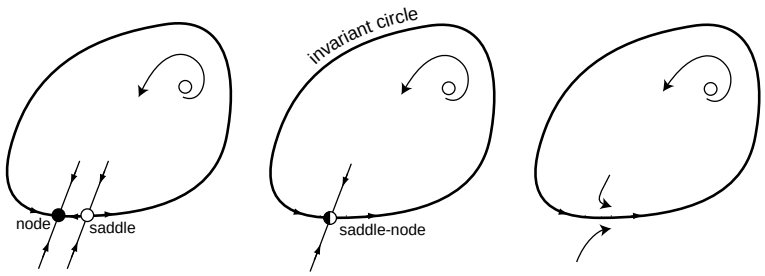


FIG. 2.2. *Saddle-node on invariant circle bifurcation.*

which was originally derived for *weakly* connected networks of Andronov–Hopf oscillators. Notice, though, that the variables u_i may significantly slow down the convergence to a synchronized state if the matrix (b_{ij}) has eigenvalues close to the imaginary axis.

2.4. Saddle-node on invariant circle bifurcation. Now we consider a slowly connected network of oscillators near a saddle-node on invariant circle bifurcation, which is illustrated in Figure 2.2. This bifurcation results in Class 1 excitable systems, i.e., systems able to oscillate with arbitrary small frequency (Hoppensteadt and Izhikevich (1997)). Without loss of generality we assume that $\bar{s} = 0$ and that each oscillator has been restricted to the center manifold and converted to the topological normal form by an appropriate change of variables,

$$\begin{aligned} \dot{y}_i &= \mu + y_i^2 + q_i(y_i, c_1, \ldots, c_n, \mu), \qquad y_i \in \mathbb{R}, \\ \dot{c}_i &= \epsilon g_i(y_i, c_i), \qquad\qquad\qquad\qquad\quad c_i \in \mathbb{R}, \end{aligned}$$

where $\epsilon \ll \sqrt{\mu} \ll 1$ and

$$q_i(y_i, 0, \ldots, 0, \mu) = 0 \qquad \text{and} \qquad g_i(0,0) = 0.$$

Here we consider only a small neighborhood of the origin, since the spike lemma (Lemma 8.1 in Hoppensteadt and Izhikevich (1997)) implies that y_i spends a negligible amount of time outside the small neighborhood; that is, action potentials generated by such a model look instantaneous on the slow time scale of order $1/\sqrt{\mu}$. Now we rescale the variables and parameters,

$$y_i = 2\sqrt{\mu}x_i, \qquad c_i = 2\sqrt{\mu}\{\partial g_i/\partial y_i\}s_i, \qquad \epsilon = 2\sqrt{\mu}\varepsilon, \qquad t_{\text{new}} = 2\sqrt{\mu}t_{\text{old}},$$

to transform the system above into the form

$$\dot{x}_i = 1/4 + x_i^2 + \sum_{j=1}^{n}(c_{ij} + h_{ij}x_i)s_j + \mathcal{O}(\sqrt{\mu}), \tag{2.6}$$

$$\dot{s}_i = \varepsilon\left\{x_i - p_i s_i + \mathcal{O}(\sqrt{\mu})\right\}, \tag{2.7}$$

where

$$\begin{aligned} c_{ij} &= \frac{1}{2}\,\frac{\partial g_i}{\partial y_i}\,\frac{\partial^2 q_i}{\partial c_j \partial \mu}, \\ h_{ij} &= \frac{\partial g_i}{\partial y_i}\,\frac{\partial^2 q_i}{\partial c_j \partial y_i}, \\ p_i &= -\frac{\partial g_i}{\partial c_i}, \end{aligned}$$

and all derivatives are evaluated at the origin. Notice that each unperturbed ($\varepsilon = 0$) oscillator has a limit cycle attractor

$$\gamma_i(t) = \frac{1}{2}\tan\frac{t}{2} \subset \mathbb{R} \cup \{\infty\}$$

with period $T = 2\pi$. From the normalization condition (1.7),

$$Q_i(t)^\top (1/4 + \gamma_i(t)^2) = 1 \qquad \text{and} \qquad P_i(t)^\top (1/4 + \gamma_i(t)^2) = -\gamma_i(t),$$

we can find directly the solutions to the adjoint problem (1.6),

$$Q_i(t) = 2(1 + \cos t) \qquad \text{and} \qquad P_i(t) = -\sin t,$$

and use them in Theorem 1.1 to find all parameters and functions. It is easy to see that

$$\begin{aligned}
a_{ij} &= \frac{1}{2\pi}\int_0^{2\pi} 2(1+\cos t)\ \left(c_{ij} + \frac{h_{ij}}{2}\tan\frac{t}{2}\right)\ dt = 2c_{ij},\\
b_{ij} &= \frac{1}{2\pi}\int_0^{2\pi} -\sin t\ \left(c_{ij} + \frac{h_{ij}}{2}\tan\frac{t}{2}\right)\ dt = -\frac{h_{ij}}{2}, \qquad i \neq j,\\
b_{ii} &= -\frac{h_{ii}}{2} - p_i.
\end{aligned}$$

Next,

$$\int_0^{t+\chi} \frac{1}{2}\tan\frac{\bar{t}}{2}\, d\bar{t} = -\ln\left|\cos\frac{t+\chi}{2}\right|,$$

and

$$\begin{aligned}
H_{ij}(\chi) &= \frac{1}{2\pi}\int_0^{2\pi} 2(1+\cos t)\ \left(c_{ij} + \frac{h_{ij}}{2}\tan\frac{t}{2}\right)\ \left(-\ln\left|\cos\frac{t+\chi}{2}\right|\right)\ dt\\
&= c_{ij}(2\ln 2 - \cos\chi) + \frac{h_{ij}}{2}\sin\chi,\\
K_{ij}(\chi) &= \frac{1}{2\pi}\int_0^{2\pi} -\sin t\ \left(c_{ij} + \frac{h_{ij}}{2}\tan\frac{t}{2}\right)\ \left(-\ln\left|\cos\frac{t+\chi}{2}\right|\right)\ dt\\
&= -\frac{c_{ij}}{2}\sin\chi - \frac{h_{ij}}{4}(2\ln 2 + \cos\chi), \qquad i \neq j,\\
K_{ii}(0) &= -\frac{h_{ii}}{4}(2\ln 2 + 1) - p_i\ \ln 2\ .
\end{aligned}$$

A typical form of the connection function $H_{ij}(\chi)$ and $K_{ij}(\chi)$ is depicted in Figure 2.3.

2.4.1. Two identical oscillators. Let us consider synchronization dynamics of two identical slowly coupled oscillators of the form

$$\begin{aligned}
&\dot{x}_1 = 1/4 + x_1^2 + (c + hx_1)s_2, && \dot{x}_2 = 1/4 + x_2^2 + (c + hx_2)s_1,\\
&\dot{s}_1 = \varepsilon(x_1 - ps_1), && \dot{s}_2 = \varepsilon(x_2 - ps_2)
\end{aligned}$$

with $p > 0$ and arbitrary c and h. The canonical phase model has the form

$$\begin{aligned}
\dot{\varphi}_1 &= 2cu_2 + H(\varphi_2 - \varphi_1),\\
\dot{u}_1 &= -p(u_1 + \ln 2) - \frac{h}{2}(u_1 + u_2) + K(\varphi_2 - \varphi_1),\\
\dot{\varphi}_2 &= 2cu_1 + H(\varphi_1 - \varphi_2),\\
\dot{u}_2 &= -p(u_2 + \ln 2) - \frac{h}{2}(u_1 + u_2) + K(\varphi_1 - \varphi_2),
\end{aligned}$$

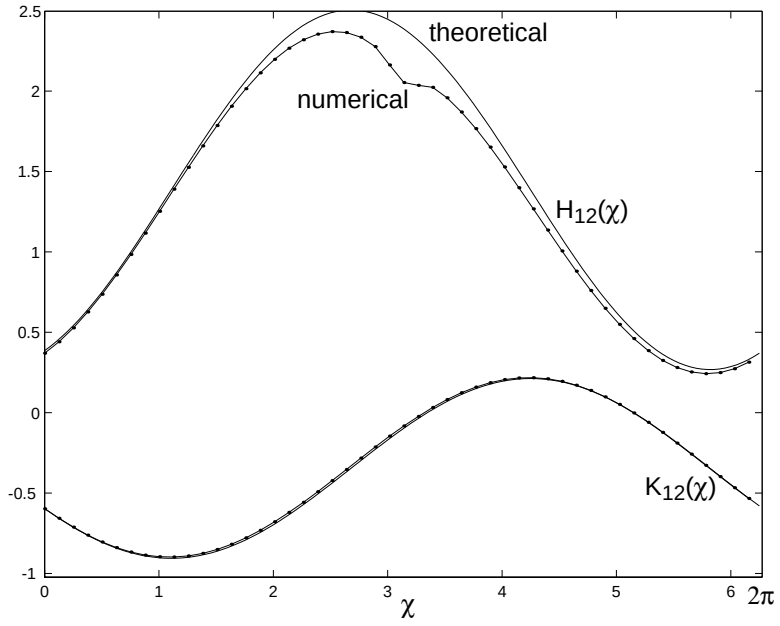


FIG. 2.3. *Numerical (see Appendix* B*) and analytical forms of the connection functions for slowly coupled oscillators near a saddle-node on invariant circle bifurcation. Parameters in* (2.6), (2.7)*:* $c_{12} = h_{12} = p_1 = 1$.

where

$$H(\chi) = c(2\ln 2 - \cos\chi) + \frac{h}{2}\sin\chi \qquad \text{and} \qquad K(\chi) = -\frac{c}{2}\sin\chi - \frac{h}{4}(2\ln 2 + \cos\chi).$$

Let $\chi = \varphi_2 - \varphi_1$ and $u = u_2 - u_1$; then

$$\begin{aligned} \dot{\chi} &= -2cu - h\sin\chi, \\ \dot{u} &= -pu + c\sin\chi. \end{aligned}$$

The in-phase synchronized solution corresponds to the equilibrium $\chi = 0$ and $u = 0$ with the Jacobian matrix

$$L = \begin{pmatrix} -h & -2c \\ c & -p \end{pmatrix}.$$

It is stable when

$$\mathrm{tr}L = -h - p < 0 \qquad \text{and} \qquad \det L = hp + 2c^2 > 0,$$

which is always the case when $h > 0$. (It is an easy exercise to check that the antiphase solution $\chi = \pi$, $u = 0$ is stable when $h < p$ and $hp + 2c^2 < 0$.)

We see that in contrast to weak coupling, which leads to a neutrally stable synchronized state of two Class 1 identical oscillators (Hansel, Mato, and Meunier (1995), Ermentrout (1996), Izhikevich (1999)), slow coupling with $h > 0$ always results in stability of the synchronized state, as we illustrate in Figure 2.4, regardless of the sign of the connection coefficient c. Notice that the convergence to the in-phase synchronized state $\chi = 0$ is oscillatory, as we illustrate in Figure 2.5(a); that is, the oscillators

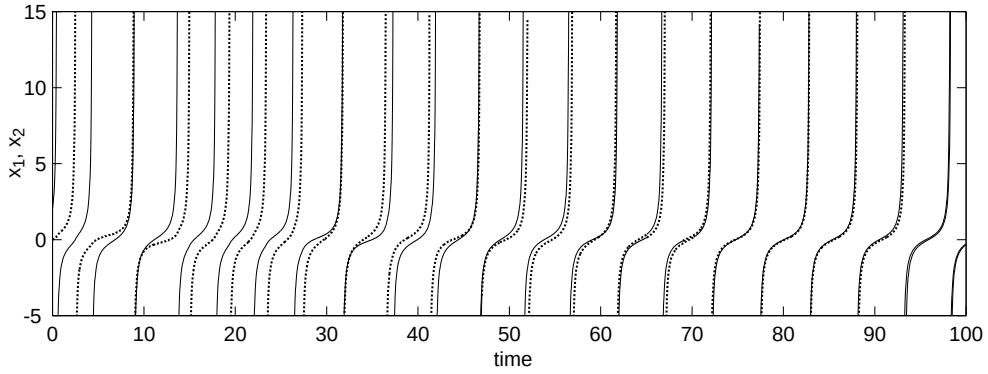


FIG. 2.4. *Synchronization dynamics of two identical oscillators near a saddle-node on invariant circle bifurcation. Parameters:* $c = 5$, $h = p = 1$, *and* $\varepsilon = 0.05$.

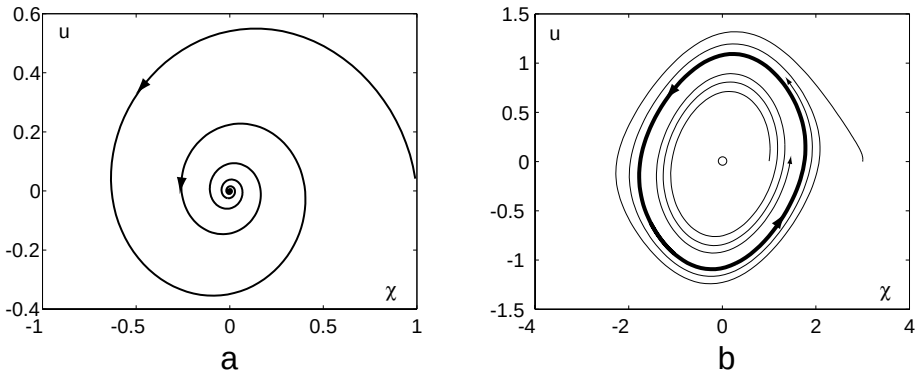


FIG. 2.5. (a) *Convergence to the in-phase synchronized state is oscillatory (parameters as in Figure* 2.4*).* (b) *The phase difference* χ *may oscillate (here* $h = -1.5$*).*

in Figure 2.4 take turns—a prominent feature of slowly connected networks that was discovered by Frankel and Kiemel (1993). If we decrease h past $-p$, a small amplitude limit cycle may appear via a supercritical Andronov–Hopf bifurcation, and the phase difference χ would exhibit sustained oscillations, as shown in Figure 2.5(b).

2.5. Relaxation oscillators. Now we consider a slowly coupled network of relaxation oscillators of the form

$$\begin{aligned} \mu\dot{x}_i &= F_i(x_i, y_i) + q_i(x_i, y_i, s_1, \ldots, s_n), \\ \dot{y}_i &= G_i(x_i, y_i) + r_i(x_i, y_i, s_1, \ldots, s_n), \\ \dot{s}_i &= \varepsilon g_i(x_i, y_i, s_i), \end{aligned}$$

where $\mu \ll 1$, $x_i, y_i, s_i \in \mathbb{R}$, and

$$q_i(x_i, y_i, 0, \ldots, 0) = 0, \quad r_i(x_i, y_i, 0, \ldots, 0) = 0, \qquad \text{and} \qquad g_i(0, 0, 0) = 0.$$

Suppose that each unperturbed ($\varepsilon = 0$, $s = 0$) system

$$\begin{aligned} \mu\dot{x}_i &= F_i(x_i, y_i), \\ \dot{y}_i &= G_i(x_i, y_i) \end{aligned}$$

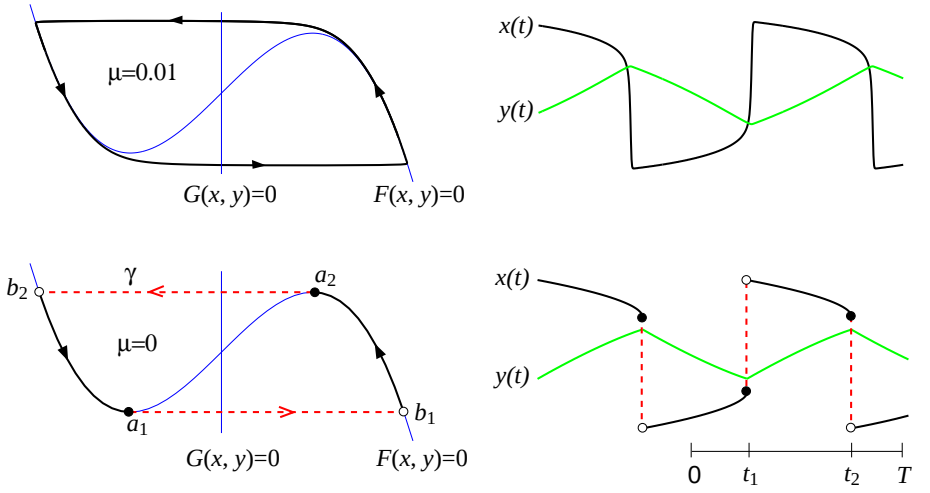


FIG. 2.6. *Top: Nullclines and periodic solution of the van der Pol relaxation oscillator* $\mu\dot{x} = -y + x - x^3/3$, $\dot{y} = x$ *for* $\mu = 0.01$. *Bottom: The periodic solution becomes discontinuous in the limit* $\mu \to 0$. *(Modified from Izhikevich (*2000*)).*

has a relaxation limit cycle attractor with period $T > 0$ converging to $\gamma_i(t) \subset \mathbb{R}^2$ similar to the one depicted in Figure 2.6 in the limit $\mu \to 0$. Such an oscillation $\gamma_i(t)$ has two discontinuities (jumps) at $t = t_1$ and $t = t_2$. Izhikevich (2000) has shown that in this case the solution to the adjoint problem (1.6) converges as $\mu \to 0$ to

$$Q_i(t) = \frac{1}{G_i(\gamma_i(t))}\left(-\frac{\partial G_i}{\partial x_i}(\gamma_i(t))\left(\frac{\partial F_i}{\partial x_i}(\gamma_i(t))\right)^{-1},\ \ 1\right)^{\top} \qquad \text{when } t \neq t_1 \text{ and } t \neq t_2$$

and

$$Q_i(t_k) = \left(c_k\delta(t - t_k),\ \ \frac{1}{G_i(\gamma_i(a_k))}\right)^{\top},$$

where

$$c_k = \left(\frac{\partial F_i}{\partial y_i}(\gamma_i(t))\right)^{-1}\left(\frac{1}{G_i(a_k)} - \frac{1}{G_i(b_k)}\right)$$

and a_k and b_k are the end points of the kth jump, $k = 1, 2$; see Figure 2.6.

Knowing $Q_i(t)$ and $P_i(t)$, one can easily find a_{ij}, b_{ij}, H_{ij}, and K_{ij}. For example,

$$H_{ij}(\chi) = -\frac{1}{T}\int_0^T \frac{\frac{\partial G_i}{\partial x_i}(\gamma_i(t))\left(\frac{\partial F_i}{\partial x_i}(\gamma_i(t))\right)^{-1}\frac{\partial q_i}{\partial s_j}(\gamma_i(t)) - \frac{\partial r_i}{\partial s_j}(\gamma_i(t))}{G_i(\gamma_i(t))}\int_0^{t+\chi} g_j(\gamma_i(\bar{t}))\,d\bar{t}\,dt$$

$$+\frac{1}{T}\sum_{k=1}^{2} c_k^{\top}\frac{\partial q_i}{\partial s_j}(a_k)\int_0^{t_k+\chi} g_j(\gamma_j(\bar{t}))\,d\bar{t}. \tag{2.8}$$

A salient feature of weakly coupled relaxation oscillators is the existence of discontinuities in the connection functions $H_{ij}(\chi)$, which result in many interesting synchronization properties, such as superconvergence, persistence under perturbations of natural frequencies, etc. (see the discussion in Izhikevich (2000)). However, the

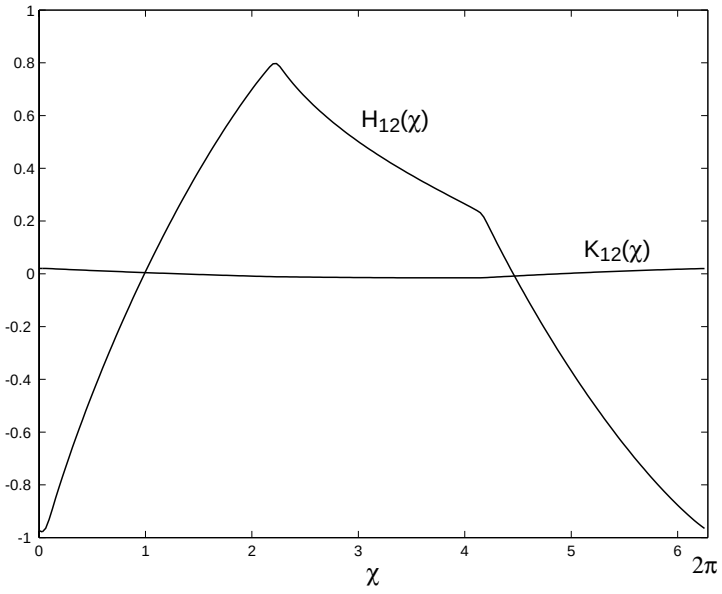


FIG. 2.7. *Numerically found connection functions for slowly coupled Bonhoeffer–van der Pol relaxation oscillators* $\mu\dot{x}_i = x_i - x_i^3/3 - y_i + s_j$, $\dot{y}_i = 0.5 + x_i$, $\dot{s}_i = \varepsilon(x_i - s_i)$, $i = 1, 2$, $j = 2, 1$, $\mu = 0.001$.

connection functions $H_{ij}(\chi)$ (and $K_{ij}(\chi)$) for slowly coupled relaxation oscillators are continuous, as shown in Figure 2.7. Hence, slowly coupled relaxation oscillators do not have those interesting synchronization properties.

3. Discussion. Much research in theoretical neuroscience is devoted to weakly coupled oscillators, as reviewed in Chapter 9 of Hoppensteadt and Izhikevich (1997). Slow connections, however, received less attention despite the fact that slow synaptic transmission is ubiquitous in the brain. Indeed, GABAb and NMDA receptor dynamics occur on the time scale of 150 ms, which are much slower than the period of firing of many neurons, which is often smaller than 10 ms. Purely NMDA synaptic connections have been found in the hippocampus (Isaac, Nicoll, and Malenka (1995)) and in the thalamocortical system (Isaac et al. (1997)). They are often referred to as being "silent synapses" since activation of NMDA receptors requires postsynaptic depolarization. (Most synapses have slow NMDA and fast AMPA glutamate receptors and hence describe slow and weak connections; see Remark 1.3.) Here we speculate that periodically firing neurons connected via NMDA or GABAb receptors could have quite different synchronization properties from the same neurons coupled via fast GABAa or AMPA receptors.

There have been only a few attempts to study rigorously slowly coupled oscillators. Notably, Rinzel and Frankel (1992) and Ermentrout (1994) used averaging to study dynamics of slowly coupled Hodgkin–Huxley-type models. They discovered new regimes that were not seen in weakly coupled networks, such as instabilities of relative phases and network bursting. Bressloff and Coombes (2000) applied similar methods to study integrate-and-fire neurons with slow synapses. The most important contribution was made by Frankel and Kiemel (1993) who showed how two slowly coupled oscillators can be transformed into a canonical phase model by an appropriate change of variables.

In this paper we extend the results of Frankel and Kiemel (1993) to a network of $n \geq 2$ slowly coupled oscillators. We confirm that synchronization properties of such a network are described by a canonical phase model in which each oscillator is represented by a pair of variables φ_i and u_i on a cylindrical phase space. Using Malkin's theorem we can derive an analytical form for all coefficients and all connection functions H_{ij} and K_{ij} for the phase model.

In the second half of the paper we consider a few analytically solvable examples. First, we study oscillators near an Andronov–Hopf bifurcation and show that dynamics of variables u_i do not depend on the phases φ_i. While the variables u_i can slow down the convergence to a synchronized state, they cannot change the stability of that state.

Next, we study synchronization properties of Class 1 oscillators; that is, oscillators that are near a saddle-node on invariant circle bifurcation (Figure 2.2). Such a bifurcation results in periodic activity with arbitrarily small frequency, and it is believed to be involved in excitable properties of neocortical neurons in mammalian brains. It is well known (Hansel, Mato, and Meunier (1995), Ermentrout (1996)) that identical Class 1 oscillators if coupled weakly do not synchronize; more precisely, the synchronized state is neutrally stable for $n = 2$ oscillators and unstable for $n > 2$ oscillators (Izhikevich (1999)) regardless of whether the connections are excitatory or inhibitory. In contrast, *slowly* connected Class 1 oscillators do synchronize for both excitatory and inhibitory connections. We show this analytically using the canonical phase model approach, and then illustrate it numerically in Figure 2.4. Notice that the oscillators take turns, i.e., change order of firing, during the convergence to the synchronized state. This is a salient feature of slowly connected oscillators found by Frankel and Kiemel (1993) that cannot occur in weakly coupled networks of $n = 2$ oscillators.

Finally, we consider slowly coupled relaxation oscillators and show that the connection functions $H_{ij}(\chi)$ and $K_{ij}(\chi)$ are continuous. Therefore, in contrast to *weak* coupling, *slow* coupling of relaxation oscillators does not lead to superstability of the in-phase synchronized solution (Izhikevich (2000)).

Appendix A. Malkin's theorem. Below we provide a general statement of Malkin's theorem (Malkin (1949), (1956)). A one-page proof can be found in (Hoppensteadt and Izhikevich (1997)).

THEOREM A.1 (Malkin). *Consider a T-periodic dynamical system of the form*

$$\dot{X} = F(X,t) + \varepsilon G(X,t,\varepsilon), \qquad X \in \mathbb{R}^m, \tag{A.1}$$

and suppose that the unperturbed system, $\dot{X} = F(X,t)$, has a k-parameter family of T-periodic solutions

$$X(t) = U(t,\alpha),$$

where $\alpha = (\alpha_1, \ldots, \alpha_k)^\top \in \mathbb{R}^k$ is a vector of independent parameters, by which we mean that the rank of the $n \times k$ matrix $D_\alpha U$ is k. Suppose the adjoint linear problem

$$\dot{R}_i = -\left\{DF(U(t,\alpha),t)\right\}^\top R_i \tag{A.2}$$

has exactly k independent T-periodic solutions $R_1(t,\alpha), \ldots, R_k(t,\alpha) \in \mathbb{R}^m$. Let R be the matrix whose columns are these solutions such that

$$R^\top D_\alpha U = I, \tag{A.3}$$

where I is the identity $k \times k$ matrix. Then the perturbed system (A.1) *has a solution of the form*

$$X(t) = U(t, \alpha(\varepsilon t)) + \mathcal{O}(\varepsilon),$$

where

$$\frac{d\alpha}{d\tau} = \frac{1}{T}\int_0^T R(t,\alpha)^\top G(U(t,\alpha),t,0)\,dt, \tag{A.4}$$

where $\tau = \varepsilon t$ is slow time. If (A.4) *has a stable equilibrium, then system* (A.1) *has a T-periodic solution.*

Appendix B. Numerical recipe. A good numerical method to solve the adjoint problem (1.6) was suggested by Williams and Bowtell (1997), and it is available in the Bard Ermentrout software package XPP. Here, for the sake of convenience, we present MATLAB script that uses the same method to determine all parameters and functions of Malkin's theorem.

The following MATLAB program consists of eight separate files, which are available at the first author's website. The user should provide the following parameters and functions:

- The period `T` and an initial point `x0` on the limit cycle $\gamma(t)$ in the file `main.m`.
- The right-hand sides of the fast and slow systems in files `f.m` and `g.m`, respectively.
- The parameters `Np`, `NT`, and `ds` in the file `main.m`, and `dx` and `dy` in the file `Df.m` control the accuracy of the numerical method. They may be changed if necessary.

File `main.m`

```
function main
% Eugene M. Izhikevich and Frank C. Hoppensteadt, December 19, 2001
% Determines all parameters and functions for two slowly
% coupled oscillators.
global s1 s2 Np T gamma
Np=200;        % The number of points on the limit cycle
NT=200;        % The number of iterations along the limit cycle
T=2*pi;        % The period of the limit cycle
x0=[0.1;0];    % An initial point on the limit cycle
s1=0; s2=0;    % Steady-state value of s
[tg,gamma] = ode23s('f',(0:Np-1)/Np*T, x0);
figure(1),plot(tg,gamma); drawnow;
%
% solve the adjoint for Q(t) as t -> -infty
[t,Qinv] = ode15s('adQ',(0:NT*Np-1)/Np*T, [1;1]);
Q = Qinv(length(t)-(0:Np-1),:);                    % Q(t) => Q(-t)
Q = Q/(Q(1,:)*f(0,gamma(1,:)));                    % Normalization
% solve the adjoint for P(t) as t -> -infty
[t,Pinv] = ode15s('adP',(0:NT*Np-1)/Np*T, [1;1]);
P = Pinv(length(t)-(0:Np-1),:);                    % P(t) => P(-t)
P = P-Q*(P(1,:)*f(0,gamma(1,:))+g(0,gamma(1,:)));   % Normalization
figure(2),plot(tg,Q,tg,P); drawnow;
```

```
%
% Determine all parameters and functions
gv=funvect('g',tg,gamma);
fv=funvect('f',tg,gamma);
intg = cumtrapz(gv)*T/Np;
ds=0.0000001;
s1=s1+ds;
dfds1 = (funvect('f',tg,gamma)-fv)/ds;
dgds1 = (funvect('g',tg,gamma)-gv)/ds;
s1 = s1-ds;
s2=s2+ds;
dfds2 = (funvect('f',tg,gamma)-fv)/ds;
s2=s2-ds;
%
for i=1:Np
   Qdf1(i)=Q(i,:)*dfds1(i,:)';
   Qdf2(i)=Q(i,:)*dfds2(i,:)';
   Pdf1(i)=P(i,:)*dfds1(i,:)';
   Pdf2(i)=P(i,:)*dfds2(i,:)';
end;
%
a11 = trapz(Qdf1)/Np                              % aii
a12 = trapz(Qdf2)/Np                              % aij
b11 = trapz(Pdf1+dgds1')/Np                       % bii
b12 = trapz(Pdf2)/Np                              % bij
H110 = trapz(Qdf1.*intg')/Np                      % Hii(0)
K110 = trapz((Pdf1+dgds1').*intg')/Np             % Kii(0)

for chi=1:Np
   H12(chi) = trapz(Qdf2.*intg')/Np;              % Hij(chi)
   K12(chi) = trapz(Pdf2.*intg')/Np;              % Kij(chi)
   intg = [intg(2:end);intg(1)];
end;
figure(3),plot(tg,H12,tg,K12);
```

File `f.m`

```
function xdot = f(t,x)
% Right-hand side of the fast system
global s1 s2
xdot = [    0.01*x(1)-x(2)-x(1)*(x(1)^2+x(2)^2)+s2;...
        x(1)+0.01*x(2)-x(2)*(x(1)^2+x(2)^2)];
```

File `g.m`

```
function sdot = g(t,x)
% Right-hand side of the slow system
global s1
sdot = 2*x(1)-s1;
```

File `Df.m`

```
function d = Df(t,x)
% Numerical evaluation of Jacobian matrix Df at the point (t,x)
dx = 0.0000001; dy = 0.0000001;
d = [(f(t,x+[dx;0])-f(t,x))/dx (f(t,x+[0;dy])-f(t,x))/dy];
```

File `Dg.m`

```
function d = Dg(t,x)
% Numerical evaluation of derivative Dg at x
dx = 0.0000001; dy = 0.0000001;
d = [(g(t,x+[dx;0])-g(t,x))/dx (g(t,x+[0;dy])-g(t,x))/dy];
```

File `adQ.m`

```
function Qdot = adQ(t,Q)
% Right-hand side of the adjoint equation
% Integrating as t -> -infty
global Np T gamma
Qdot=Df(t,gamma(ceil(Np*mod(-t/T+0.5/Np,1)),:)')'*Q;
```

File `adP.m`

```
function Pdot = adP(t,P)
% Right-hand side of the adjoint equation for P
% Integrating as t -> -infty
global Np T gamma
gmt = gamma(ceil(Np*mod(-t/T+0.5/Np,1)),:)';
Pdot=Df(t,gmt)'*P + Dg(t,gmt)';
```

File `funvect.m`

```
function ans = funvect(fname,t,x)
% Applies function fname to the vector of arguments t,x
ans = zeros(length(t),length(feval(fname,t(1),x(1,:))));
for i=1:length(t)
   ans(i,:) = feval(fname,t(i),x(i,:))';
end;
```

REFERENCES

P.C. Bressloff and S. Coombes (2000), *Dynamics of strongly coupled spiking neurons*, Neural Computation, 12, pp. 91–129.

G.B. Ermentrout (1994), *Reduction of conductance-based models with slow synapses to neural nets*, Neural Computation, 6, pp. 679–695.

G.B. Ermentrout (1996), *Type* I *membranes, phase resetting curves, and synchrony*, Neural Computation, 8, pp. 979–1001.

P. Frankel and T. Kiemel (1993), *Relative phase behavior of two slowly coupled oscillators*, SIAM J. Appl. Math., 53, pp. 1436–1446.

D. Hansel, G. Mato, and C. Meunier (1995), *Synchrony in excitatory neural networks,* Neural Computations, 7, pp. 307–335.

F.C. Hoppensteadt and E.M. Izhikevich (1997), *Weakly Connected Neural Networks*, Springer-Verlag, New York.

J.T. Isaac, R.A. Nicoll, and R.C. Malenka (1995), *Evidence for silent synapses: Implications for the expression of LTP*, Neuron, 15, pp. 427–34.

J.T. Isaac, M.C. Crair, R.A. Nicoll, and R.C. Malenka (1997), *Silent synapses during development of thalamocortical inputs*, Neuron, 18, pp. 269–280.

E.M. Izhikevich (2000), *Phase equations for relaxation oscillators*, SIAM J. Appl. Math., 60, pp. 1789–1804.

E.M. Izhikevich (1999), *Class* 1 *neural excitability, conventional synapses, weakly connected networks, and mathematical foundations of pulse-coupled models*, IEEE Trans. Neural Networks, 10, pp. 499–507.

Y. Kuramoto (1984), *Chemical Oscillations, Waves, and Turbulence*, Springer-Verlag, New York.

I.G. Malkin (1949), *Methods of Poincare and Liapunov in Theory of Non-linear Oscillations*, Gostexizdat, Moscow (in Russian).

I.G. Malkin (1956), *Some Problems in Nonlinear Oscillation Theory*, Gostexizdat, Moscow (in Russian).

J. Rinzel and P. Frankel (1992), *Activity patterns of a slow synapse network predicted by explicitly averaging spike dynamics*, Neural Computation, 4, pp. 534–545.

T.L. Williams and G. Bowtell (1997), *The calculation of frequency-shift functions for chains of coupled oscillators, with application to a network model of the lamprey locomotor pattern generator*, J. Comput. Neurosci., 4, pp. 47–55.

SIAM J. APPL. MATH.
Vol. 63, No. 6, pp. 1954–1971

OPTIMAL CONTROL APPLIED TO COMPETING CHEMOTHERAPEUTIC CELL-KILL STRATEGIES*

K. RENEE FISTER† AND JOHN CARL PANETTA‡

Abstract. Optimal control techniques are used to develop optimal strategies for chemotherapy. In particular, we investigate the qualitative differences between three different cell-kill models: log-kill hypothesis (cell-kill is proportional to mass); Norton–Simon hypothesis (cell-kill is proportional to growth rate); and, E_{max} hypothesis (cell-kill is proportional to a saturable function of mass). For each hypothesis, an optimal drug strategy is characterized that minimizes the cancer mass and the cost (in terms of total amount of drug). The cost of the drug is nonlinearly defined in one objective functional and linearly defined in the other. Existence and uniqueness for the optimal control problems are analyzed. Each of the optimality systems, which consists of the state system coupled with the adjoint system, is characterized. Finally, numerical results show that there are qualitatively different treatment schemes for each model studied. In particular, the log-kill hypothesis requires less drug compared to the Norton–Simon hypothesis to reduce the cancer an equivalent amount over the treatment interval. Therefore, understanding the dynamics of cell-kill for specific treatments is of great importance when developing optimal treatment strategies.

Key words. optimal control, cancer, cell-kill

AMS subject classifications. 49K20, 35F20

DOI. 10.1137/S0036139902413489

1. Introduction.

When developing effective treatment strategies, understanding the effects of chemotherapeutic drugs on tumors is of primary importance. Several approaches to modeling chemotherapeutic induced cell-kill (killing of tumor cells) have been developed. One of the early approaches was by Schabel, Skipper, and Wilcox [1] who proposed that cell-kill due to a chemotherapeutic drug was proportional to the tumor population. This hypothesis is based on in vitro studies in the murine leukemia cell-line L1210. It states that for a fixed dose, the reduction of large tumors occurred more rapidly than for smaller tumors. Skipper's concept is referred to as the log-kill mechanism. Norton and Simon [2, 3] find this model to be inconsistent with clinical observations of Hodgkin's disease and acute lymphoblastic leukemia which showed that, in some cases, reduction in large tumors was slower than in histologically similar smaller tumors. Therefore, Norton and Simon hypothesize that the cell-kill is proportional to the growth rate (e.g., exponential, logistic, or Gompertz) of the tumor. A third hypothesis notes that some chemotherapeutic drugs must be metabolized by an enzyme before being activated. This reaction is saturable due to the fixed amount of enzyme. Thus, Holford and Sheiner [4] develop the E_{max} model which describes cell-kill in terms of a saturable function of Michaelis–Menton form.

In this study, we use optimal control theory to evaluate and compare effective treatment strategies for each of these models by developing formal mathematical

*Received by the editors August 28, 2002; accepted for publication (in revised form) February 21, 2003; published electronically September 4, 2003.
http://www.siam.org/journals/siap/63-6/41348.html

†Department of Mathematics and Statistics, Murray State University, Murray, KY 42071 (renee.fister@murraystate.edu). The research of this author was supported by a KY NSF EPSCoR Research Enhancement grant.

‡Department of Pharmaceutical Sciences, St. Jude Children's Research Hospital, 332 North Lauderdale St., Memphis, TN 38105-2794 (carl.panetta@stjude.org). The research of this author was supported by Cancer Center CORE grant CA21765, a Center of Excellence grant from the State of Tennessee, and American Lebanese Syrian Associated Charities (ALSAC).

criteria to be minimized. These include tumor mass and dose of drug. We give a mathematically detailed development of optimal control forms for the various growth and drug terms that are subject to different objective functionals. We also show therapeutically significant differences between the cell-kill hypotheses and their effect on treatment schedules.

We have previously developed a treatment strategy using optimal control techniques for the use of cell-cycle specific drugs such as Taxol for the reduction of breast and ovarian cancers [5]. The model included a resting phase which made it more realistic in the clinical setting. Among other things, the model showed that treating with repeated shorter periods allows more drug to be given without excess damage to the bone marrow. Similar results were also observed in [6]. Several other models where optimal control methods have been utilized in analyzing effective chemotherapeutic treatments include Swan [7, 8] and Murray [9]. Swan [7, 8] obtained feedback treatment control drug characterizations for cancer models under a quadratic performance criterion. Murray [9] has considered systems of normal and tumor cells under the hypotheses of Gompertzian and logistic growth in which he controls the rate of administration of drugs. Murray has minimized the tumor burden at the end of treatment and, in another application, the toxicity level, defined as the area under the drug concentration curve.

2. The model. Mathematically, the general form of the model under investigation is depicted by the differential equation:

$$\frac{dN}{dt} = rNF(N) - G(N,t), \tag{2.1}$$

where N is the tumor volume, r is the growth rate of the tumor, $F(N)$ is the generalized growth function. For the proposed model, we allow for Gompertzian growth:

$$F(N) = \ln\left(\frac{\Theta}{N}\right). \tag{2.2}$$

The function $G(N,t)$ describes the pharmacokinetic and pharmacodynamic effects of the drug on the system. In this study, we compare three cell-kill strategies. These include the following:

- $G(N) = \delta u(t)N$: Skipper's log-kill (i.e., percentage kill) hypothesis,
- $G(N) = \delta u(t)N/(K+N)$: E_{max} model, and
- $G(N) = \delta u(t)F(N)$: Norton–Simon hypothesis,

where δ is the magnitude of the dose and the control, $u(t)$, describes the time dependent pharmacokinetics of the drug; i.e., $u(t) = 0$ implies no drug effect is present and $u(t) > 0$ implies the amount or strength of the drug effect. We investigated the differences and similarities among the three drug effects via optimal control techniques for ordinary differential equations.

We considered two objective functionals when determining the minimum amount of drug needed to reduce or eliminate the tumor mass. One criterion considered is

$$J_\alpha(u) = \int_0^T \left[a(N - N_d)^2 + bu^2\right]\, dt, \tag{2.3}$$

where the measure of the "closeness" of the tumor mass to the desired tumor density, N_d, and the cost of the control, $u(t)$, are minimized over the class of measurable,

nonnegative controls. Here, a and b are positive weight parameters. The second criterion we considered (previously used by Boldrini and Costa [10] with variations by Murray [11] and Martin and Teo [12]) is

$$J_\beta(u) = aN(T) + b\int_0^T u(t)\,dt, \tag{2.4}$$

in which the tumor burden at the end of treatment (the first term in (2.4)) and the toxicity (the second term in (2.4)), in terms of area under the drug concentration curve, are minimized over the class of measurable, nonnegative controls.

After scaling the models using $\overline{N} = N/\Theta$, $\overline{k} = k/\Theta$, and $\overline{\delta} = \delta/\Theta$ and dropping the bars, we have the following three state equations (which will henceforth be referred to as P1, P2, and P3, respectively), all with the same initial condition of $N(0) = N_0$, where $0 < N_0 < 1$ since the tumor cells have been normalized via the above change of variables:

P1
$$\frac{dN}{dt} = rN\ln\left(\frac{1}{N}\right) - u(t)\delta N,$$

P2
$$\frac{dN}{dt} = rN\ln\left(\frac{1}{N}\right) - u(t)\frac{\delta N}{k+N},$$

P3
$$\frac{dN}{dt} = rN\ln\left(\frac{1}{N}\right)[1 - \delta u(t)].$$

Ultimately, we determine the unique characterization of the optimal control $u(t)$ in the admissible control class,

$$U = \{u \text{ measurable } |0 \le u(t), t \in [0,T]\} \tag{2.5}$$

or

$$V = \{u \text{ measurable } |0 \le u(t) \le M, t \in [0,T]\}, \tag{2.6}$$

such that the objective functionals J_α and J_β are minimized over the class of controls, U and V, respectively.

In sections 3.1–3.3, we consider the existence issues, the characterization of the optimal control, and the uniqueness concept in association with problems P1–P3 such that the objective functional (2.3) involving the nonlinear control term is minimized over the class of controls, U. In sections 4.1–4.2, we discuss the existence of an optimal control and its characterization such that it minimizes the second objective functional (2.4) subject to each of the differential equations represented in P1–P3. Also, in section 5, numerical simulations representing the control situations in relation to the two objective functionals as well as the different cell-kill hypotheses depicted in the differential equations are analyzed.

3. Nonlinear control.

3.1. Existence. First, the existence of the state solution to each of problems P1–P3 given an optimal control in the admissible set, U, is shown. Also, the existence of the optimal control for the state system is analyzed.

THEOREM 3.1. *Given $u \in U$, there exists a bounded solution solving each of the problems* (P1)–(P3).

Proof. We consider the following differential equations in relation to P1, P2, P3, respectively. The state variables $N_1(t)$, $N_2(t)$, and $N_3(t)$ represent supersolutions for problems P1, P2, and P3.

$$\frac{dN_1}{dt} = r, \tag{3.1}$$

$$\frac{dN_2}{dt} = r + u(t)\delta N_2, \tag{3.2}$$

$$\frac{dN_3}{dt} = -rN_3 \ln N_3(1 - u(t)\delta). \tag{3.3}$$

Since $N(t) > 0$ and $\ln \frac{1}{N} \leq \frac{1}{N}$, then equation (3.3) follows from P1.

Using that $0 \leq t \leq T$, we can show that

$$N_1(t) \leq rT + N_0, \tag{3.4}$$

$$N_2(t) \leq (N_0 + rT)e^{\delta \int_0^T u(s)\,ds}, \tag{3.5}$$

$$N_3(t) \leq N_0^{\left(e^{-rt+r\delta \int_0^T u(s)\,ds}\right)}. \tag{3.6}$$

Since $u(t) \in U$, then, along with $N_1(t)$, $N_2(t)$ and $N_3(t)$ are bounded above. Via a maximum principle [13] and standard existence theory for first-order nonlinear differential equations, we obtain the existence of a solution to each of the problems P1–P3. □

Next, the existence of an optimal control for the state system is analyzed. Using the fact that the solution to each state equation is bounded, the existence of an optimal control for each problem can be determined using the theory developed by Fleming and Rishel [14].

THEOREM 3.2. *Given the objective functional, $J_\alpha(u) = \int_0^T \left[a(N - N_d)^2 + bu^2\right] dt$, where*

$$U = \{u \text{ measurable } |0 \leq u(t), t \in [0, T]\} \tag{3.7}$$

and each of the problems P1–P3 *with $N(0) = N_0$, then there exists an optimal control u^* associated with each problem* P1–P3 *such that $\min_{u \in U} J_\alpha(u) = J_\alpha(u^*)$ if the following conditions are met:*

(i) *The class of all initial conditions with a control u in the admissible control set along with each state equation being satisfied is not empty.*

(ii) *The admissible control set U is closed and convex.*

(iii) *Each right-hand side of* P1–P3 *is continuous, is bounded above by a sum of the bounded control and the state, and can be written as a linear function of u with coefficients depending on time and the state.*

(iv) *The integrand of* (2.3) *is convex on U and is bounded below by $-c_2 + c_1|u|^\eta$ with $c_1 > 0$ and $\eta > 1$.*

Proof. Since each problem has a bounded solution for the initial condition, given an optimal control, by Theorem 3.1, then part (i) is established. By definition, U is

closed and convex. To complete part (iii) we first reconsider the right-hand sides of P1–P3 below:

$$f(t,N(t),u(t)) = rN\ln\frac{1}{N} - \delta N u(t),$$
$$g(t,N(t),u(t)) = rN\ln\frac{1}{N} - \frac{\delta N}{k+N}u(t),$$
$$h(t,N(t),u(t)) = rN\ln\frac{1}{N} - \delta rN\ln\frac{1}{N}u(t).$$

We see by the representations of f, g, and h that they are continuous in t, u, and N since $N(t) > 0$. Also, they are each written as a linear function of the control with coefficients depending on time and the state. For the boundedness requirement, we use the bounds in the proof of Theorem 3.1 to obtain the result. Consequently,

$$\begin{aligned}|f(t,N(t),u(t))| &\le \left|rN\ln\frac{1}{N}\right| + |\delta N(t)u(t)|\\ &\le r + \delta|u(t)|(N_0+rT)\\ &\le C_1(1+|u(t)|+|N(t)|),\end{aligned}$$

where C_1 depends on r, d, N_0, and T,

$$\begin{aligned}|g(t,N(t),u(t))| &\le \left|rN\ln\frac{1}{N}\right| + \left|\frac{\delta N(t)}{k+N(t)}u(t)\right|\\ &\le r + \delta|u(t)|\\ &\le r + \delta|u(t)| + |N(t)|,\end{aligned}$$

and

$$\begin{aligned}|h(t,N(t),u(t))| &\le \left|rN\ln\frac{1}{N}\right| + \left|\delta rN(t)\ln\frac{1}{N}u(t)\right|\\ &\le r + \delta r|u(t)|\\ &\le C_2(1+|u(t)|+|N(t)|),\end{aligned}$$

where C_2 depends on r and d. Hence, the right-hand side of each state equation is bounded above by a sum of the control and the state. Lastly, the integrand of the objective functional is convex on U. One can consider the second partial of the integrand of the objective functional with respect to the control and find that it is positive. To obtain the necessary lower bound for the integrand, we see the $a(N-N_d)^2 + bu^2 \ge bu^2 \ge -c + bu^2$ for any $c > 0$. Therefore, part (iv) is complete and so is the proof. □

3.2. Characterization of optimal control. Since an optimal control exists for minimizing the objective functional (2.3) subject to each of the three equations P1–P3 with the initial conditions, the necessary conditions for an optimal control for each problem are determined. For brevity, we derive the conditions using a version of Pontryagin's maximum principle for P3 [15, 16]. Then we give the optimality system, which is the state system coupled with the adjoint system, for each problem.

In order to derive the necessary conditions, we define the Lagrangian associated with $J_\alpha(u)$ subject to P3 as

$$L(N,u,\lambda_3,w_1) = a(N-N_d)^2 + bu^2 + \lambda_3\Big(rN\ln\frac{1}{N}(1-\delta r u(t))\Big) - w_1(t)u(t), \tag{3.8}$$

where $w_1(t) \ge 0$ is a penalty multiplier satisfying $w_1(t)u(t) = 0$ at the optimal u^*.

Similar definitions for hold for $J_\alpha(u)$ subject to P1 and P2.

THEOREM 3.3. *Given an optimal control u^* and solution of the corresponding state equation* (P3)*, there exists an adjoint variable λ_3 satisfying the following:*

$$\frac{d\lambda_3}{dt} = -\frac{\partial L}{\partial N} = -\left[2a(N - N_d) + \lambda_3 r(1 - u\delta)\left[\ln\frac{1}{N} - 1\right]\right] \tag{3.9}$$

with $\lambda_3(T) = 0$. Further, $u^*(t)$ can be represented by

$$u^*(t) = \Big(\frac{-\lambda_3 \delta r N \ln N}{2b}\Big)^+.$$

Proof. The existence of the adjoint solution is found via a maximum principle satisfying [13]. Using the Lagrangian (3.8), we complete the representation for u^* by analyzing the optimality condition $\frac{\partial L}{\partial u} = 0$. Upon some algebraic manipulation, the representation of u^* becomes

$$u^*(t) = \frac{-\lambda_3 \delta r N \ln N + w_1}{2b}.$$

To determine an explicit expression for the optimal control, without w_1, a standard optimality technique is utilized. The optimal control is characterized as

$$u^*(t) = \Big(\frac{-\lambda_3 \delta r N \ln N}{2b}\Big)^+, \tag{3.10}$$

where

$$r^+ = \begin{cases} r & \text{if} \quad r > 0, \\ 0 & \text{if} \quad r \le 0. \end{cases} \tag{3.11}$$

Similarly, we can find the representations for the controls associated with problems P1 and P2 that are subject to J_α. The associated control for P1 is $u^*(t) = \frac{\delta}{2b}(\lambda_1 N)^+$ and the control for P2 is $u^*(t) = \frac{\delta}{2b}(\frac{\lambda_2 N}{k+N})^+$.

Using this explicit representation for the control, the adjoint equation coupled with the state equation and the initial and transversality conditions form the optimality system. The optimality systems associated with each of the state equations and their associated adjoint equations are given below. We note that Optimality System 1 is associated with P1 and its adjoint, Optimality System 2 is associated with P2 and its adjoint, and Optimality System 3 is associated with P3 and its adjoint.

Optimality System 1 (OS1).

$$\begin{aligned} \frac{dN}{dt} &= rN\ln\frac{1}{N} - N\frac{\delta^2}{2b}(\lambda_1 N)^+, \\ \frac{d\lambda_1}{dt} &= -\left[2a(N - N_d) + \lambda_1\left(r\ln\frac{1}{N} - r - \frac{\delta^2}{2b}(\lambda_1 N)^+\right)\right] \end{aligned}$$

with $N(0) = N_0$ and $\lambda_1(T) = 0$.

Optimality System 2 (OS2).

$$\begin{aligned} \frac{dN}{dt} &= rN\ln\frac{1}{N} - \frac{\delta^2}{2b}\frac{N}{k+N}\Big(\frac{\lambda_2 N}{k+N}\Big)^+, \\ \frac{d\lambda_2}{dt} &= -\left[2a(N - N_d) + \lambda_2\left(r\ln\frac{1}{N} - r - \frac{k\delta^2}{2b(k+N)^2}\Big(\frac{\lambda_2 N}{k+N}\Big)^+\right)\right] \end{aligned}$$

with $N(0) = N_0$ and $\lambda_2(T) = 0$.

Optimality System 3 (OS3).

$$\frac{dN}{dt} = rN\ln\frac{1}{N}\Big[1 - \frac{\delta^2 r}{2b}(-\lambda_3 N\ln N)^+\Big],$$
$$\frac{d\lambda_3}{dt} = -\left[2a(N-N_d) + \lambda_3 r\left(1 - \frac{\delta^2 r}{2b}(-\lambda_3 N\ln N)^+\right)\left(\ln\frac{1}{N} - 1\right)\right]$$

with $N(0) = N_0$ and $\lambda_3(T) = 0$.

3.3. Uniqueness. Here, we focus our attention on OS1 and note that similar analysis gives the uniqueness of the OS2 and OS3. The optimal control depends on the adjoint and the state variables. By proving the optimality system has a unique solution, we will argue that the optimal control is unique as well. We recognize that since the tumor mass, $N(t)$, is bounded, then the adjoint equation (3.12) has a bounded right-hand side that is dependent on the final time T. Hence, there exists a $D > 0$, depending on the coefficients of the state equation and the uniform bound for $N(t)$ such that $|\lambda_1(t)| < DT$ on [0,T].

THEOREM 3.4. *For T sufficiently small, the solution to Optimality System* 1 *is unique.*

Proof. We suppose that (N, λ_1) and (M, ψ) are two distinct solutions to OS1. Let $m > 0$ be chosen such that $N = e^{mt}v$, $M = e^{mt}w$, $\lambda_1 = e^{-mt}y$, and $\psi = e^{-mt}z$. Also, we have that $u = \frac{d}{2b}(yv)^+$ and $f = \frac{\delta}{2b}(wz)^+$. Upon substitution of the representations for N, M, λ_1, and ψ into the state and adjoint equations followed by simplification, we obtain the following equations:

$$\frac{dv}{dt} + mv = rv(-mt - \ln v) - \frac{\delta^2}{2b}v(yv)^+,$$
$$\frac{dw}{dt} + mw = rw(-mt - \ln w) - \frac{\delta^2}{2b}w(wz)^+,$$
$$\frac{dy}{dt} - my = -2av + 2ae^{mt}N_d - yr(-mt - \ln v) + ry + \frac{\delta^2}{2b}y(yv)^+,$$
$$\frac{dz}{dt} - mz = -2aw + 2ae^{mt}N_d - zr(-mt - \ln w) + rz + \frac{\delta^2}{2b}z(wz)^+$$

with $v(0) = N_0$, $w(0) = N_0$, $y(T) = 0$, and $z(T) = 0$.

The next step is to subtract the equations corresponding to v, w, y, z. Then each of these differences are multiplied by an appropriate function and are integrated from 0 to T. We obtain the following two equations for the modified state and the adjoint:

$$\frac{1}{2}[v(T) - w(T)]^2 + m\int_0^T (v-w)^2\,dt = mr\int_0^T t(v-w)^2\,dt$$
$$- r\int_0^T [v\ln v - w\ln w](v-w)\,dt$$
$$- \frac{\delta^2}{2b}\int_0^T (v(yv)^+ - w(wz)^+)(v-w)\,dt$$

and

$$\frac{1}{2}[y(0)-z(0)]^2+m\int_0^T(y-z)^2\,dt=2a\int_0^T(v-w)(y-z)\,dt$$
$$+mr\int_0^T t(y-z)^2\,dt+r\int_0^T(y-z)^2\,dt$$
$$+r\int_0^T[y\ln v-z\ln w](y-z)\,dt$$
$$+\frac{\delta^2}{2b}\int_0^T(y(yv)^+-z(wz)^+)(y-z)\,dt.$$

We need to estimate several terms in order to obtain the uniqueness result. For explanation, we include one estimate below where the boundedness of the state variables and Cauchy's inequality are used:

$$\left|\int_0^T(v\ln v-w\ln w)(v-w)\,dt\right|=\left|\int_0^T[v(\ln v-\ln w)+(v-w)\ln w](v-w)\,dt\right|$$
$$=\left|\int_0^T\Big[v\Big(\ln\frac{v}{w}\Big)(v-w)+(v-w)^2\ln w\Big]\,dt\right|$$
$$\le\int_0^T\Big[\frac{v^2}{w}|(v-w)|+|w|(v-w)^2\Big]\,dt$$
$$\le TC_1\int_0^T(v-w)^2\,dt.$$

In the estimate above, C_1 depends on the bounds of the state variables and the coefficients.

To complete this uniqueness proof, we add the two integral equations together and bounds the terms to obtain

$$\frac{1}{2}[v(T)-w(T)]^2+\frac{1}{2}[y(0)-z(0)]^2+m\int_0^T[(v-w)^2+(y-z)^2]\,dt$$
$$\le((mr+C_2)T)\int_0^T[(v-w)^2+(y-z)^2]\,dt,$$

where C_2 depends on the coefficients and the bounds of the state and the adjoint variables.

Since the variable expressions evaluated at the initial and the terminal times are nonnegative, the inequality reduces to

$$(m-mrT-C_2T)\int_0^T\Big[(v-w)^2+(y-z)^2\Big]\,dt\le 0. \tag{3.12}$$

For the optimality system to be unique, we must choose m such that $m>\frac{C_2}{1-r}$ and, thus,

$$m-mrT-C_2T>0.$$

For this choice of m we have that $T<\frac{m}{mr+C_2}$. Moreover, OS1 has a unique solution. Since the characterization of the optimal control directly depends on the state and the adjoint solutions, which are unique, then the optimal control corresponding to OS1 is unique. Similar results give uniqueness for Optimality Systems 2 and 3.

4. Linear control. In this case we still consider the same three differential equations, P1–P3, subject to their initial conditions. However, in this case, we determine the existence and the characterization of an optimal control in the admissible control class, V, such that the objective functional (2.4),

$$J_\beta(u) = aN(T) + b\int_0^T u(t)\,dt, \tag{4.1}$$

is minimized over this class of controls. The goal is to find an optimal control, u^*, such that

$$\min_{0\le u\le M} J(u) = J(u^*).$$

4.1. Existence. In subsection 3.1, we obtain the existence of the state solution for each problem P1–P3 given an optimal control in U. This work can be extended directly because the only change is that $u(t)$ is bounded above by a maximum amount of drug M.

For simpler discussions, we transform the original problems P1–P3 via $x = \ln N$. Consequently, we minimize

$$J_1(u) = ae^{x(T)} + b\int_0^T u(t)\,dt \tag{4.2}$$

over the class of admissible controls, V, subject to each of the three differential equations that correspond to P1, P2, and P3, respectively. We note that $k \neq N_0$ and that the initial condition is $x(0) = \ln N_0$ and is negative since $0 < N_0 < 1$.

$$\frac{dx}{dt} = -rx - u(t)\delta, \tag{4.3}$$

$$\frac{dx}{dt} = -rx - u(t)\frac{\delta}{k+e^x}, \tag{4.4}$$

$$\frac{dx}{dt} = rx(u(t)\delta - 1). \tag{4.5}$$

The theorem for the existence of an optimal control for the appropriate objective functional is stated below. Since the proof involves standard arguments, it is omitted. For further information, see [14].

THEOREM 4.1. *There exists an optimal control in V that minimizes the objective functional $J_1(u)$ subject to* (4.3)*,* (4.4)*, and* (4.5)*, respectively.*

4.2. Characterization of optimal control. Since an optimal control exists, we determine the characterization for each optimal control $u(t)$ associated with each problem (4.3), (4.4), and (4.5) that minimizes $J_1(u)$. We use Pontryagin's maximum principle [17] to obtain the necessary conditions for optimality for each problem.

THEOREM 4.2. *Given an optimal control $u^*(t)$ and a solution, $x(t)$ to* (4.3)*, there exists an adjoint ψ_1 satisfying*

$$\frac{d\psi_1}{dt} = r\psi_1(t),$$
$$\psi_1(T) = ae^{x(T)}.$$

Furthermore,

$$u^*(t) = \begin{cases} M & \text{if} \quad \psi_1(t) > \frac{b}{\delta}, \\ 0 & \text{if} \quad \psi_1(t) < \frac{b}{\delta}. \end{cases} \tag{4.6}$$

Note that this is a problem similar to Swierniak and Duda [18]. We note that $\psi_1(t) = ae^{x(T)-r(T-t)}$ and a singular control could not exist. (A singular control exists when the Hamiltonian is linear in the control and the coefficient of the control is zero for some time interval.) If one were to exist, then $\psi_1(t)$ must equal $\frac{b}{\delta}$ on some interval inclusive to [0,T]. This cannot occur since $\psi_1(t)$ would be constant only for one instant in time. Furthermore, Swierniak and Duda provide conditions for the representation of the control $u^*(t)$ in terms of the model parameters. Please see [18, 19] for complete details.

For problem (4.4), we have the following result.

THEOREM 4.3. *Given an optimal control $u^*(t)$ and corresponding solution $x^*(t)$ to* (4.4)*, there exists an adjoint ψ_2 satisfying*

$$\frac{d\psi_2}{dt} = \psi_2(t)\Big[r - \frac{u(t)\delta e^{x(t)}}{(k+e^{x(t)})^2}\Big], \tag{4.7}$$

$$\psi_2(T) = ae^{x^*(T)}. \tag{4.8}$$

In addition,

$$u^*(t) = \begin{cases} 0 & \text{if} \quad b - \frac{\psi_2(t)\delta}{(k+e^{x^*(t)})} > 0, \\ M & \text{if} \quad b - \frac{\psi_2(t)\delta}{(k+e^{x^*(t)})} < 0. \end{cases} \tag{4.9}$$

Proof. To determine the representation for $u^*(t)$ and the differential equation associated with $\psi_2(t)$, we form the Hamiltonian. We note via Pontryagin's maximum principle [17] that if u^* is an optimal control associated with a corresponding trajectory on $[0,T]$, then there exists $\lambda_0 \geq 0$ and an absolutely continuous function $\lambda : [0,T] \to R$ such that $(\lambda_0, \lambda(t) \neq (0,0))$ for all $t \in [0,T]$ and $\lambda(t)$ satisfies (4.7) and $\lambda(T) = \lambda_0 ae^{x(T)}$. This optimal control minimizes $H = \lambda_0 bu(t) + \lambda(-rx(t) - \frac{u(t)\delta}{(k+e^{x(t)})^2})$ over V. Yet, λ_0 cannot vanish for this problem because, if it did, then $\lambda(T) = 0$ and hence $\lambda(t) = 0$ on $[0,T]$. This contradicts the nontriviality of the multipliers. Therefore, without loss of generality, we let $\lambda_0 = 1$.

Consequently, we consider the following Hamiltonian, where we omit the asterisks for simplicity:

$$H(t, x(t), u(t), \psi_2(t)) = bu(t) + \psi_2(t)\Big(-rx(t) - \frac{u(t)\delta}{(k+e^{x(t)})}\Big).$$

We note that from standard existence theory we obtain the existence of $\psi_2(t)$ solving (4.7) given that $x(t)$ is bounded. The necessary conditions of optimality give that

$$u(t) = \begin{cases} 0 & \text{if} \quad b - \frac{\psi_2(t)\delta}{(k+e^{x^*(t)})} > 0, \\ M & \text{if} \quad b - \frac{\psi_2(t)\delta}{(k+e^{x^*(t)})} < 0, \\ \text{singular} & \text{if} \quad b - \frac{\psi_2(t)\delta}{(k+e^{x^*(t)})} = 0. \end{cases} \tag{4.10}$$

We note that

$$\psi_2(t) = ae^{x(T) - \int_t^T \left[r - \frac{u(t)\delta e^{x(t)}}{(k+e^{x(t)})^2}\right] ds}$$

is always positive on [0,T] since $a > 0$.

We suppose that the control is singular on $(t_0, t_1) \subset [0, T]$, i.e.,

$$b - \frac{\psi_2(t)d}{(k + e^{x(t)})} = 0 \tag{4.11}$$

on that interval. We take the derivative with respect to time of (4.11) and obtain, after simplification,

$$(k + e^{x(t)})\frac{d\psi_2(t)}{dt} - \psi_2(t)e^{x(t)}\frac{dx}{dt} = 0. \tag{4.12}$$

Next, we substitute the right-hand sides of the differential equations for $\psi_2(t)$ and for $x(t)$ associated with problem (4.4) and find that

$$(k + e^{x(t)})\Big[\psi_2(t)\Big(r - \frac{u(t)\delta e^{x(t)}}{(k + e^{x(t)})^2}\Big)\Big] - e^x\psi_2(t)\Big(-rx(t) - \frac{u(t)\delta}{(k + e^{x(t)})}\Big) = 0,$$
$$\psi_2(t)r(k + e^{x(t)} + x(t)e^{x(t)}) = 0.$$

Since $\psi_2(t) > 0$ on [0,T] and $r > 0$, then

$$k + e^{x(t)} + x(t)e^{x(t)} = 0. \tag{4.13}$$

This immediately gives that $x(t)$ is constant.

Since $x(t)$ is constant, then $u(t)$ is constant here. We make note that with the fixed final time that $H(t, x, u, \psi_2)$ is constant [20]. Since $H(t, x, u, \psi_2) = bu + \psi_2\frac{dx}{dt} = bu$ in this case and we are minimizing H, then $u(t) = 0$. However, for this to occur $b - \frac{\psi_2\delta}{(k+e^x)} > 0$, in (4.10), which contradicts our original assumption for the control to be singular. Thus, a singular control does not exist and our control is

$$u^*(t) = \begin{cases} 0 & \text{if} \quad b - \frac{\psi_2(t)\delta}{(k+e^{x^*(t)})} > 0, \\ M & \text{if} \quad b - \frac{\psi_2(t)\delta}{(k+e^{x^*(t)})} < 0. \end{cases} \tag{4.14}$$

□

We now determine the characterization of the optimal control to minimize $J_1(u)$ subject to (4.5).

THEOREM 4.4. *Given an optimal control u^* and a corresponding solution $x^*(t)$ to* (4.5)*, there exists an adjoint $\psi_3(t)$ satisfying*

$$\frac{d\psi_3}{dt} = -\psi_3(t)r(u^*(t)\delta - 1), \tag{4.15}$$

$$\psi_3(T) = ae^{x^*(T)} \tag{4.16}$$

with

$$u^*(t) = \begin{cases} 0 & \textit{if} \quad x(T)e^{x(T)} > \frac{-b}{a\delta r}, \\ M & \textit{if} \quad x(T)e^{x(T)} < \frac{-b}{a\delta r}. \end{cases} \tag{4.17}$$

Proof. To determine the representation for the control, we form the Hamiltonian in a similar fashion as we did in the proof of Theorem 4.3,

$$H(t, x(t), u(t), \psi_3(t)) = bu(t) + \psi_3(t)(rx(t)(u(t)\delta - 1)).$$

As before, a solution to the adjoint exists and is given by

$$\psi_3(t) = ae^{x(T)-r\int_t^T (1-u(s)\delta)\,ds}, \tag{4.18}$$

and a solution to the problem (4.5) is

$$x(t) = x(0)e^{-rt+r\delta\int_0^t u(s)\,ds}. \tag{4.19}$$

Since $x(0) < 0$ and $a > 0$, then $\psi_3(t)x(t) < 0$ on [0,T].

The necessary conditions for optimality give that

$$u(t) = \begin{cases} 0 & \text{if} \quad b + \psi_3(t)\delta r x(t) > 0, \\ M & \text{if} \quad b + \psi_3(t)\delta r x(t) < 0, \\ \text{singular} & \text{if} \quad b + \psi_3(t)\delta r x(t) = 0. \end{cases} \tag{4.20}$$

We see that

$$\begin{aligned} \psi_3(t)x(t) &= (ae^{x(T)-r\int_t^T (1-u(s)\delta)\,ds})(x(0)e^{-rt+r\delta\int_0^t u(s)\,ds}) \\ &= ax(0)e^{x(T)-r(T-t)+r\delta\int_t^T u(s)\,ds-rt+r\delta\int_0^t u(s)\,ds} \\ &= ax(0)e^{x(T)-rT+r\delta\int_0^T u(s)\,ds} \\ &= ax(T)e^{x(T)}. \end{aligned}$$

Hence, $\psi_3(t)x(t)$ is constant on [0,T]. This means that $u^*(t)$ must be either zero, its maximum value—M, or its singular representation on the entire interval [0,T].

Using that $\psi_3(t)x(t)$ is constant and that the Hamiltonian is to be minimized, we can exclude the singular case. If the control is singular, then the Hamiltonian is equal to $-\psi_3(t)x(t)r$. We can see for the case, $u = M$, that $H(t, x(t), u(t), \psi_3(t)) < -\psi_3(t)x(t)r$. Moreover, for the case $u = 0$ we see that the Hamiltonian is bounded strictly above by $\frac{b}{\delta}$, which is the value of the Hamiltonian if the control is singular. Consequently, a singular control will not generate the minimum value for the Hamiltonian.

Therefore, the necessary conditions for optimality are

$$u^*(t) = \begin{cases} 0 & \text{if} \quad x(T)e^{x(T)} > \frac{-b}{a\delta r}, \\ M & \text{if} \quad x(T)e^{x(T)} < \frac{-b}{a\delta r}. \end{cases} \tag{4.21}$$

We simply need to check if the expression $x(T)e^{x(T)}$ is smaller or larger than $\frac{-b}{ar\delta}$. Then this determines the constant control on the interval [0,T]. □

Using the representation of the control in terms of the state and adjoint solutions to the transformed problems (4.3), (4.4), and (4.5), we have the associated Optimality Systems 4, 5, and 6.

Optimality System 4 (OS4).

$$\begin{aligned} \frac{dx}{dt} &= -rx(t) - \delta u^*(t), \\ \frac{d\psi_1}{dt} &= r\psi_1(t), \\ x(0) &= \ln N_0, \text{ and } \psi_1(T) = ae^{x(T)}, \end{aligned}$$

where

$$u^*(t) = \begin{cases} M & \text{if} \quad \psi_1(t) > \frac{b}{\delta}, \\ 0 & \text{if} \quad \psi_1(t) < \frac{b}{\delta}. \end{cases}$$

Optimality System 5 (OS5).

$$\begin{aligned} \frac{dx}{dt} &= -rx(t) - \frac{\delta u^*(t)}{(k + e^{x(t)})}, \\ \frac{d\psi_2}{dt} &= \psi_2\Big[r - \frac{u^*(t)\delta e^{x(t)}}{(k + e^{x(t)})^2}\Big], \\ x(0) &= \ln N_0, \text{ and } \psi_2(T) = ae^{x(T)}, \end{aligned}$$

where

$$u^*(t) = \begin{cases} 0 & \text{if} \quad b - \frac{\psi_2(t)\delta}{(k+e^{x^*(t)})} > 0, \\ M & \text{if} \quad b - \frac{\psi_2(t)\delta}{(k+e^{x^*(t)})} < 0. \end{cases}$$

Optimality System 6 (OS6).

$$\begin{aligned} \frac{dx}{dt} &= rx(t)(u^*(t)\delta - 1), \\ \frac{d\psi_3}{dt} &= -r\psi_3(t)(u^*(t)\delta - 1), \\ x(0) &= \ln N_0, \text{ and } \psi_3(T) = ae^{x(T)}, \end{aligned}$$

where

$$u^*(t) = \begin{cases} 0 & \text{if} \quad x(T)e^{x(T)} > \frac{-b}{a\delta r}, \\ M & \text{if} \quad x(T)e^{x(T)} < \frac{-b}{a\delta r}. \end{cases}$$

5. Numerical results. We obtain numerical solutions to each of the optimality systems using the two-point boundary value solver in Matlab [21, 22].

5.1. OS1–OS3. We observe several interesting differences among the three systems. First, there are significant differences between the optimal solutions of systems OS1–OS3 based on the initial tumor volume. For initial conditions near the carrying capacity (97.5% of carrying capacity), the optimal solutions for OS1 require less treatment to reduce the tumor volume the same amount as in OS2 or OS3 (Figure 5.1(A), (B)). In particular, the optimal solution for OS1 allows for the treatment to be reduced quickly while for OS3 the treatment remains higher for a more extended period. This difference is due to the different cell-kill hypotheses of the three methods. In particular, OS1 hypothesizes that cell-kill is proportional to tumor size, thus larger tumors are effectively reduced by the drug and the optimality scheme can quickly reduce the dose needed to keep the tumor size small over the treatment interval. However, OS3 hypothesizes that cell-kill is proportional to the growth rate, which is small when the tumor is near its carrying capacity. Therefore, the optimality scheme requires more drug for a longer period of time to reduce the tumor an equivalent amount as in OS1.

For initial conditions at half the carrying capacity the differences between the optimal solutions of the three strategies are less significant (Figure 5.1(C), (D)). OS3

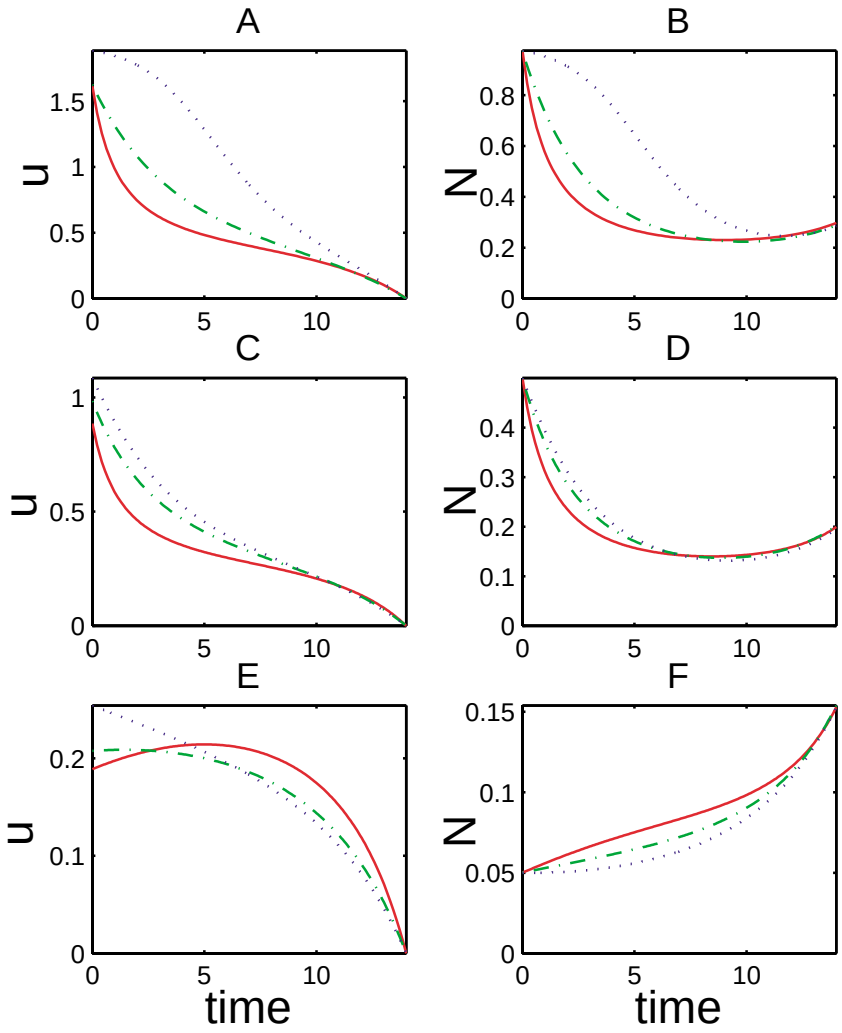


FIG. 5.1. *Comparison of hypotheses OS*1 ——, *OS*2 — - —, *and OS*3 - - -. *The parameters used are given in Table* 5.1. *The left column is the control and the right column is the tumor volume.* A *and* B *represent a large tumor near its carrying capacity (*97.5% *of carrying capacity).* C *and* D *represent a midsized tumor at half its carrying capacity (*50% *of carrying capacity).* E *and* F *represent a small tumor relative to its carrying capacity (*5% *of carrying capacity). In all cases the parameter "δ" was set such that all three methods would have the same tumor volume at the end time "T." The units for time is days.*

still requires more drug early, when the growth rate was slower, to obtain the same overall cell-kill as OS1, but the differences are much smaller.

As for smaller tumors where the initial conditions are much smaller than the carrying capacity (5% of carrying capacity), the optimal solution for OS1 requires more drug later compared to the optimal solution for OS3 to have the same effect on the tumor volume (Figure 5.1(E), (F)). But again, the differences in this case are much less significant compared to tumors near carrying capacity.

We also consider the effects of the weights a, b, and δ in the objective function (2.3). In particular, we fix b and considered how varying a and δ affects results.

TABLE 5.1
Model parameters. All units are nondimensional.

Parameter	OS1	OS2	OS3
r	0.1	0.1	0.1
δ	0.45	0.225	4.0
a	3	3	3
b	1	1	1
N_d	0	0	0
k		0.25	

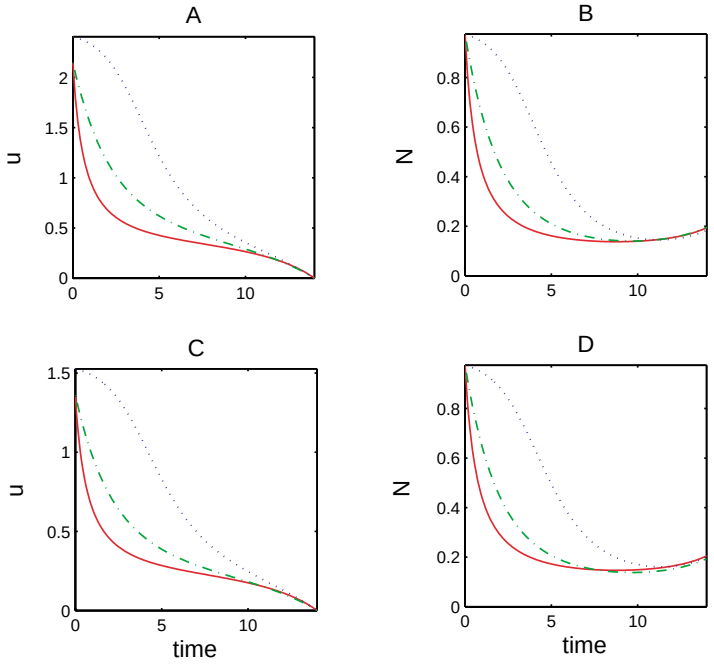


FIG. 5.2. *OS1 ——, OS2 — - —, and OS3 - - -. In* A *and* B *the parameter "a" is larger (*2.5 *times) and the parameter "δ" is smaller (*50%–60%*) compared to* C *and* D*. In all cases the initial condition for the tumor size is* 97.5% *of the carrying capacity and the remaining parameters are given in Table* 5.1*.*

(The remaining parameters are given in Table 5.1.) In general, a smaller a, 2.5 times smaller (i.e., less weight in the objective function on minimizing the tumor volume) requires more drug (or a more effective drug) via an increase in δ (50%–60% increase) to give equivalent results (Figure 5.2). In general, a, b, and δ alter the quantitative but not the qualitative results.

5.2. OS4–OS6. Systems OS4–OS6, with the linear control functional (2.4), give qualitatively different results compared to OS1–OS3. For example, OS4 and OS5 force treatment to be delayed until the later portion of the treatment interval (Figure 5.3). This delay in treatment is due to the choice of the objective functional (2.4), which minimizes N at the end time T. Since (based on the Gompertz growth model) larger tumors grow much slower than smaller ones, it is more advantageous for the tumor to remain larger (and thus slower growing) for the first portion of the treatment

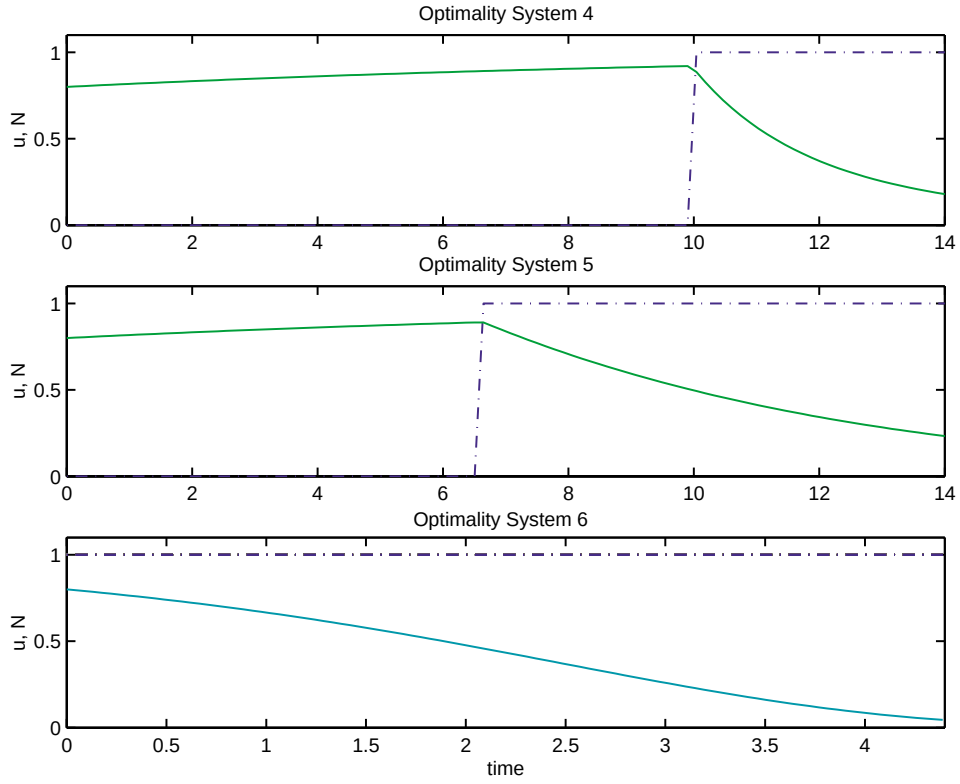


FIG. 5.3. *Tumor volume "N" ——, control "u(t)" — - —. Note that for OS6 the graph shows only the portion where the treatment is "on."*

interval and treat during the remaining portion of the treatment interval. If instead the treatment is given over the first portion of the treatment interval, then the tumor would be able to recover during the second portion of the interval at a much faster rate (since the smaller tumor after treatment has a faster growth rate) and thus not optimize the objective functional (2.4).

OS6 could only determine the length of treatment and not when treatment should be started or stopped due to the end condition being only dependent on the end time T (Figure 5.3). However, there are also some similarities between OS1–OS3 and OS4–OS6. These include OS4, which like OS1 hypothesizes that cell-kill is proportional to the tumor volume and requires drug to be given over a shorter period (3 days) compared to the other two methods (OS5 approximately 7.5 days and OS6 approximately 4.5 days).

6. Conclusions.

Theoretical. There are several important differences in the objective functionals that are minimized over the two classes of controls. For the objective functional with the quadratic control (2.3), the representation for the optimal control involves both the state and the adjoint variables for all time, t. For the second objective functional (2.4), the control is explicitly dependent on the adjoint, which in turn does implicitly depend on the state, for OS4 and OS5. In OS6, the control is dependent only on the final time and the evaluation of the state at the final time.

Within OS1–OS3, the existence, uniqueness, and characterization of the optimal control are easier to obtain because of the nonlinear control. The biological validity of the quadratic cost term has been debated [7]. However, its use in order to incorporate the nonlinear flavor of the problem has given results that are qualitatively significant. For OS4–OS6, an interesting component is the bang-bang structure of the optimal controls. An intriguing difference occurs in OS6 in which the optimal control depends on the length of the treatment. Once this is known, the control will either be given at the maximum or the minimum level. A future problem relating to OS6 could involve the minimization of the time of treatment in conjunction with the minimization of the drug needed.

The mathematical significance of these problems lies in the similarities and differences of the construction of the controls. Basic concepts using a Lagrangian or Hamiltonian are needed in each. Yet, the possibility of singular controls in section 4 required more explorations of the control representations. In this investigation, the strategies employed in both the nonlinear and the linear control settings attempt to qualitatively answer the questions relating to the appropriate drug treatment to impose under the given three hypotheses of cell-kill.

Clinical. The most important clinical question that this study addresses is, When can drug treatment be reduced to reduce toxicity? If the log-kill hypothesis is used, then the optimal control systems suggest that treatment can be given for a shorter period of time relative to the Norton–Simon hypothesis. However, the consequence of choosing the incorrect hypothesis is to either under- or over-treat the patient, causing ineffective reduction of the tumor or toxicity, respectively. Therefore, more studies are needed to determine the specific dynamics of various drugs on tumors relative to these (or other) cell-kill hypotheses.

We also observe qualitatively different treatment strategies based on the use of different objective functionals. These differences show the importance of defining an objective functional that most accurately reflects the toxicities of a particular drug along with the objective of the treatment strategy, e.g., reduce the tumor mass at the end of the treatment interval, reduce the overall tumor burden over the treatment interval, or some other clinically relevant criteria.

REFERENCES

[1] F. Schabel, Jr., H. Skipper, and W. Wilcox, *Experimental evaluation of potential anticancer agents.* XIII. *On the criteria and kinetics associated with curability of experimental leukemia*, Cancer Chemo. Rep., 25 (1964), pp. 1–111.

[2] L. Norton and R. Simon, *Tumor size, sensitivity to therapy, and design of treatment schedules*, Cancer Treat. Rep., 61 (1977), pp. 1307–1317.

[3] L. Norton and R. Simon, *The Norton-Simon hypothesis revisited*, Cancer Treat. Rep., 70 (1986), pp. 163–169.

[4] N. Holford and L. Sheiner, *Pharmacokinetic and pharmacodynamic modeling* in vivo, CRC Crit. Rev. Bioeng., 5 (1981), pp. 273–322.

[5] K. R. Fister and J. C. Panetta, *Optimal control applied to cell-cycle-specific chemotherapy*, SIAM J. Appl. Math., 60 (2000), pp. 1059–1072.

[6] J. C. Panetta and J. A. Adam, *A mathematical model of cycle-specific chemotherapy*, Math. Comput. Modelling, 22 (1995), pp. 67–82.

[7] G. Swan, *General applications of optimal control theory in cancer chemotherapy*, IMA J. Math. Appl. Med. Biol., 5 (1988), pp. 303–316.

[8] G. W. Swan, *Role of optimal control theory in cancer chemotherapy*, Math. Biosci., 101 (1990), pp. 237–284.

[9] J. M. Murray, *Optimal control for a cancer chemotherapy problem with general growth and loss functions*, Math. Biosci., 98 (1990), pp. 273–287.

[10] J. L. Boldrini and M. I. S. Costa, *The influence of fixed and free final time of treatment of optimal chemotherapeutic protocols*, Math. Comput. Modelling, 27 (1998), pp. 59–72.
[11] J. M. Murray, *Some optimal control problems in cancer chemotherapy with a toxicity limit*, Math. Biosci., 100 (1990), pp. 49–67.
[12] R. Martin and K. Teo, *A worst-case optimal parameter selection model of cancer chemotherapy*, IEEE Trans. Biomed. Eng., 39 (1992), pp. 1081–1085.
[13] M. Protter and H. Weinberger, *Maximum Principles in Differential Equations*, Prentice–Hall, Englewood Cliffs, NJ, 1967.
[14] W. H. Fleming and R. W. Rishel, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York, 1975.
[15] M. Kamien and N. Schwartz, *Dynamic Optimization: The Calculus of Variations and Optimal Control in Economics and Management*, North–Holland, Amsterdam, 1991.
[16] F. H. Clarke, *Optimization and Nonsmooth Analysis*, SIAM, Philadelphia, 1990.
[17] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mishchenko, *Mathematical Theory of Optimal Processes*, Macmillan, New York, 1964.
[18] A. Swierniak and Z. Duda, *Singularity of optimal control in some problems related to optimal chemotherapy*, Math. Comput. Modelling, 19 (1994), pp. 255–262.
[19] A. Swierniak and Z. Duda, *Some control problems for simplest differential models of proliferation cycle*, Appl. Math. Comput. Sci., 4 (1994), pp. 223–232.
[20] D. Kirk, *Optimal Control Theory: An Introduction*, Prentice–Hall, Englewood Cliffs, NJ, 1970.
[21] *Matlab*, Version 6.1, The MathWorks, Inc., Natick, MA, 2002.
[22] L. Shamine, J. Kierzenka, and M. W. Reichelt, *Solving Boundary Value Problems for Ordinary Differential Equations in Matlab with bvp4c*, 2000. Available via anonymous ftp from ftp://ftp.mathworks.com/pub/doc/papers/bvp/.

SIAM J. APPL. MATH.
Vol. 63, No. 6, pp. 1972–1997

ON THE MOTION OF SOLIDS THROUGH AN IDEAL LIQUID: APPROXIMATED EQUATIONS FOR MANY BODY SYSTEMS*

CLODOALDO GROTTA RAGAZZO†

Abstract. The problem of motion of many solids through an unbounded ideal liquid (inviscid and irrotational) is considered. A Lagrangian formulation of the equations of motion leads to a set of ordinary differential equations (ODEs) coupled to an elliptic partial differential equation (PDE) [H. Lamb, *Hydrodynamics*, 6th ed., Dover, New York, 1932]. Here, using a variational approach, an approximated solution for the PDE is presented, and the problem is reduced to the study of a system of ODEs. As a consequence one can get approximate forces and torques due to hydrodynamic interaction of rigid bodies of arbitrary shapes. Some examples are discussed at the end.

Key words. multibody systems, fluid-solid interactions

AMS subject classifications. 74F10, 76B99, 70E55

DOI. 10.1137/S003613990139427X

1. Introduction and main result. This paper is concerned with the motion of solids through an infinite mass of an ideal liquid (inviscid and irrotational). Since the 1800s this problem has been treated in Lagrangian form considering the solids and the liquid as a single system in which dynamics are determined only by their inertia. The equations of motion obtained in this way are given by a set of ordinary differential equations (ODEs) (borne upon the rigid bodies) coupled to an elliptic partial differential equation (PDE) (borne upon the fluid). The fundamental contributions to the subject were given by Kirchhoff, Thomson (Lord Kelvin), and Tait. The book of Lamb [1] contains an excellent presentation of these and related works. A critical discussion on the physical validity of the hypotheses behind this classical approach can be found in the book by Birkhoff [2, paragraph 109].

The problem of motion of solids in ideal liquids has recently received renewed interest, in particular, due to its significance in the study of two phase flows (see, for instance, [3], [4], [5], [6]) and in the study of motion of submerged bodies (see, for instance, [7], [8], [9]). In a few situations of very simple geometry (all bodies are balls) and under the hypothesis of not close approach, some authors were able to approximately solve the elliptic PDE in the equations of motion and to get an explicit set of ODEs for the dynamics of the rigid bodies (see, for instance, [1], [10], [4], [5], [6], [7]). This was particularly important in numerical studies of systems with a great number of bodies (for instance, [4], [6]). In this paper the idea of approximately solving the elliptic part of the problem is extended to systems containing bodies of arbitrary shape. The goal is to obtain an explicit set of ODEs (and/or the Lagrangian function that determines it) for the dynamics of the rigid bodies. This finite set of ODEs depends on geometric parameters, volume, center of volume, and "added-mass" coefficients of each body. In general (except for balls and ellipsoids) these added-mass coefficients are yet to be determined numerically through the solution of an elliptic

*Received by the editors August 22, 2001; accepted for publication (in revised form) March 11, 2003; published electronically September 4, 2003. This work was partially supported by FAPESP (São Paulo-Brazil) grant 96/12284-6 and PETROBRAS (the Brazilian oil company) PROCAP2000. The author is partially supported by CNPq (Brazil) grant 301817/96-0.

http://www.siam.org/journals/siap/63-6/39427.html

†Instituto de Matemática e Estatística, Universidade de São Paulo, 05508-090, São Paulo, SP, Brazil (ragazzo@ime.usp.br).

PDE. Therefore, the main contribution given here is the reduction of the original coupled ODE-PDE system for the dynamics of rigid bodies in a fluid to a finite system of ODEs with some time-independent parameters. These parameters can be determined by solving a finite number of PDEs only once. Notice that in the original problem at each time step of integration of the ODE, an elliptic PDE has to be solved in a domain that changes according to the motion of the bodies.

The hypotheses assumed in this paper are very restrictive from a physical point of view. A real fluid has viscosity and its dynamic is in most cases very different from that of an inviscid irrotational fluid. A first motivation for this paper was research on the dynamics of ships used as oil floating production storage and offloading (FPSO) units (see, for instance, [11] [12]) carried out by a group of people at the University of São Paulo and PETROBRAS (the Brazilian oil company) under the leadership of J. A. P. Aranha. This group has developed a physical model for the forces and torques (due to current, wind, and waves) acting on a single FPSO under some typical environmental conditions. The hydrodynamic part of the model was developed in two steps. At first expressions for the force and torque functions, depending on a minimal number of unknown parameters, were obtained by means of a theoretical analysis. The guideline for this analysis was a decomposition proposed by Lighthill [13] of the forces acting on the ship into a viscous drag force and an inviscid inertial force (see also [14] and [9]). Then the unknown parameters were determined using either statistical data obtained from similar ships or towing tank experiments. During an offloading operation two ships get close to each other. In this case a new hydrodynamic interaction force between the ships appears. However, in this case not even an estimate of the inviscid potential part of the hydrodynamic interaction functions was available in the literature, at least in a form sufficiently explicit to be used as it was in the single ship case. Therefore, the main motivation for this paper was to come up with such an estimate. Following the general idea of Lighthill's force decomposition the expressions given here provide a starting point for the buildup of more realistic expressions for the force and torque interaction functions. In this context it is clear the importance of having explicit formulas for bodies of arbitrary shape. Of course many terms will have to be incorporated to these interaction functions due to viscous effects as it was in the single ship case. In the end tests and adjustments of the model will have to be done experimentally.

A second more direct application of the results of this paper is to the problem of hydrodynamic interaction between fast oscillating bodies. For instance, consider a system of two bodies inside a fluid. To the first body a fast small amplitude oscillation is imposed. The second one is allowed to move freely. How do they interact? Do they attract or repel each other? These questions were addressed in [15]. There the equations of motion were conveniently averaged and both Theorems 1.2 and 1.4 presented below played an important role. (In the context of the previous paragraph only Theorem 1.2 is relevant.) In this case the fluid flow induced by the fast oscillating body has a small "particle displacement/particle acceleration" ratio, so that the inertial forces are quite relevant (see [2, paragraph 103]). On the other hand eddies are not generated near the bodies (there are no wakes) due to their small fast net displacements. So, eddy formation, the usual major way of body-fluid interaction through viscosity, is not relevant in this case. There may also be the influence of viscosity due to skin friction and of fluid compressibility if the imposed oscillation of the first body is too fast (acoustic effects). However, at least for certain frequency ranges of oscillation, these are second order effects. So, in this context the force and torque expressions given here seem to be the physically most relevant part of the real

hydrodynamic interaction functions.

The hypotheses and notation used in this paper are as follows. Let us consider a system of N bounded rigid bodies whose boundaries are smooth surfaces (or, more precisely, *the boundaries are continuously differentiable*). Each body will be labeled by a Greek letter $\alpha \in \{1, 2, \dots, N\}$. In each body we fix a reference point and a three orthogonal reference frame centered on it. We denote by K_α the reference frame of body α and by $\{\vec{e}_{\alpha 1}, \vec{e}_{\alpha 2}, \vec{e}_{\alpha 3}\}$ its unit vectors. We also consider a space fixed three orthogonal reference frame K and denote by $\{\vec{e}_1, \vec{e}_2, \vec{e}_3\}$ its unit vectors. The configuration of the system is determined by N position vectors $\vec{R}_\alpha$ and N orthogonal transformations T_α, where $\vec{R}_\alpha$ is the position of the reference point of solid α with respect to K and $T_\alpha : K_\alpha \to K$ describes the orientation of body α with respect to K. (T_α is the attitude matrix of body α.) The set of points of body α will be denoted by B_α (sometimes we will just use B_α to refer to the body itself), and the boundary of B_α will be denoted by ∂B_α. The bodies will be supposed to move without collision through an infinite mass of an ideal liquid (inviscid, incompressible) of density ρ which is at rest at infinity. It will be assumed that the motion of the liquid is entirely due to its interaction with the solid, namely there is no external forces acting on the fluid. The fluid flow will be supposed irrotational, and if a body is bounded by a multiply connected surface, then it will be supposed that there is no circulation about any irreducible path on this surface.

We start the study of the dynamics of a system of solids in a fluid by the simplest case of a single body. Let α be the index of this body. Under the above hypothesis it can be shown [1] that the velocity field $\vec{u}$ of the fluid is the gradient of a potential function Φ, $\vec{u} = \nabla\Phi$. For each time t the incompressibility of the liquid implies that the Laplacian of Φ is equal to zero, $\Delta\Phi = 0$. In addition to this equation, it is well known that the following boundary conditions completely determine $\vec{u} = \nabla\Phi$: (1) the components of $\nabla\Phi$ normal to the body surface points are equal to the normal velocity of these points, and (2) the fluid is at rest at infinity. Moreover, these boundary conditions imply that $||\nabla\Phi||$, $\vec{x} \in K$, is of order $1/||\vec{x}||^3$ as $||\vec{x}|| \to \infty$. The time dependence enters parametrically in this elliptic problem. The total kinetic energy of the system solid plus fluid is given by

$$W = \text{kinetic energy of body} + \frac{\rho}{2}\int_{\mathbb{R}^3 - B_\alpha} ||\nabla\Phi||^2,$$

where B_α denotes the set of points of solid α. The position and velocity of each point of B_α are determined by the position vector $\vec{R}_\alpha$, the attitude matrix T_α, the velocity vector $\dot{\vec{R}}_\alpha \stackrel{\text{def}}{=} \vec{V}_\alpha$, and the angular velocity vector $\vec{\Omega}_\alpha$, whose components in the fixed reference frame K are given by

$$\dot{T}_\alpha T_\alpha^\dagger \stackrel{\text{def}}{=} \begin{pmatrix} 0 & -\Omega_{\alpha 3} & \Omega_{\alpha 2} \\ \Omega_{\alpha 3} & 0 & -\Omega_{\alpha 1} \\ -\Omega_{\alpha 2} & \Omega_{\alpha 1} & 0 \end{pmatrix},$$

where $T_\alpha^\dagger$ is the transpose of matrix T_α (which coincides with its inverse because T_α is orthogonal). Since $\nabla\Phi$ is determined by positions and velocities of points in the boundary of B_α, which we denote by ∂B_α, we conclude that the kinetic energy W is completely determined by $\vec{R}_\alpha$, $\vec{V}_\alpha$, T_α, and $\vec{\Omega}_\alpha$. In order to write W explicitly in terms of these positions and velocities it is convenient to represent them in the body reference frame K_α. Using the transformation $T_\alpha : K_\alpha \to K$ we define

$$\vec{r}_\alpha = T_\alpha^{-1}\vec{R}_\alpha, \qquad \vec{v}_\alpha = T_\alpha^{-1}\vec{V}_\alpha, \qquad \vec{\omega}_\alpha = T_\alpha^{-1}\vec{\Omega}_\alpha.$$

Sometimes it will be convenient to denote the components of $\vec{v}_\alpha$ and $\vec{\omega}_\alpha$ in a combined way, so we define

$$c_{\alpha i} = v_{\alpha i}, \qquad c_{\alpha(i+3)} = \omega_{\alpha i}, \qquad i = 1,2,3 \tag{1.1}$$

(we will always use Latin indices i, j, l, k to denote vector components). It is also convenient to use the following decomposition for Φ (see [1, article 118]):

$$\Phi(T_\alpha \vec{x}_\alpha) = \sum_{i=1}^{6} c_{\alpha i}\phi_{\alpha i}(\vec{x}_\alpha), \qquad \vec{x}_\alpha \in K_\alpha,$$

where for $i = 1,2,3$

$$\begin{cases} \Delta\phi_{\alpha i}(\vec{x}_\alpha) = 0 & \text{for} \quad \vec{x}_\alpha \in \mathbb{R}^3 - B_\alpha, \\ \nabla\phi_{\alpha i}(\vec{x}_\alpha)\cdot\vec{n}_\alpha(\vec{x}_\alpha) = \vec{\mathrm{e}}_{\alpha i}\cdot\vec{n}_\alpha(\vec{x}_\alpha) & \text{for} \quad \vec{x}_\alpha \in \partial B_\alpha, \\ ||\nabla\phi_{\alpha i}(\vec{x}_\alpha)|| \to 0 & \text{as} \quad ||\vec{x}_\alpha|| \to \infty \end{cases} \tag{1.2}$$

($\vec{n}_\alpha(\vec{x}_\alpha) \in K_\alpha$ denotes the unit normal vector at the point $\vec{x}_\alpha \in \partial B_\alpha$ that points out of the body) and for $i = 4,5,6$

$$\begin{cases} \Delta\phi_{\alpha i}(\vec{x}_\alpha) = 0 & \text{for} \quad \vec{x}_\alpha \in \mathbb{R}^3 - B_\alpha, \\ \nabla\phi_{\alpha i}(\vec{x}_\alpha)\cdot\vec{n}_\alpha(\vec{x}_\alpha) = \vec{\mathrm{e}}_{\alpha(i-3)}\times\vec{x}_\alpha\cdot\vec{n}_\alpha(\vec{x}_\alpha) & \text{for} \quad \vec{x}_\alpha \in \partial B_\alpha, \\ ||\nabla\phi_{\alpha i}(\vec{x}_\alpha)|| \to 0 & \text{as} \quad ||\vec{x}_\alpha|| \to \infty. \end{cases} \tag{1.3}$$

Using the above definitions it is possible to show that the kinetic energy W of the system body plus fluid can be written as

$$W = \frac{1}{2}\left(m_\alpha||\vec{v}_\alpha||^2 + 2m_\alpha\vec{v}_\alpha\cdot(\vec{\omega}_\alpha\times\vec{\tau}_\alpha) + \vec{\omega}_\alpha\cdot I^\alpha\vec{\omega}_\alpha\right) + \frac{1}{2}\sum_{i,j=1}^{6} A_{ij}^\alpha c_{\alpha i}c_{\alpha j}, \tag{1.4}$$

where

m_α is the mass of body α,
$\vec{\tau}_\alpha$ is the position of the center of mass of body α in the reference frame K_α,
I^α is the moment of inertia tensor of body α with respect to the reference frame K_α, and
A_{ij}^α are the added mass coefficients of body α,

$$A_{ij}^\alpha \stackrel{\text{def}}{=} \rho\int_{\mathbb{R}^3 - B_\alpha} \nabla\phi_{\alpha i}\cdot\nabla\phi_{\alpha j}.$$

Finally, using this expression for W and D'Alambert's equation,

$$\frac{d}{dt}\frac{\partial}{\partial\dot{s}_{\alpha i}}W - \frac{\partial}{\partial s_{\alpha i}}W = Q_{\alpha i},$$

where $s_{\alpha i}$ is some generalized coordinate of the body α and $Q_{\alpha i}$ is its related generalized external force, we can write the equations of motion for the dynamics of a single body in a fluid. Notice that the effect of the fluid on the body dynamics is represented by the set of constants A_{ij}^α that depends only on the geometry of solid α. These constants either were already analytically computed for some simple interesting geometries (like ellipsoids; see [1]) or can be numerically estimated with high precision (for instance, with the commercial software "Wamit"; see http://www.wamit.com).

Therefore, in the context of ideal fluids, the problem of the dynamics of a single solid in a fluid reduces to the problem of studying some known set of ODEs. We just remark that by no means is this a trivial task. For instance, it has been proved by Kozlov [16] that even the "free motion" ($Q_{\alpha i} = 0$) of a single rigid body with sufficiently complex geometry leads to a nonintegrable set of ODEs.

Now, let us turn to the problem of motion of several solids through a liquid. As above, the idea is to write the kinetic energy of the system solid plus fluid in terms of generalized coordinates of the solids only. The same arguments as before lead to the following expression [1]:

$$W = \frac{1}{2}\sum_{\alpha=1}^{N}\left[\left(m_\alpha||\vec{v}_\alpha||^2 + 2m_\alpha\vec{v}_\alpha\cdot(\vec{\omega}_\alpha\times\vec{\tau}_\alpha) + \vec{\omega}_\alpha\cdot I^\alpha\vec{\omega}_\alpha\right)\right] + \frac{\rho}{2}\int_S ||\nabla\Phi||^2,$$

where S is the open set of points in $\mathbb{R}^3$ outside $B_1 \cup \ldots \cup B_N$ and

$$\text{(1.5)}\qquad \begin{cases} \Delta\Phi = 0 & \text{for} \quad \vec{x}\in S, \\ \nabla\Phi\cdot\vec{n} = [\vec{V}_\alpha + \vec{\Omega}_\alpha\times(\vec{x}-\vec{R}_\alpha)]\cdot\vec{n} & \text{for} \quad \vec{x}\in\partial B_\alpha, \\ ||\nabla\Phi|| \to 0 & \text{as} \quad ||\vec{x}||\to\infty. \end{cases}$$

Now, in contrast to the case of a single solid, it is not possible to write the integral

$$\text{(1.6)}\qquad \frac{\rho}{2}\int_S ||\nabla\Phi||^2$$

in terms of a set of constants that depend only on the geometry of each body. In fact this integral is a quadratic function of the velocities and angular velocities of each body and also a function of the relative positions and orientations of the bodies. Except for very simple geometries, like two spheres moving on a line of centers ([10, Chapter XI]), the form of this function is unknown. As we already mentioned the goal of this paper is to write this integral approximately and explicitly in the case where the bodies are not close to each other. This approximation will depend on certain geometrical parameters of each body α which are combinations of the added mass coefficients A^α_{ij}; the volume of body α,

$$\eta_\alpha \overset{\text{def}}{=} \int_{B_\alpha} d^3x_\alpha;$$

and the center of volume of body α,

$$\vec{\xi}_\alpha \overset{\text{def}}{=} \int_{B_\alpha}\vec{x}_\alpha d^3x_\alpha.$$

In order to motivate the definition of these geometrical parameters let us present a proposition concerning the solutions of (1.2) and (1.3).

We say that a function $f:\mathbb{R}\to\mathbb{R}$ is of order $1/|x|^n$ or

$$f(x) = \mathcal{O}\left(\frac{1}{|x|^n}\right)$$

if there exist constants C and C' such that

$$|f(x)| < \frac{C'}{|x|^n} \qquad \text{for} \quad |x| > C.$$

PROPOSITION 1.1. *Let $\phi_{\alpha i}$ be the solution of problem* (1.2) *for $i = 1, 2, 3$ and of problem* (1.3) *for $i = 4, 5, 6$. Then the following expansion holds:*

$$\phi_{\alpha i}(\vec{x}_\alpha) = -\frac{1}{4\pi}\frac{1}{||\vec{x}_\alpha||}\left\{\frac{\vec{\lambda}_{\alpha i}\cdot\vec{x}_\alpha}{||\vec{x}_\alpha||^2} + \mathcal{R}(\vec{x}_\alpha)\right\},$$

where

$$\vec{\lambda}_{\alpha i} \stackrel{\text{def}}{=} \left\{\eta_\alpha \vec{e}_{\alpha i} + \frac{1}{\rho}\sum_{j=1}^{3} A_{ij}^{\alpha}\vec{e}_{\alpha j}\right\} \in K_\alpha \qquad \textit{for} \quad i = 1, 2, 3,$$

$$\vec{\lambda}_{\alpha i} \stackrel{\text{def}}{=} \left\{\vec{e}_{\alpha(i-3)} \times \vec{\xi}_\alpha + \frac{1}{\rho}\sum_{j=1}^{3} A_{ij}^{\alpha}\vec{e}_{\alpha j}\right\} \in K_\alpha \qquad \textit{for} \quad i = 4, 5, 6,$$

$$\vec{x}_\alpha = T_\alpha^{-1}(\vec{x} - \vec{R}_\alpha).$$

Moreover, function $\mathcal{R}$ can be expanded in a convergent power series of $\vec{x}_\alpha/||\vec{x}_\alpha||^2$ and

$$|\mathcal{R}(\vec{x}_\alpha)| = \mathcal{O}\left(\frac{1}{||\vec{x}_\alpha||^2}\right).$$

The expansion in this proposition in cases $i = 1, 2, 3$ is essentially given in article 121a of Lamb's book [1]. The expansion in cases $i = 4, 5, 6$ is obtained in a similar way. The proof of the analyticity of $\phi_{\alpha i}$ at infinity can be found, for instance, in Folland's book [18]. Notice that Lamb defines the velocity of the fluid as $-\nabla\phi$ while we use $\nabla\phi$. This explains the difference in sign between the formula above and the one in Lamb's book.

Notice that the geometric parameter $\vec{\lambda}_{\alpha i}$ appearing in Proposition 1.1 is the dipole coefficient of the multipole expansion of $\phi_{\alpha i}$. The main idea in this paper is to use Proposition 1.1 in the following way. Function Φ appearing in (1.5) is unknown but it is well approximated by a sum of solutions of one-body problems $\sum_\alpha \sum_i c_{\alpha i}\phi_{\alpha i}$ if the bodies are sufficiently far apart. Since each $\phi_{\alpha i}$ behaves like a dipole at infinity, the error at the boundary conditions of (1.5) after replacing Φ by its approximation is of order $\mathcal{O}(1/R^3)$, where

$$R \stackrel{\text{def}}{=} \min\{||\vec{R}_\beta - \vec{R}_\alpha|| : \text{for} \quad \alpha \neq \beta,\ \alpha = 1, \dots, N,\ \beta = 1, \dots, N\}.$$

Then a well-known variational principle for (1.5) (see [17, exercise 8.4]) implies that the error in the kinetic energy (1.6) after replacing Φ by its approximation is of order $\mathcal{O}(1/R^6)$. Some further approximations, necessary to obtain more explicit formulas, lead to an error estimate of order $\mathcal{O}(1/R^4)$ for the kinetic energy (1.6). These are the main ideas in the proof of the following theorem.

THEOREM 1.2. *Let us define*

$$\vec{R}_{\alpha\beta} = \vec{R}_\beta - \vec{R}_\alpha \quad \textit{for} \quad \alpha \neq \beta,\ \alpha = 1, \dots, N,\ \beta = 1, \dots, N.$$

For each pair α, β, with $\alpha \neq \beta$, we define a transformation $F^{\alpha\beta} : K_\beta \to K_\alpha$ which depends on T_α, T_β, and $\vec{R}_{\alpha\beta}$, whose 3×3 matrix has the following elements:

$$\begin{aligned} F_{lk}^{\alpha\beta} &= (F^{\alpha\beta}\vec{e}_{\beta k}\cdot\vec{e}_{\alpha l}) \\ &= \frac{1}{4\pi}\frac{1}{||\vec{R}_{\alpha\beta}||^3}\left\{(T_\beta\vec{e}_{\beta k}\cdot T_\alpha\vec{e}_{\alpha l}) - 3\frac{(\vec{e}_{\beta k}\cdot T_\beta^{-1}\vec{R}_{\alpha\beta})(T_\alpha^{-1}\vec{R}_{\alpha\beta}\cdot\vec{e}_{\alpha l})}{||\vec{R}_{\alpha\beta}||^2}\right\} \\ &= F_{kl}^{\beta\alpha} = \mathcal{O}\left(\frac{1}{R^3}\right). \end{aligned}$$

Then

$$\frac{\rho}{2}\int_S ||\nabla\Phi||^2 = \frac{1}{2}\sum_{\alpha=1}^{N}\sum_{i=1}^{6}\sum_{j=1}^{6} A_{ij}^{\alpha} c_{\alpha i} c_{\alpha j} + \frac{\rho}{2}\sum_{\alpha=1}^{N}\sum_{\substack{\beta\neq\alpha\\ \beta=1}}^{N}\sum_{i=1}^{6}\sum_{j=1}^{6}(\vec{\lambda}_{\alpha i}\cdot F^{\alpha\beta}\vec{\lambda}_{\beta j})c_{\alpha i}c_{\beta j} + \mathcal{O}\left(\frac{1}{R^4}\right),$$

where

$$c_{\alpha i} = v_{\alpha i}, \qquad c_{\alpha(i+3)} = \omega_{\alpha i}, \qquad i = 1, 2, 3.$$

The result in Theorem 1.2 can be written in different ways. At first let us represent $\vec{\lambda}_{\alpha i}$ as a column vector with 3 components $\lambda_{\alpha 1 i}, \lambda_{\alpha 2 i}, \lambda_{\alpha 3 i}$ and define a 3×6 matrix λ_α, where the ith column is given by $\vec{\lambda}_{\alpha i}$. We call λ_α the "form matrix." Let us represent $c_{\alpha 1}, \ldots, c_{\alpha 6}$ as a column vector c_α with 6 elements. Notice that $\lambda_\alpha c_\alpha$ is a column vector in K_α with 3 elements. Then the kinetic energy due to the interaction of bodies α and β is given by the quadratic function $(\lambda_\alpha c_\alpha \cdot F^{\alpha\beta}\lambda_\beta c_\beta)$. Notice that matrix $F^{\alpha\beta}$ does not depend on the geometric properties of bodies α and β but just on their relative position and orientation. We call $F^{\alpha\beta}$ the "interaction matrix." It is interesting to write the interaction matrix in a different way. Let $R_{\alpha\beta 1}, R_{\alpha\beta 2}, R_{\alpha\beta 3}$ be the three components of $\vec{R}_{\alpha\beta}$ and $P^{\alpha\beta}$ be the following matrix:

$$P^{\alpha\beta} \stackrel{\text{def}}{=} \frac{1}{||\vec{R}_{\alpha\beta}||^2}\begin{pmatrix} R_{\alpha\beta1}R_{\alpha\beta1} & R_{\alpha\beta1}R_{\alpha\beta2} & R_{\alpha\beta1}R_{\alpha\beta3} \\ R_{\alpha\beta2}R_{\alpha\beta1} & R_{\alpha\beta2}R_{\alpha\beta2} & R_{\alpha\beta2}R_{\alpha\beta3} \\ R_{\alpha\beta3}R_{\alpha\beta1} & R_{\alpha\beta3}R_{\alpha\beta2} & R_{\alpha\beta3}R_{\alpha\beta3} \end{pmatrix}.$$

This matrix defines a transformation $P^{\alpha\beta}: K \to K$ that projects any vector in K on a unit vector with the direction of $\vec{R}_{\alpha\beta}$ ($P^{\alpha\beta}P^{\alpha\beta} = \mathit{identity}$ and $P^{\alpha\beta}\vec{R}_{\alpha\beta} = \vec{R}_{\alpha\beta}$). Let $\mathbb{1}$ be the identity matrix and $G^{\alpha\beta}: K \to K$ be the transformation defined as

$$G^{\alpha\beta} \stackrel{\text{def}}{=} \frac{1}{4\pi||\vec{R}_{\alpha\beta}||^3}[\mathbb{1} - 3P^{\alpha\beta}]. \tag{1.7}$$

Then the interaction matrix can be factorized as

$$F^{\alpha\beta} = T_\alpha^{-1} G^{\alpha\beta} T_\beta, \tag{1.8}$$

where $G^{\alpha\beta}$ depends only on $\vec{R}_{\alpha\beta}$ (notice that $F^{\alpha\beta} = F^{\beta\alpha\dagger}$). As we will see this factorization simplifies a lot the derivation of the equations of motion from the Lagrangian function. Finally, using the form and the interaction matrices we can rewrite the result in Theorem 1.2 as

$$\frac{\rho}{2}\int_S ||\nabla\Phi||^2 = \frac{1}{2}\sum_{\alpha=1}^{N}\sum_{i=1}^{6}\sum_{j=1}^{6} A_{ij}^{\alpha} c_{\alpha i} c_{\alpha j} + \frac{\rho}{2}\sum_{\alpha=1}^{N}\sum_{\substack{\beta\neq\alpha\\ \beta=1}}^{N}(\lambda_\alpha c_\alpha \cdot F^{\alpha\beta}\lambda_\beta c_\beta) + \mathcal{O}\left(\frac{1}{R^4}\right). \tag{1.9}$$

Now, in order to explain the origin of the interaction matrix we differentiate the expansion in Proposition 1.1 to obtain the following proposition. (The symbol ∇

before a function refers to the gradient of the function with respect to the variables in its argument.)

PROPOSITION 1.3. *Let $\phi_{\alpha i}$ be the solution of problem* (1.2) *for $i = 1, 2, 3$ and of problem* (1.3) *for $i = 4, 5, 6$. Suppose that $\vec{x}_\alpha \in \partial B_\beta$, $\beta \neq \alpha$. Then the following expansions hold:*

$$T_\beta^{-1} T_\alpha \nabla \phi_{\alpha i}(\vec{x}_\alpha) = -\vec{f}_i^{\alpha\beta} + \mathcal{O}\left(\frac{1}{R^4}\right)$$

and

$$\phi_{\alpha i}(\vec{x}_\alpha) = \phi_{\alpha i}(T_\alpha^{-1}\vec{R}_{\alpha\beta}) - \vec{f}_i^{\alpha\beta} \cdot \vec{x}_\beta + \mathcal{O}\left(\frac{1}{R^4}\right),$$

where

$$\vec{f}_i^{\alpha\beta} = F^{\beta\alpha}\vec{\lambda}_{\alpha i} = \frac{1}{4\pi}\frac{1}{||\vec{R}_{\alpha\beta}||^5}\left\{||\vec{R}_{\alpha\beta}||^2 T_\beta^{-1} T_\alpha \vec{\lambda}_{\alpha i} - 3(T_\alpha \vec{\lambda}_{\alpha i} \cdot \vec{R}_{\alpha\beta}) T_\beta^{-1}\vec{R}_{\alpha\beta}\right\} \tag{1.10}$$

and

$$\vec{x}_\beta = T_\beta^{-1}(\vec{x} - \vec{R}_\beta) \quad \Rightarrow \quad \vec{x}_\alpha = T_\alpha^{-1}(\vec{R}_{\alpha\beta} + T_\beta \vec{x}_\beta).$$

Proposition 1.3 implies that $-\vec{f}_i^{\alpha\beta} = -F^{\beta\alpha}\vec{\lambda}_{\alpha i} \in K_\beta$ essentially represents the velocity of the fluid at the reference point of body β due to the unit velocity motion of body α in the "direction i," as if the latter were moving in the absence of all other bodies. The vector $\vec{f}_i^{\alpha\beta}$ is nondimensional for $i = 1, 2, 3$, and it has dimension of length for $i = 4, 5, 6$.

Our second main result concerns the problem of motion of a single body, referred as "body α," in $\mathbb{R}^3$ in the presence of N other rigid bodies which are held at rest. If we try to apply the formula in Theorem 1.2 to this problem, we just obtain the kinetic energy for the motion of a single isolated body given in (1.4). This problem requires a better approximation of Φ than that used in the proof of Theorem 1.2. This approximation is presented in section 2. The approximated expression for the kinetic energy in this case is given in the following theorem. The notation follows the one above. In this case $\vec{R}_\beta$, T_β, $\beta = 1, \ldots, N$, do not depend on time.

THEOREM 1.4. *Let us consider a system of $N + 1$ rigid bodies in an ideal irrotational fluid in $\mathbb{R}^3$. The body denoted by the letter α is free to move, but all others with indices $\beta = 1, \ldots, N$ are held at rest. Then*

$$\frac{\rho}{2}\int_S ||\nabla\Phi||^2 = \frac{1}{2}\sum_{i=1}^{6}\sum_{j=1}^{6} c_{\alpha i} c_{\alpha j}(A_{ij}^\alpha + C_{ij}^\alpha) + \mathcal{O}\left(\frac{1}{R^7}\right),$$

where

$$\begin{aligned} C_{ij}^\alpha = C_{ji}^\alpha &= \rho \sum_{\beta=1}^{N}\sum_{k=1}^{3}(\vec{\lambda}_{\beta k} \cdot F^{\beta\alpha}\vec{\lambda}_{\alpha j})(\vec{e}_{\beta k} \cdot F^{\beta\alpha}\vec{\lambda}_{\alpha i}) \\ &= \rho \sum_{\beta=1}^{N}\sum_{k=1}^{3}(\vec{\lambda}_{\beta k} \cdot \vec{f}_j^{\alpha\beta})(\vec{e}_{\beta k} \cdot \vec{f}_i^{\alpha\beta}) \\ &= \mathcal{O}\left(\frac{1}{R^6}\right), \end{aligned}$$

where $F^{\beta\alpha}$ *is defined in Theorem* 1.2 *and* $\vec{f}_i^{\alpha\beta}$ *is defined in* (1.10).

Again the result in Theorem 1.4 can be written in a more concise way. Let $\overline{\lambda}_{\beta\ (3\times3)}$ be the form matrix $\lambda_{\beta\ (3\times6)}$ restricted to its first three first columns. Let $Q : K \to K$ be the transformation defined by

$$Q = \sum_{\beta=1}^{N} G^{\alpha\beta} T_\beta \overline{\lambda}_\beta T_\beta^{-1} G^{\alpha\beta}, \tag{1.11}$$

where $G^{\alpha\beta}$ is defined in (1.7). Notice that Q depends on the position of body α but not on its orientation. These definitions imply that the result in Theorem 1.4 can be written as

$$\frac{\rho}{2}\int_S ||\nabla\Phi||^2 = \frac{1}{2}\left\{\sum_{i=1}^{6}\sum_{j=1}^{6} c_{\alpha i} c_{\alpha j} A_{ij}^{\alpha}\right\} + \frac{\rho}{2}(\lambda_\alpha c_\alpha \cdot T_\alpha^{\dagger} Q T_\alpha \lambda_\alpha c_\alpha) + \mathcal{O}\left(\frac{1}{R^7}\right). \tag{1.12}$$

The rest of this paper is organized as follows. In section 2 we present a variational principle which provides a "good" framework for obtaining approximations for integral (1.6). We then use this variational principle to obtain first approximations for integral (1.6) which, unfortunately, do not provide expressions as explicit and simple as those given in the above theorems. In section 3 we do further approximations to the expressions obtained in section 2 and compute the formulas in Theorems 1.2 and 1.4. In section 4 we obtain the equations of motion for a system of N rigid bodies in the case in which each body has three orthogonal planes of symmetry (like, for instance, a system of ellipsoids). The results in this section indicate how to obtain the equations of motion in the general case. In section 5 we apply Theorem 1.2 to some examples: a system of N spheres, a single sphere moving in a fluid containing other fixed spheres (it could be, for instance, an infinite cubic lattice of identical equally spaced fixed spheres), and a planar system of two bodies. For this last system we also present the equations of motion explicitly. It is interesting to point out that for the problem of two spheres moving on the line of centers, Smereka compared the approximation of Theorem 1.2 to a higher order one, of order $1/R^{12}$, due to Basset (see [6, eqn. (3.26) and Figure 2] and [10, Chapter XI, especially paragraphs 232, 233]). His conclusion was that the agreement between both approximations was very good (the relative error was of order 2/1000) even when the spheres touched. Other applications of Theorems 1.2 and 1.4 can be found in [15].

Finally, it is worth mentioning some words about the important issue of estimating the errors in the first and higher order derivatives of the approximated expressions presented in the above theorems. It is not hard to prove that the function defined by integral (1.6) depends analytically on $\vec{R}_\alpha$ and T_α for $\alpha = 1, \ldots, N$. This proof was essentially presented to me by the late Prof. Daniel Henry. It is a consequence of a general theory developed by himself for the dependence of solutions of elliptic problems on the shape of the domain. (This theory is explained in a book of his which will be published by Cambridge University Press.) This implies that we can differentiate function Φ with respect to $\vec{R}_\alpha$ infinitely many times and that these derivatives satisfy elliptic problems similar to those studied in this paper. In order to estimate the derivatives of the expressions in the above theorems we have to estimate solutions of these new elliptic problems. Though interesting, this problem will not be considered here.

2. The variational principle. Let S be the open set given by $\mathbb{R}^3$ minus the union of B_α, $\alpha = 1, \dots, N$. Let Φ be the solution of problem (1.5) which we write in a simpler way as

$$\begin{cases} \Delta\Phi(\vec{x}) = 0 & \text{for} \quad \vec{x} \in S, \\ \nabla\Phi(\vec{x}) \cdot \vec{n}(\vec{x}) = g(\vec{x}) & \text{for} \quad \vec{x} \in \partial S, \\ ||\nabla\Phi(\vec{x})|| \to 0 & \text{as} \quad ||\vec{x}|| \to \infty. \end{cases}$$

We recall that $\vec{n}$ is the unit normal to ∂S *pointing "inside"* S. We will denote by $\overline{S}$ the closure of S. Using Green's identities it is easy to prove the following theorem (see [17, exercise 8.4]).

THEOREM 2.1. *Let $\theta : \overline{S} \to \mathbb{R}$ be a continuously differentiable function such that the integral of $||\nabla\theta||^2$ over S exists. Then*

$$G(\Phi) \stackrel{\text{def}}{=} \int_S ||\nabla\Phi||^2 \geq -2\int_{\partial S} g\theta - \int_S ||\nabla\theta||^2 \stackrel{\text{def}}{=} F(\theta).$$

Moreover, if $\theta = \Phi + \delta$, then

$$G(\Phi) - F(\theta) = \int_S ||\nabla\delta||^2.$$

Proof. In order to prove the theorem we write $G(\Phi) - F(\Phi + \delta)$ explicitly. Then we use Green's identity to obtain

$$\int_{\partial S} \delta\nabla\Phi \cdot \vec{n} = -\int_S (\nabla\delta \cdot \nabla\Phi),$$
$$\int_{\partial S} \Phi\nabla\Phi \cdot \vec{n} = -\int_S ||\nabla\Phi||^2.$$

With these identities we remove the integrals over the boundary from $G(\Phi) - F(\Phi + \delta)$ and get $G(\Phi) - F(\Phi + \delta) = G(\delta) \geq 0$. Therefore, $F(\theta)$ takes its minimum value when $\theta = \Phi$. □

Notice that if $\Delta\theta = 0$, then using Green's identity again we obtain

$$\left| \int_S ||\nabla\Phi||^2 - F(\theta) \right| = \left| \int_S ||\nabla\delta||^2 \right| = \left| \int_{\partial S} \delta\nabla\delta \cdot \vec{n} \right|, \tag{2.1}$$

which means that the difference $|G(\Phi) - F(\theta)|$ depends quadratically on δ over the boundary ∂S. This enables us to compute the kinetic energy of the fluid with a great precision even if we only know the fluid velocity field approximately. We remark that variational tools similar to the one described above are used to compute capacities in electrostatics (see, for instance, [17]).

In order to prove Theorems 1.2 and 1.4 we use approximations θ_1 and θ_2 that are convenient sums of solutions of one-body problems. At first we restrict attention to θ_1, which is chosen as

$$\theta_1(\vec{x}) = \sum_{\alpha=1}^{N} \sum_{i=1}^{6} c_{\alpha i}\phi_{\alpha i}(\vec{x}_\alpha), \tag{2.2}$$

where $c_{\alpha i}$ is given in (1.1). Notice that θ_1 is harmonic in S, and for each given α it can be decomposed as

$$\begin{aligned}\theta_1(\vec{x}) &= \sum_{i=1}^{6} c_{\alpha i}\phi_{\alpha i}(\vec{x}_\alpha) + \sum_{\substack{\beta\neq\alpha\\ \beta=1}}^{N}\sum_{i=1}^{6} c_{\beta i}\phi_{\beta i}(\vec{x}_\beta)\\ &\stackrel{\text{def}}{=} \sum_{i=1}^{6} c_{\alpha i}\phi_{\alpha i}(\vec{x}_\alpha) + h_\alpha(\vec{x}).\end{aligned}$$

The estimates in Proposition 1.1 imply that for $\vec{x}\in\partial B_\alpha$,

$$h_\alpha(\vec{x}) = \mathcal{O}(1/R^2), \qquad \nabla h_\alpha(\vec{x}) = \mathcal{O}(1/R^3). \tag{2.3}$$

For $\vec{x}\in\partial B_\alpha$, using that

$$\begin{aligned}&\nabla\theta_1(\vec{x}) = \sum_{i=1}^{6} c_{\alpha i}T_\alpha\nabla\phi_{\alpha i}(\vec{x}_\alpha) + \nabla h_\alpha(\vec{x}),\\ &\nabla\phi_{\alpha i}(\vec{x}_\alpha)\cdot\vec{n}_\alpha(\vec{x}_\alpha) = \vec{\mathrm{e}}_{\alpha i}\cdot\vec{n}_\alpha(\vec{x}_\alpha) \qquad \text{for} \quad i=1,2,3,\\ &\nabla\phi_{\alpha i}(\vec{x}_\alpha)\cdot\vec{n}_\alpha(\vec{x}_\alpha) = \vec{\mathrm{e}}_{\alpha(i-3)}\times\vec{x}_\alpha\cdot\vec{n}_\alpha(\vec{x}_\alpha) \qquad \text{for} \quad i=4,5,6,\\ &T_\alpha\vec{n}_\alpha(\vec{x}_\alpha) = \vec{n}(\vec{x}),\end{aligned}$$

we obtain

$$\begin{aligned}\nabla\theta_1(\vec{x})\cdot\vec{n}(\vec{x}) &= \sum_{i=1}^{6} c_{\alpha i}[T_\alpha\nabla\phi_{\alpha i}(\vec{x}_\alpha)\cdot\vec{n}(\vec{x})] + \nabla h_\alpha(\vec{x})\cdot\vec{n}(\vec{x})\\ &= \sum_{i=1}^{6} c_{\alpha i}[\nabla\phi_{\alpha i}(\vec{x}_\alpha)\cdot\vec{n}_\alpha(\vec{x}_\alpha)] + \nabla h_\alpha(\vec{x})\cdot\vec{n}(\vec{x})\\ &= \left[\sum_{i=1}^{3} v_{\alpha i}\vec{\mathrm{e}}_{\alpha i}\cdot\vec{n}_\alpha(\vec{x}_\alpha)\right] + \left[\sum_{i=1}^{3}\omega_{\alpha i}\vec{\mathrm{e}}_{\alpha i}\times\vec{x}_\alpha\cdot\vec{n}_\alpha(\vec{x}_\alpha)\right] + \nabla h_\alpha(\vec{x})\cdot\vec{n}(\vec{x})\\ &= [\vec{v}_\alpha\cdot\vec{n}_\alpha(\vec{x}_\alpha)] + [\omega_\alpha\times\vec{x}_\alpha\cdot\vec{n}_\alpha(\vec{x}_\alpha)] + \nabla h_\alpha(\vec{x})\cdot\vec{n}(\vec{x})\\ &= [\vec{V}_\alpha + \vec{\Omega}_\alpha\times(\vec{x}-\vec{R}_\alpha)]\cdot\vec{n}(\vec{x}) + \nabla h_\alpha(\vec{x})\cdot\vec{n}(\vec{x}).\end{aligned}$$

This and the boundary conditions (1.5) verified by Φ implies that $\delta_1 = \theta_1 - \Phi$ satisfies

$$\begin{cases}\Delta\delta_1(\vec{x}) = 0 & \text{for} \quad \vec{x}\in S,\\ \nabla\delta_1(\vec{x})\cdot\vec{n}(\vec{x}) = \nabla h_\alpha(\vec{x})\cdot\vec{n}(\vec{x}) & \text{for} \quad \vec{x}\in\partial B_\alpha \quad \alpha=1,\dots,N,\\ ||\nabla\delta_1(\vec{x})|| \to 0 & \text{as} \quad ||\vec{x}||\to\infty.\end{cases} \tag{2.4}$$

So, from (2.3) we get that the normal derivative of δ_1 at ∂S is of the order $\mathcal{O}(1/R^3)$.

Now, let θ_2 be given by

$$\theta_2(\vec{x}) = \sum_{i=1}^{6} c_{\alpha i}\phi_{\alpha i}(\vec{x}_\alpha) + \sum_{\beta=1}^{N}\sum_{i=1}^{6}\sum_{j=1}^{3} c_{\alpha i}f_{ij}^{\alpha\beta}\phi_{\beta j}(\vec{x}_\beta), \tag{2.5}$$

where

$$f_{ij}^{\alpha\beta} \stackrel{\text{def}}{=} (\vec{f}_i^{\alpha\beta}\cdot\vec{\mathrm{e}}_{\beta j})$$

and $\vec{f}_i^{\alpha\beta}$ is defined in (1.10). Using Proposition 1.3 and an argument analogous to the one above we obtain that $\delta_2 = \theta_2 - \Phi$ satisfies

$$\text{(2.6)} \qquad \begin{cases} \Delta\delta_2(\vec{x}) = 0 & \text{for} \quad \vec{x} \in S, \\ \nabla\delta_2(\vec{x}) \cdot \vec{n}(\vec{x}) = \mathcal{O}(1/R^4) & \text{for} \quad \vec{x} \in \partial S, \\ ||\nabla\delta_2(\vec{x})|| \to 0 & \text{as} \quad ||\vec{x}|| \to \infty. \end{cases}$$

In order to estimate the difference $|G(\Phi) - F(\theta_l)|$, $l = 1, 2$, using (2.1) it is still necessary to find an upper bound for δ_l on ∂S. This is provided by the lemma below. Its proof is given in the appendix.

LEMMA 2.2. *Suppose that δ is a solution of either problem* (2.4) *or* (2.6) *with $\nabla\delta(\vec{x}) \cdot \vec{n}(\vec{x}) = \mathcal{O}(1/R^k)$ for $\vec{x} \in \partial S$. Then*

$$\max\{|\delta(\vec{x})| : \vec{x} \in \partial S\} = \mathcal{O}(1/R^k).$$

Finally, this lemma, the estimates for $\nabla\delta \cdot \vec{n}$, and (2.1) imply the following theorem.

THEOREM 2.3. *Let θ_1 and θ_2 be the approximations of Φ given in* (2.2) *and* (2.5)*, respectively. Then*

$$\left| \int_S ||\nabla\Phi||^2 + 2\int_{\partial S} \theta_k(\nabla\Phi \cdot \vec{n}) + \int_S ||\nabla\theta_k||^2 \right| = \mathcal{O}(1/R^{4+2k})$$

for $k = 1, 2$.

3. Proofs of Theorems 1.2 and 1.4. In general the functions $\phi_{\alpha i}$ are not known. Thus, it is not easy to directly obtain from Theorem 2.3 expressions that are so explicit as those presented in Theorems 1.2 and 1.4. So, in this section we do some approximations to the integrals in Theorem 2.3 and obtain Theorems 1.2 and 1.4. At first, we will prove Theorem 1.2.

From the definition of $F(\theta)$ in Theorem 2.1 and Green's identity we get

$$\text{(3.1)} \qquad F(\theta) = -2\int_{\partial S} \theta(\nabla\Phi \cdot \vec{n}) - \int_S ||\nabla\theta||^2 = \sum_{\alpha=1}^{N} \int_{\partial B_\alpha} \theta[-2\nabla\Phi \cdot \vec{n} + \nabla\theta \cdot n].$$

Using the definition of θ_1, (2.2), and Proposition 1.3 for $\vec{x} \in \partial B_\alpha$, we get

$$\nabla\theta_1(\vec{x}) \cdot \vec{n}(\vec{x}) = \nabla\Phi(\vec{x}) \cdot \vec{n}(\vec{x}) - \sum_{\substack{\beta=1 \\ \beta\neq\alpha}}^{N} \sum_{i=1}^{6} c_{\beta i} \vec{f}_i^{\beta\alpha} \cdot \vec{n}_\alpha(\vec{x}_\alpha) + \mathcal{O}\left(\frac{1}{R^4}\right).$$

This equation, the definition of θ_1, Proposition 1.3, and

$$\text{(3.2)} \qquad \int_{\partial B_\alpha} [\nabla\phi_{\alpha i}(\vec{x}_\alpha) \cdot \vec{n}_\alpha(\vec{x}_\alpha)] = 0$$

for $i = 1, \ldots, 6$ imply that $F(\theta_1)$ given by (3.1) can be written as

$$\begin{aligned} F(\theta_1) = &- \sum_{\alpha=1}^{N} \sum_{i=1}^{6} \sum_{j=1}^{6} c_{\alpha i} c_{\alpha j} \int_{\partial B_\alpha} \phi_{\alpha j}(\vec{x}_\alpha) \nabla\phi_{\alpha i}(\vec{x}_\alpha) \cdot \vec{n}_\alpha(\vec{x}_\alpha) \\ &+ \sum_{\alpha=1}^{N} \sum_{\substack{\beta=1 \\ \beta\neq\alpha}}^{N} \sum_{i=1}^{6} \sum_{j=1}^{6} c_{\alpha j} c_{\beta i} \left\{ - \int_{\partial B_\alpha} \phi_{\alpha j}(\vec{x}_\alpha) \vec{f}_i^{\beta\alpha} \cdot \vec{n}_\alpha(\vec{x}_\alpha) \right. \\ &\left. + \int_{\partial B_\alpha} (\vec{f}_i^{\beta\alpha} \cdot \vec{x}_\alpha)(\nabla\phi_{\alpha j}(\vec{x}_\alpha) \cdot \vec{n}_\alpha(\vec{x}_\alpha)) \right\} + \mathcal{O}\left(\frac{1}{R^4}\right). \end{aligned}$$

The formula in Theorem 1.2 is a consequence of this equation, Theorem 2.3,

$$-\int_{\partial B_\alpha} \phi_{\alpha j}(\vec{x}_\alpha)\nabla\phi_{\alpha i}(\vec{x}_\alpha)\cdot\vec{n}_\alpha(\vec{x}_\alpha) = \int_{\mathbb{R}^3-B_\alpha} \nabla\phi_{\alpha i}\cdot\nabla\phi_{\alpha j} = \frac{1}{\rho}A^\alpha_{ij}, \tag{3.3}$$

which is the ij added-mass coefficient of body α, and the following proposition.

PROPOSITION 3.1.

$$-\int_{\partial B_\alpha} \phi_{\alpha j}(\vec{x}_\alpha)\vec{f}_i^{\beta\alpha}\cdot\vec{n}_\alpha(\vec{x}_\alpha) + \int_{\partial B_\alpha}(\vec{f}_i^{\beta\alpha}\cdot\vec{x}_\alpha)(\nabla\phi_{\alpha j}(\vec{x}_\alpha)\cdot\vec{n}_\alpha(\vec{x}_\alpha)) = (\vec{\lambda}_{\alpha j}\cdot\vec{f}_i^{\beta\alpha}).$$

Proof. Let us consider a large sphere Γ of radius R_Γ centered on B_α. Then using that $\phi_{\alpha j}$ and $\vec{f}_i^{\beta\alpha}\cdot\vec{x}_\alpha$ are harmonic outside B_α from Green's identity, we get

$$\begin{aligned}\int_{\partial B_\alpha} \phi_{\alpha j}(\vec{x}_\alpha)\vec{f}_i^{\beta\alpha}\cdot\vec{n}_\alpha(\vec{x}_\alpha) &+ \int_\Gamma \phi_{\alpha j}(\vec{x}_\alpha)\vec{f}_i^{\beta\alpha}\cdot\vec{n}_\alpha(\vec{x}_\alpha)\\ &= \int_{\partial B_\alpha}(\vec{f}_i^{\beta\alpha}\cdot\vec{x}_\alpha)(\nabla\phi_{\alpha j}(\vec{x}_\alpha)\cdot\vec{n}_\alpha(\vec{x}_\alpha)) + \int_\Gamma(\vec{f}_i^{\beta\alpha}\cdot\vec{x}_\alpha)(\nabla\phi_{\alpha j}(\vec{x}_\alpha)\cdot\vec{n}_\alpha(\vec{x}_\alpha)),\end{aligned}$$

which implies

$$\begin{aligned}-\int_{\partial B_\alpha} \phi_{\alpha j}(\vec{x}_\alpha)\vec{f}_i^{\beta\alpha}\cdot\vec{n}_\alpha(\vec{x}_\alpha) &+ \int_{\partial B_\alpha}(\vec{f}_i^{\beta\alpha}\cdot\vec{x}_\alpha)(\nabla\phi_{\alpha j}(\vec{x}_\alpha)\cdot\vec{n}_\alpha(\vec{x}_\alpha))\\ &= \int_\Gamma \phi_{\alpha j}(\vec{x}_\alpha)\vec{f}_i^{\beta\alpha}\cdot\vec{n}_\alpha(\vec{x}_\alpha) - \int_\Gamma(\vec{f}_i^{\beta\alpha}\cdot\vec{x}_\alpha)(\nabla\phi_{\alpha j}(\vec{x}_\alpha)\cdot\vec{n}_\alpha(\vec{x}_\alpha)).\end{aligned}$$

Using Proposition 1.1 we obtain

$$\int_\Gamma \phi_{\alpha j}(\vec{x}_\alpha)\vec{f}_i^{\beta\alpha}\cdot\vec{n}_\alpha(\vec{x}_\alpha) = -\frac{1}{4\pi}\int_\Gamma \frac{\vec{\lambda}_{\alpha j}\cdot\vec{x}_\alpha}{||\vec{x}_\alpha||^3}\vec{f}_i^{\beta\alpha}\cdot\vec{n}_\alpha(\vec{x}_\alpha) + \mathcal{O}\left(\frac{1}{R_\Gamma}\right). \tag{3.4}$$

Differentiating the approximation for $\phi_{\alpha j}$ in Proposition 1.1 we obtain

$$\nabla\phi_{\alpha j}(\vec{x}_\alpha) = -\frac{1}{4\pi||\vec{x}_\alpha||^5}\left\{||\vec{x}_\alpha||^2\vec{\lambda}_{\alpha j} - 3(\vec{\lambda}_{\alpha j}\cdot\vec{x}_\alpha)\vec{x}_\alpha\right\} + \mathcal{O}\left(\frac{1}{||\vec{x}_\alpha||^4}\right).$$

Using this approximation we obtain

$$\begin{aligned}\int_\Gamma(\vec{f}_i^{\beta\alpha}\cdot\vec{x}_\alpha)(\nabla\phi_{\alpha j}(\vec{x}_\alpha)\cdot\vec{n}_\alpha(\vec{x}_\alpha)) = &-\frac{1}{4\pi}\int_\Gamma\frac{\vec{\lambda}_{\alpha j}\cdot\vec{n}_\alpha(\vec{x}_\alpha)}{||\vec{x}_\alpha||^3}\vec{f}_i^{\beta\alpha}\cdot\vec{x}_\alpha\\ &+\frac{3}{4\pi}\int_\Gamma\frac{(\vec{f}_i^{\beta\alpha}\cdot\vec{x}_\alpha)(\vec{\lambda}_{\alpha j}\cdot\vec{x}_\alpha)(\vec{x}_\alpha\cdot\vec{n}_\alpha(\vec{x}_\alpha))}{||\vec{x}_\alpha||^5}\\ &+\mathcal{O}\left(\frac{1}{R_\Gamma}\right).\end{aligned}$$

This equation, equation (3.4), and the fact that for $\vec{x}_\alpha\in\Gamma$, $\vec{n}_\alpha(\vec{x}_\alpha) = -\vec{x}_\alpha/||\vec{x}_\alpha||$ imply that

$$\begin{aligned}\int_\Gamma \phi_{\alpha j}(\vec{x}_\alpha)\vec{f}_i^{\beta\alpha}\cdot\vec{n}_\alpha(\vec{x}_\alpha) &- \int_\Gamma(\vec{f}_i^{\beta\alpha}\cdot\vec{x}_\alpha)(\nabla\phi_{\alpha j}(\vec{x}_\alpha)\cdot\vec{n}_\alpha(\vec{x}_\alpha))\\ &= -\frac{3}{4\pi}\int_\Gamma\frac{(\vec{f}_i^{\beta\alpha}\cdot\vec{x}_\alpha)(\vec{\lambda}_{\alpha j}\cdot\vec{x}_\alpha)(\vec{x}_\alpha\cdot\vec{n}_\alpha(\vec{x}_\alpha))}{||\vec{x}_\alpha||^5} + \mathcal{O}\left(\frac{1}{R_\Gamma}\right).\end{aligned} \tag{3.5}$$

Using that $\vec{n}_\alpha(\vec{x}_\alpha) = -\vec{x}_\alpha/||\vec{x}_\alpha||$ we obtain

$$\frac{3}{4\pi}\int_\Gamma \frac{(\vec{f}_i^{\beta\alpha}\cdot\vec{x}_\alpha)(\vec{\lambda}_{\alpha j}\cdot\vec{x}_\alpha)(\vec{x}_\alpha\cdot\vec{n}_\alpha(\vec{x}_\alpha))}{||\vec{x}_\alpha||^5} = \frac{3}{4\pi R_\Gamma^3}\int_\Gamma (\vec{f}_i^{\beta\alpha}\cdot\vec{x}_\alpha)(\vec{\lambda}_{\alpha j}\cdot\vec{n}_\alpha(\vec{x}_\alpha)).$$

Using that $(\vec{f}_i^{\beta\alpha}\cdot\vec{x}_\alpha)$ and $(\vec{\lambda}_{\alpha j}\cdot\vec{x}_\alpha)$ are harmonic functions in $\mathbb{R}^3$ from Green's identity, we get

$$\frac{3}{4\pi R_\Gamma^3}\int_\Gamma (\vec{f}_i^{\beta\alpha}\cdot\vec{x}_\alpha)(\vec{\lambda}_{\alpha j}\cdot\vec{n}_\alpha(\vec{x}_\alpha)) = -\frac{3}{4\pi R_\Gamma^3}\int_{Ball\ R_\Gamma} (\vec{f}_i^{\beta\alpha}\cdot\vec{\lambda}_{\alpha j}) = -(\vec{f}_i^{\beta\alpha}\cdot\vec{\lambda}_{\alpha j}).$$

Thus, taking the limit as $R_\Gamma \to \infty$ in (3.5), we prove the proposition. □

Now, we turn to the proof of Theorem 1.4. Using (3.1), the definition of θ_2, (2.5), Propositions 1.1 and 1.3, and (3.2) we obtain

$$\begin{aligned} F(\theta_2) = &-\sum_{i=1}^{6}\sum_{j=1}^{6} c_{\alpha i}c_{\alpha j}\int_{\partial B_\alpha}\phi_{\alpha j}(\vec{x}_\alpha)\nabla\phi_{\alpha i}(\vec{x}_\alpha)\cdot\vec{n}_\alpha(\vec{x}_\alpha) \\ &+\sum_{\beta=1}^{N}\sum_{i=1}^{6}\sum_{j=1}^{6}\sum_{k=1}^{3} c_{\alpha j}c_{\beta i}f_{ik}^{\alpha\beta}\left\{-\int_{\partial B_\alpha}\phi_{\alpha j}(\vec{x}_\alpha)\vec{f}_k^{\beta\alpha}\cdot\vec{n}_\alpha(\vec{x}_\alpha)\right. \\ &\left.+\int_{\partial B_\alpha}(\vec{f}_k^{\beta\alpha}\cdot\vec{x}_\alpha)(\nabla\phi_{\alpha j}(\vec{x}_\alpha)\cdot\vec{n}_\alpha(\vec{x}_\alpha))\right\}+\mathcal{O}\left(\frac{1}{R^7}\right). \end{aligned}$$

This equation, equation (3.3), Proposition 3.1, and Theorem 2.3 imply the formula in Theorem 1.4. The statement $C_{ij}^\alpha = C_{ji}^\alpha$ in Theorem 1.4 is a consequence of (1.12) and the symmetry of matrix Q given in (1.11). (Notice that both $G^{\alpha\beta}$ and $\overline{\lambda}_\beta$ are symmetric.)

4. The equations of motion. Let us consider a system of N rigid bodies such that each body has three orthogonal planes of symmetry. This implies that we can choose each reference frame K_α such that the added mass matrix A_{ij}^α is diagonal. We will assume this choice. In this case all coefficients of the form matrix $\lambda_{\alpha ij}$ are null for $i = 1, 2, 3$ and $j = 4, 5, 6$. So we can restrict the form matrix to its first three columns. To simplify the notation we keep denoting this restricted matrix by λ_α. Notice that λ_α is diagonal: $\lambda_{\alpha ii} = \eta_\alpha + \rho^{-1}A_{ii}^\alpha$, $i = 1, 2, 3$, $\alpha = 1, \dots, N$. Let us write the kinetic energy of the system as

$$W = \sum_{\alpha=1}^{N} W_\alpha + \frac{1}{2}\sum_{\alpha=1}^{N}\sum_{\substack{\beta=1\\ \beta\neq\alpha}}^{N} W_{\alpha\beta}, \tag{4.1}$$

where W_α is the kinetic energy of body α as if it were isolated and $W_{\alpha\beta}$ is the approximation for the kinetic energy of hydrodynamic interaction given in Theorem 1.2,

$$W_{\alpha\beta} = \rho(\lambda_\alpha\vec{v}_\alpha\cdot F^{\alpha\beta}\lambda_\beta\vec{v}_\beta) = W_{\beta\alpha}.$$

The kinetic energy W_α (equation (1.4)) can be written as

$$W_\alpha = \frac{1}{2}\left(\vec{v}_\alpha\cdot b^\alpha\vec{v}_\alpha + 2m_\alpha\vec{v}_\alpha\cdot(\vec{\omega}_\alpha\times\vec{\tau}_\alpha) + \vec{\omega}_\alpha\cdot\xi^\alpha\vec{\omega}_\alpha\right),$$

where b^α and ξ^α are matrices given by

$$
\begin{aligned}
&b^\alpha_{ij} = m_\alpha \delta_{ij} + A^\alpha_{ij}\delta_{ij}, \\
&\xi^\alpha_{ij} = I^\alpha_{ij} + A^\alpha_{(i+3)(j+3)}\delta_{ij} \qquad \text{for} \quad i = 1,2,3,\ j = 1,2,3,
\end{aligned} \tag{4.2}
$$

where $\delta_{ij} = 1$ if $i = j$, otherwise it is zero. In order to find the equations of motion we have to compute

$$
\frac{d}{dt}\nabla_{\vec{V}_\alpha} W - \nabla_{\vec{R}_\alpha} W,
$$
$$
\frac{d}{dt}\nabla_{\dot{\vec{\psi}}_\alpha} W - \nabla_{\psi_\alpha} W
$$

for $\alpha = 1, \dots, N$, where $V_\alpha = \dot{\vec{R}}_\alpha$ and ψ_α is any set of three angles $\psi_{\alpha 1}, \psi_{\alpha 2}, \psi_{\alpha 3}$ that parameterize the attitude matrix T_α. (It can be the Euler angles, for instance.) Due to the decomposition (4.1) of W, the equations of motion are sums of terms of the following types:

$$
\frac{d}{dt}\nabla_{\vec{V}_\alpha} W_\alpha - \nabla_{\vec{R}_\alpha} W_\alpha, \tag{4.3}
$$

$$
\frac{d}{dt}\nabla_{\dot{\vec{\psi}}_\alpha} W_\alpha - \nabla_{\psi_\alpha} W_\alpha, \tag{4.4}
$$

$$
\frac{d}{dt}\nabla_{\vec{V}_\alpha} W_{\alpha\beta} - \nabla_{\vec{R}_\alpha} W_{\alpha\beta}, \tag{4.5}
$$

$$
\frac{d}{dt}\nabla_{\dot{\vec{\psi}}_\alpha} W_{\alpha\beta} - \nabla_{\psi_\alpha} W_{\alpha\beta}, \tag{4.6}
$$

where $\beta \neq \alpha$. There are two main difficulties in computing these derivatives. The first difficulty is the amount of terms that have to be differentiated and that appear in the final equation of motion. In order to partially overcome this problem we define some tensors that allow us to simplify the computations and the final result. The second difficulty is related to the angle parameterization of the attitude matrices. We overcome this problem by using some identities for derivatives of orthogonal matrices that are independent of the particular choice of angle parameterization. We remark that the expressions resulting from (4.3) and (4.4) are well known. They correspond to the equations of motion for an isolated body (Kirchhoff's equations). Here we indicate how to obtain them from the Lagrangian W_α to show how to handle derivatives with respect to angular velocities. Notice that angular velocities do not appear in the particular $W_{\alpha\beta}$ considered in this section, but they do appear in $W_{\alpha\beta}$ for systems of bodies of more complex geometry.

Let us start computing (4.5). Using (1.8), that $\vec{V}_\alpha = T_\alpha \vec{v}_\alpha$, and the identification $T_\alpha^{-1} = T_\alpha^\dagger$, we write

$$
W_{\alpha\beta} = \rho(\vec{V}_\alpha \cdot T_\alpha \lambda_\alpha T_\alpha^\dagger G^{\alpha\beta} T_\beta \lambda_\beta T_\beta^\dagger \vec{V}_\beta). \tag{4.7}
$$

This implies that

$$
\frac{1}{\rho}\nabla_{\vec{V}_\alpha} W_{\alpha\beta} = T_\alpha \lambda_\alpha T_\alpha^\dagger G^{\alpha\beta} T_\beta \lambda_\beta T_\beta^\dagger \vec{V}_\beta
$$

and that

$$\frac{1}{\rho}\frac{d}{dt}\nabla_{\vec{V}_\alpha} W_{\alpha\beta} = \dot{T}_\alpha\lambda_\alpha T_\alpha^\dagger G^{\alpha\beta} T_\beta\lambda_\beta T_\beta^\dagger \vec{V}_\beta + T_\alpha\lambda_\alpha \dot{T}_\alpha^\dagger G^{\alpha\beta} T_\beta\lambda_\beta T_\beta^\dagger \vec{V}_\beta \tag{4.8}$$

$$+ T_\alpha\lambda_\alpha T_\alpha^\dagger G^{\alpha\beta} \dot{T}_\beta\lambda_\beta T_\beta^\dagger \vec{V}_\beta + T_\alpha\lambda_\alpha T_\alpha^\dagger G^{\alpha\beta} T_\beta\lambda_\beta \dot{T}_\beta^\dagger \vec{V}_\beta \tag{4.9}$$

$$+ T_\alpha\lambda_\alpha T_\alpha^\dagger \dot{G}^{\alpha\beta} T_\beta\lambda_\beta T_\beta^\dagger \vec{V}_\beta \tag{4.10}$$

$$+ T_\alpha\lambda_\alpha T_\alpha^\dagger G^{\alpha\beta} T_\beta\lambda_\beta T_\beta^\dagger \dot{\vec{V}}_\beta. \tag{4.11}$$

The time derivatives of the attitude matrices are related to the angular velocities in the following way. Matrix T_α satisfies $T_\alpha T_\alpha^\dagger = \mathbb{1}$. Differentiating this relation we get $\dot{T}_\alpha T_\alpha^\dagger = -T_\alpha \dot{T}_\alpha^\dagger$, which implies that $\dot{T}_\alpha T_\alpha^\dagger$ is antisymmetric. The angular velocities $\vec{\Omega}_\alpha$, $\vec{\omega}_\alpha$ are vectors whose components satisfy

$$\dot{T}_\alpha T_\alpha^\dagger \stackrel{\text{def}}{=} \hat{\Omega}_\alpha = \begin{pmatrix} 0 & -\Omega_{\alpha 3} & \Omega_{\alpha 2} \\ \Omega_{\alpha 3} & 0 & -\Omega_{\alpha 1} \\ -\Omega_{\alpha 2} & \Omega_{\alpha 1} & 0 \end{pmatrix} \tag{4.12}$$

and

$$T_\alpha^\dagger(\dot{T}_\alpha T_\alpha^\dagger)T_\alpha = T_\alpha^\dagger \dot{T}_\alpha \stackrel{\text{def}}{=} \hat{\omega}_\alpha = \begin{pmatrix} 0 & -\omega_{\alpha 3} & \omega_{\alpha 2} \\ \omega_{\alpha 3} & 0 & -\omega_{\alpha 1} \\ -\omega_{\alpha 2} & \omega_{\alpha 1} & 0 \end{pmatrix}. \tag{4.13}$$

Notice that $\hat{\Omega}_\alpha : K \to K$ and $\hat{\omega}_\alpha : K_\alpha \to K_\alpha$ are antisymmetric transformations. For any vector $\vec{U} \in K$ or $\vec{u} \in K_\alpha$ the following identities hold:

$$\dot{T}_\alpha T_\alpha^\dagger \vec{U} = \hat{\Omega}_\alpha \vec{U} = \vec{\Omega}_\alpha \times \vec{U}, \qquad \dot{T}_\alpha^\dagger \dot{T}_\alpha \vec{u} = \hat{\omega}_\alpha \vec{u} = \vec{\omega}_\alpha \times \vec{u}.$$

Sometimes we write $\dot{T}_\alpha^\dagger \dot{T}_\alpha \vec{u} = \hat{\omega}_\alpha \vec{u}$, sometimes $\dot{T}_\alpha^\dagger \dot{T}_\alpha \vec{u} = \vec{\omega}_\alpha \times \vec{u}$. For each $\alpha = 1, \ldots, N$, we write the form matrix in the inertial reference frame K as

$$T_\alpha \lambda_\alpha T_\alpha^\dagger = \Lambda_\alpha. \tag{4.14}$$

With this notation we get that line (4.8) becomes

$$\dot{T}_\alpha\lambda_\alpha T_\alpha^\dagger G^{\alpha\beta} T_\beta\lambda_\beta T_\beta^\dagger \vec{V}_\beta + T_\alpha\lambda_\alpha \dot{T}_\alpha^\dagger G^{\alpha\beta} T_\beta\lambda_\beta T_\beta^\dagger \vec{V}_\beta = [\hat{\Omega}_\alpha\Lambda_\alpha - \Lambda_\alpha\hat{\Omega}_\alpha] G^{\alpha\beta}\Lambda_\beta \vec{V}_\beta \tag{4.15}$$

and line (4.9) becomes

$$T_\alpha\lambda_\alpha T_\alpha^\dagger G^{\alpha\beta} \dot{T}_\beta\lambda_\beta T_\beta^\dagger \vec{V}_\beta + T_\alpha\lambda_\alpha T_\alpha^\dagger G^{\alpha\beta} T_\beta\lambda_\beta \dot{T}_\beta^\dagger \vec{V}_\beta = \Lambda_\alpha G^{\alpha\beta}[\hat{\Omega}_\beta\Lambda_\beta - \Lambda_\beta\hat{\Omega}_\beta]\vec{V}_\beta. \tag{4.16}$$

In order to handle the derivative $\dot{G}^{\alpha\beta}$ we define a tensor $H^{\alpha\beta}$ that with each vector $\vec{U} \in K$ associates a linear transformation $H^{\alpha\beta}(\vec{U})$ in K. The components of $H^{\alpha\beta}$ are

$$H_{ijk}^{\alpha\beta} = \frac{\partial G_{ij}^{\alpha\beta}}{\partial R_{\alpha\beta k}} = -\frac{3}{4\pi||\vec{R}_{\alpha\beta}||^5}\left\{ R_{\alpha\beta k}\delta_{ij} + R_{\alpha\beta i}\delta_{kj} + R_{\alpha\beta j}\delta_{ik} - 5\frac{R_{\alpha\beta i}R_{\alpha\beta j}R_{\alpha\beta k}}{||\vec{R}_{\alpha\beta}||^2}\right\}, \tag{4.17}$$

where i, j, k take values in $1, 2, 3$ and $\delta_{ij} = 1$ if $i = j$, otherwise it is zero. Notice that $H^{\alpha\beta}$ is a totally symmetric tensor. If $\vec{U}$ and $\vec{W}$ are arbitrary vectors in K, then

$$H^{\alpha\beta}(\vec{U})\vec{W} = -\frac{3}{4\pi||\vec{R}_{\alpha\beta}||^5}\Bigg\{\vec{R}_{\alpha\beta}(\vec{U}\cdot\vec{V}) + \vec{U}(\vec{R}_{\alpha\beta}\cdot\vec{V}) + \vec{V}(\vec{R}_{\alpha\beta}\cdot\vec{U}) - 5\frac{\vec{R}_{\alpha\beta}(\vec{R}_{\alpha\beta}\cdot\vec{U})(\vec{R}_{\alpha\beta}\cdot\vec{V})}{||\vec{R}_{\alpha\beta}||^2}\Bigg\}. \tag{4.18}$$

With this definition, after working with each coordinate separately, we get

$$\begin{aligned}\Lambda_\alpha\dot{G}^{\alpha\beta}\Lambda_\beta V_\beta - \nabla_{\vec{R}_{\alpha\beta}}W_{\alpha\beta} \\ = \Lambda_\alpha H^{\alpha\beta}(\Lambda_\beta\vec{V}_\beta)\vec{V}_\beta + [H^{\alpha\beta}(\Lambda_\beta\vec{V}_\beta)\Lambda_\alpha - \Lambda_\alpha H^{\alpha\beta}(\Lambda_\beta\vec{V}_\beta)]\vec{V}_\alpha.\end{aligned} \tag{4.19}$$

Finally, from (4.19), (4.16), (4.15), (4.11), (4.10), (4.9), and (4.8) we get

$$\begin{aligned}&\frac{d}{dt}\nabla_{\vec{V}_\alpha}W_{\alpha\beta} - \nabla_{\vec{R}_\alpha}W_{\alpha\beta} \\ &= \rho[\hat{\Omega}_\alpha\Lambda_\alpha - \Lambda_\alpha\hat{\Omega}_\alpha]G^{\alpha\beta}\Lambda_\beta\vec{V}_\beta + \rho\Lambda_\alpha G^{\alpha\beta}[\hat{\Omega}_\beta\Lambda_\beta - \Lambda_\beta\hat{\Omega}_\beta]\vec{V}_\beta + \rho\Lambda_\alpha G^{\alpha\beta}\Lambda_\beta\dot{\vec{V}}_\beta \\ &+ \rho\Lambda_\alpha H^{\alpha\beta}(\Lambda_\beta\vec{V}_\beta)\vec{V}_\beta + \rho[H^{\alpha\beta}(\Lambda_\beta\vec{V}_\beta)\Lambda_\alpha - \Lambda_\alpha H^{\alpha\beta}(\Lambda_\beta\vec{V}_\beta)]\vec{V}_\alpha.\end{aligned} \tag{4.20}$$

We remark that the force terms in line (4.20) are of order $\mathcal{O}(1/R^4)$ while the terms in the previous line are of order $\mathcal{O}(1/R^3)$.

Let us compute (4.6). Now, we have to differentiate T_α with respect to the angles $\psi_{\alpha i}$, $i = 1, 2, 3$. In the same way we defined the angular velocities Ω_α, (4.12), and ω_α, (4.13), we define vectors $\vec{\Gamma}_{\alpha i}$ and $\vec{\gamma}_{\alpha i}$, $i = 1, 2, 3$, that have components $\Gamma_{\alpha i1}, \Gamma_{\alpha i2}, \Gamma_{\alpha i3}$ and $\gamma_{\alpha i1}, \gamma_{\alpha i2}, \gamma_{\alpha i3}$, respectively, given by

$$\left(\frac{\partial T_\alpha}{\partial\psi_{\alpha i}}\right)T_\alpha^\dagger = \begin{pmatrix} 0 & -\Gamma_{\alpha i3} & \Gamma_{\alpha i2} \\ \Gamma_{\alpha i3} & 0 & -\Gamma_{\alpha i1} \\ -\Gamma_{\alpha i2} & \Gamma_{\alpha i1} & 0 \end{pmatrix} \tag{4.21}$$

and

$$T_\alpha^\dagger\frac{\partial T_\alpha}{\partial\psi_{\alpha i}} = \begin{pmatrix} 0 & -\gamma_{\alpha i3} & \gamma_{\alpha i2} \\ \gamma_{\alpha i3} & 0 & -\gamma_{\alpha i1} \\ -\gamma_{\alpha i2} & \gamma_{\alpha i1} & 0 \end{pmatrix}. \tag{4.22}$$

This definition implies that for any vectors $\vec{U} \in K$, $\vec{u} \in K_\alpha$ the following identities hold:

$$\left(\frac{\partial T_\alpha}{\partial\psi_{\alpha i}}\right)T_\alpha^\dagger\vec{U} = \vec{\Gamma}_{\alpha i}\times\vec{U}, \qquad T_\alpha^\dagger\frac{\partial T_\alpha}{\partial\psi_{\alpha i}}\vec{u} = \vec{\gamma}_{\alpha i}\times\vec{u}.$$

Using that $W_{\alpha\beta}$ does not depend on $\dot{\psi}_{\alpha i}$, the above definition of $\vec{\Gamma}_{\alpha i}$, and some easy vector identities, we get

$$\begin{aligned}\frac{d}{dt}\frac{\partial}{\partial\dot{\psi}_{\alpha i}}W_{\alpha\beta} - \frac{\partial}{\partial\psi_{\alpha i}}W_{\alpha\beta} \\ = \rho(\vec{\Gamma}_{\alpha i}\cdot\{[\vec{V}_\alpha\times\Lambda_\alpha G^{\alpha\beta}\Lambda_\beta\vec{V}_\beta] - [(\Lambda_\alpha\vec{V}_\alpha)\times G^{\alpha\beta}\Lambda_\beta\vec{V}_\beta]\}).\end{aligned}$$

Using the linear independence of $\vec{\Gamma}_{\alpha i}$, $i = 1, 2, 3$, we get

$$\begin{aligned} &\frac{d}{dt}\nabla_{\dot{\psi}_\alpha} W_{\alpha\beta} - \nabla_{\psi_\alpha} W_{\alpha\beta} \\ &\qquad = \rho[\vec{V}_\alpha \times \Lambda_\alpha G^{\alpha\beta}\Lambda_\beta \vec{V}_\beta] - \rho[(\Lambda_\alpha \vec{V}_\alpha) \times G^{\alpha\beta}\Lambda_\beta \vec{V}_\beta]. \end{aligned} \tag{4.23}$$

We remark that the torque on the right-hand side of this equation is of order $\mathcal{O}(1/R^3)$.

The computation of (4.3) uses only arguments similar to those given above. So in this case we just present the final result. Let B^α and Ξ^α be the inertia and moment of inertia matrices given in the inertial reference frame, namely

$$B^\alpha = T_\alpha b^\alpha T_\alpha^\dagger, \qquad \Xi^\alpha = T_\alpha \xi^\alpha T_\alpha^\dagger, \tag{4.24}$$

where b^α and ξ^α were defined in (4.2). Then

$$\begin{aligned} &\frac{d}{dt}\nabla_{\vec{V}_\alpha} W_\alpha - \nabla_{\vec{R}_\alpha} W_\alpha \\ &\qquad = B\dot{\vec{V}}_\alpha + m_\alpha \dot{\vec{\Omega}}_\alpha \times (T_\alpha \vec{\tau}_\alpha) + m_\alpha \vec{\Omega}_\alpha \times [\vec{\Omega}_\alpha \times (T_\alpha \vec{\tau}_\alpha)] \\ &\qquad\quad + [\vec{\Omega}_\alpha \times B\vec{V}_\alpha] - B[\vec{\Omega}_\alpha \times \vec{V}_\alpha]. \end{aligned} \tag{4.25}$$

The computation of (4.4) involves derivatives of W_α with respect to $\dot{\psi}_{\alpha i}$, which did not appear yet. In this case, in order to get a final expression that does not depend on a particular parameterization of T_α we have to use the following identities:

$$\vec{\Gamma}_{\alpha i} \times \vec{\Gamma}_{\alpha j} = \frac{\partial \vec{\Gamma}_{\alpha j}}{\partial \psi_{\alpha i}} - \frac{\partial \vec{\Gamma}_{\alpha i}}{\partial \psi_{\alpha j}}, \tag{4.26}$$

$$\vec{\gamma}_{\alpha i} \times \vec{\gamma}_{\alpha j} = \frac{\partial \vec{\gamma}_{\alpha i}}{\partial \psi_{\alpha j}} - \frac{\partial \vec{\gamma}_{\alpha j}}{\partial \psi_{\alpha i}}, \tag{4.27}$$

and

$$\begin{aligned} \vec{\Omega}_\alpha &= \dot{\psi}_{\alpha 1}\vec{\Gamma}_{\alpha 1} + \dot{\psi}_{\alpha 2}\vec{\Gamma}_{\alpha 2} + \dot{\psi}_{\alpha 3}\vec{\Gamma}_{\alpha 3}, \\ \vec{\omega}_\alpha &= \dot{\psi}_{\alpha 1}\vec{\gamma}_{\alpha 1} + \dot{\psi}_{\alpha 2}\vec{\gamma}_{\alpha 2} + \dot{\psi}_{\alpha 3}\vec{\gamma}_{\alpha 3}. \end{aligned} \tag{4.28}$$

These identities are used in the following way. Let us consider the part of (4.4) given by

$$\frac{d}{dt}\frac{\partial}{\partial \dot{\psi}_{\alpha i}}\frac{\vec{\omega}_\alpha \cdot \xi^\alpha \vec{\omega}_\alpha}{2} - \frac{\partial}{\partial \psi_{\alpha i}}\frac{\vec{\omega}_\alpha \cdot \xi^\alpha \vec{\omega}_\alpha}{2}.$$

Using (4.28) and that ξ^α does not depend on time, we get

$$\frac{d}{dt}\frac{\partial}{\partial \dot{\psi}_{\alpha i}}\frac{\vec{\omega}_\alpha \cdot \xi^\alpha \vec{\omega}_\alpha}{2} = \vec{\gamma}_{\alpha i} \cdot \xi^\alpha \dot{\vec{\omega}}_\alpha + (\xi^\alpha \vec{\omega}_\alpha) \cdot \sum_{j=1}^{3} \dot{\psi}_{\alpha j}\frac{\partial \vec{\gamma}_{\alpha i}}{\partial \psi_{\alpha j}}.$$

Then using (4.27) we get

$$\begin{aligned}
&\frac{d}{dt}\frac{\partial}{\partial\dot{\psi}_{\alpha i}}\frac{\vec{\omega}_\alpha\cdot\xi^\alpha\vec{\omega}_\alpha}{2}-\frac{\partial}{\partial\psi_{\alpha i}}\frac{\vec{\omega}_\alpha\cdot\xi^\alpha\vec{\omega}_\alpha}{2}\\
&\qquad=\vec{\gamma}_{\alpha i}\cdot\xi^\alpha\dot{\vec{\omega}}_\alpha+(\xi^\alpha\vec{\omega}_\alpha)\cdot\sum_{j=1}^{3}\dot{\psi}_{\alpha j}\left[\frac{\partial\vec{\gamma}_{\alpha i}}{\partial\psi_{\alpha j}}-\frac{\partial\vec{\gamma}_{\alpha j}}{\partial\psi_{\alpha i}}\right]\\
&\qquad=\vec{\gamma}_{\alpha i}\cdot\xi^\alpha\dot{\vec{\omega}}_\alpha+(\xi^\alpha\vec{\omega}_\alpha)\cdot\sum_{j=1}^{3}\dot{\psi}_{\alpha j}[\vec{\gamma}_{\alpha i}\times\vec{\gamma}_{\alpha j}]\\
&\qquad=\vec{\gamma}_{\alpha i}\cdot[\xi^\alpha\dot{\vec{\omega}}_\alpha+(\vec{\omega}_\alpha\times\xi^\alpha\vec{\omega}_\alpha)]\\
&\qquad=\vec{\Gamma}_{\alpha i}\cdot[\Xi^\alpha\dot{\vec{\Omega}}_\alpha+(\vec{\Omega}_\alpha\times\Xi^\alpha\vec{\Omega}_\alpha)],
\end{aligned}$$

where in the last line we used that $\vec{\Gamma}_{\alpha i}=T_\alpha\vec{\gamma}_{\alpha i}$ and $\dot{\vec{\Omega}}_\alpha=T_\alpha\dot{\vec{\omega}}_\alpha$. Using these ideas and some well-known vector identities, we obtain

$$\begin{aligned}
&\frac{d}{dt}\nabla_{\dot{\vec{\psi}}_\alpha}W_\alpha-\nabla_{\psi_\alpha}W_\alpha\\
&\qquad=\Xi^\alpha\dot{\vec{\Omega}}_\alpha+m_\alpha(T_\alpha\vec{\tau}_\alpha)\times\dot{\vec{V}}_\alpha+\vec{\Omega}_\alpha\times(\Xi^\alpha\vec{\Omega}_\alpha)+\vec{V}_\alpha\times(B\vec{V}_\alpha).
\end{aligned}\tag{4.29}$$

Equations (4.20), (4.23), (4.25), and (4.29) are given in the inertial reference frame K. It is also possible to write the full system of equations choosing for each equation containing W_α the reference frame K_α. This is the approach taken in the trivial case of a system with only one body, since the attitude matrix disappears from the equations of motion. This advantage is obviously lost when there is more then one body. Nevertheless, it is still interesting to write the equations in the reference frame of each body especially when the bodies are far a way. In this case the attitude matrices appear only in the weak coupling term that corresponds to $W_{\alpha\beta}$. Below, we write equations (4.20), (4.23), (4.25), and (4.29) in the reference frame K_α:

$$\begin{aligned}
&T_\alpha^\dagger\left\{\frac{d}{dt}\nabla_{\vec{V}_\alpha}W_{\alpha\beta}-\nabla_{\vec{R}_\alpha}W_{\alpha\beta}\right\}\\
&\qquad=\rho[\hat{\omega}_\alpha\lambda_\alpha-\lambda_\alpha\hat{\omega}_\alpha]F^{\alpha\beta}\lambda_\beta\vec{v}_\beta+\rho\lambda_\alpha F^{\alpha\beta}\hat{\omega}_\beta\lambda_\beta\vec{v}_\beta\\
&\qquad\quad+\rho\lambda_\alpha F^{\alpha\beta}\lambda_\beta\dot{\vec{v}}_\beta+\rho\lambda_\alpha T_\alpha^\dagger H^{\alpha\beta}(T_\beta\lambda_\beta\vec{v}_\beta)T_\beta\vec{v}_\beta\\
&\qquad\quad+\rho[T_\alpha^\dagger H^{\alpha\beta}(T_\beta\lambda_\beta\vec{v}_\beta)T_\alpha\lambda_\alpha-\lambda_\alpha T_\alpha^\dagger H^{\alpha\beta}(T_\beta\lambda_\beta\vec{v}_\beta)T_\alpha]\vec{v}_\alpha,
\end{aligned}\tag{4.30}$$

$$\begin{aligned}
&T_\alpha^\dagger\left\{\frac{d}{dt}\nabla_{\dot{\psi}_\alpha}W_{\alpha\beta}-\nabla_{\psi_\alpha}W_{\alpha\beta}\right\}\\
&\qquad=\rho[\vec{v}_\alpha\times\lambda_\alpha F^{\alpha\beta}\lambda_\beta\vec{v}_\beta]-\rho[(\lambda_\alpha\vec{v}_\alpha)\times F^{\alpha\beta}\lambda_\beta\vec{v}_\beta],
\end{aligned}\tag{4.31}$$

$$\begin{aligned}
&T_\alpha^\dagger\left\{\frac{d}{dt}\nabla_{\vec{V}_\alpha}W_\alpha-\nabla_{\vec{R}_\alpha}W_\alpha\right\}\\
&\qquad=b^\alpha\dot{\vec{v}}_\alpha+m_\alpha\dot{\vec{\omega}}_\alpha\times\vec{\tau}_\alpha+m_\alpha\vec{\omega}_\alpha\times[\vec{\omega}_\alpha\times\tau_\alpha]+[\vec{\omega}_\alpha\times b^\alpha\vec{v}_\alpha],
\end{aligned}\tag{4.32}$$

$$\begin{aligned}
&T_\alpha^\dagger\left\{\frac{d}{dt}\nabla_{\dot{\vec{\psi}}_\alpha}W_\alpha-\nabla_{\psi_\alpha}W_\alpha\right\}\\
&\qquad=\xi^\alpha\dot{\vec{\omega}}_\alpha+m_\alpha\vec{\tau}_\alpha\times\dot{\vec{v}}_\alpha+m_\alpha\vec{\tau}_\alpha\times(\omega_\alpha\times\vec{v}_\alpha)\\
&\qquad\quad+\vec{\omega}_\alpha\times(\xi^\alpha\vec{\omega}_\alpha)+\vec{v}_\alpha\times(b^\alpha\vec{v}_\alpha).
\end{aligned}\tag{4.33}$$

5. Examples.

5.1. Many balls. The added-mass coefficients for a single ball of radius a_α with respect to its center are $A^\alpha_{ii} = \frac{2}{3}\pi\rho a^3_\alpha$ for $i = 1, 2, 3$, and all the remaining coefficients are zero. The volume of the ball is $\eta_\alpha = \frac{4}{3}\pi a^3_\alpha$, and its center of volume is $\vec{\xi}_\alpha = \vec{0}$. Therefore, $\vec{\lambda}_{\alpha i} = 2\pi a^3_\alpha \vec{e}_{\alpha i}$, $i = 1, 2, 3$, $\vec{\lambda}_{\alpha i} = \vec{0}$, $i = 4, 5, 6$, and $\lambda_\alpha c_\alpha = 2\pi a^3_\alpha \vec{v}_\alpha$. Let us consider a system of N balls of radii a_α, $\alpha = 1, \dots, N$. Then from Theorem 1.2, equations (1.7) and (1.8) the kinetic energy of the system is approximately given by

$$\frac{1}{2}\sum_{\alpha=1}^{N}\left(m_\alpha + \frac{2}{3}\pi a^3_\alpha \rho\right)||\vec{V}_\alpha||^2 + \frac{\rho}{2}\sum_{\alpha=1}^{N}\sum_{\substack{\beta\neq\alpha\\ \beta=1}}^{N}\frac{\pi a^3_\alpha a^3_\beta}{||\vec{R}_{\alpha\beta}||^3}\left\{(\vec{V}_\alpha\cdot\vec{V}_\beta) - 3\frac{(\vec{V}_\alpha\cdot\vec{R}_{\alpha\beta})(\vec{V}_\beta\cdot\vec{R}_{\alpha\beta})}{||\vec{R}_{\alpha\beta}||^2}\right\},$$

where m_α is the mass of ball α. The formulas in this paragraph had been previously obtained in [4] and [6].

5.2. A ball moving in an environment containing many fixed balls. Let us assume that a ball of radius a_α is moving in a fluid of density ρ in $\mathbb{R}^3$ in the presence of N fixed balls of radius b_β, $\beta = 1, \dots, N$, which are not close to each other. Then, from Theorem 1.4, equations (1.11) and (1.12) we get that the kinetic energy of the system is approximately

$$\frac{1}{2}\left(m_\alpha + \frac{2}{3}\pi a^3_\alpha \rho\right)||\vec{V}_\alpha||^2 + \frac{\rho\pi a^6_\alpha}{4}\sum_{\beta=1}^{N}\frac{b^3_\beta}{||\vec{R}_{\alpha\beta}||^6}\left\{||\vec{V}_\alpha||^2 + 3\frac{(\vec{V}_\alpha\cdot\vec{R}_{\alpha\beta})(\vec{V}_\alpha\cdot\vec{R}_{\alpha\beta})}{||\vec{R}_{\alpha\beta}||^2}\right\},$$

where $\vec{R}_{\alpha\beta} = \vec{R}_\beta - \vec{R}_\alpha$ and $\vec{R}_\beta$ is fixed. Notice that using this formula we can get an approximation for the kinetic energy of a ball of radius a moving in a fluid that contains an infinite cubic lattice of fixed equal balls of radius b, provided the lattice spacing is sufficiently larger than a and b. It is easy to check that the infinite sum that appears in this case converges absolutely. This is in contrast to the case of a ball in a box with periodic boundary conditions, where certain sum rules have to be assumed in order to overcome the problem of nonabsolute convergence of an analogous series (see [6]).

5.3. Two bodies moving on a plane. Let us consider a system of two bodies that can freely move on a plane X, Y. For simplicity we will assume that each body is symmetric with respect to two orthogonal lines. We choose the reference point in each body at the intersection of these two lines. In order to make easier the distinction between body labels and coordinate labels, we use α and β to indicate the first and the second body, respectively. Let $\vec{R}_\alpha$, $\vec{R}_\beta$, ψ_α, and ψ_β be the configuration coordinates of the system as shown in Figure 5.1. Notice that

$$T_\alpha(\psi_\alpha) = \begin{pmatrix} \cos\psi_\alpha & -\sin\psi_\alpha \\ \sin\psi_\alpha & \cos\psi_\alpha \end{pmatrix}, \quad T_\beta(\psi_\beta) = \begin{pmatrix} \cos\psi_\beta & -\sin\psi_\beta \\ \sin\psi_\beta & \cos\psi_\beta \end{pmatrix}$$

and $\vec{\omega}_\alpha = \dot{\psi}_\alpha \vec{e}_3$, $\vec{\omega}_\beta = \dot{\psi}_\beta \vec{e}_3$, where $\vec{e}_3$ is a fixed unit vector perpendicular to the plane of motion. The inertia and form coefficients of each body are as follows:

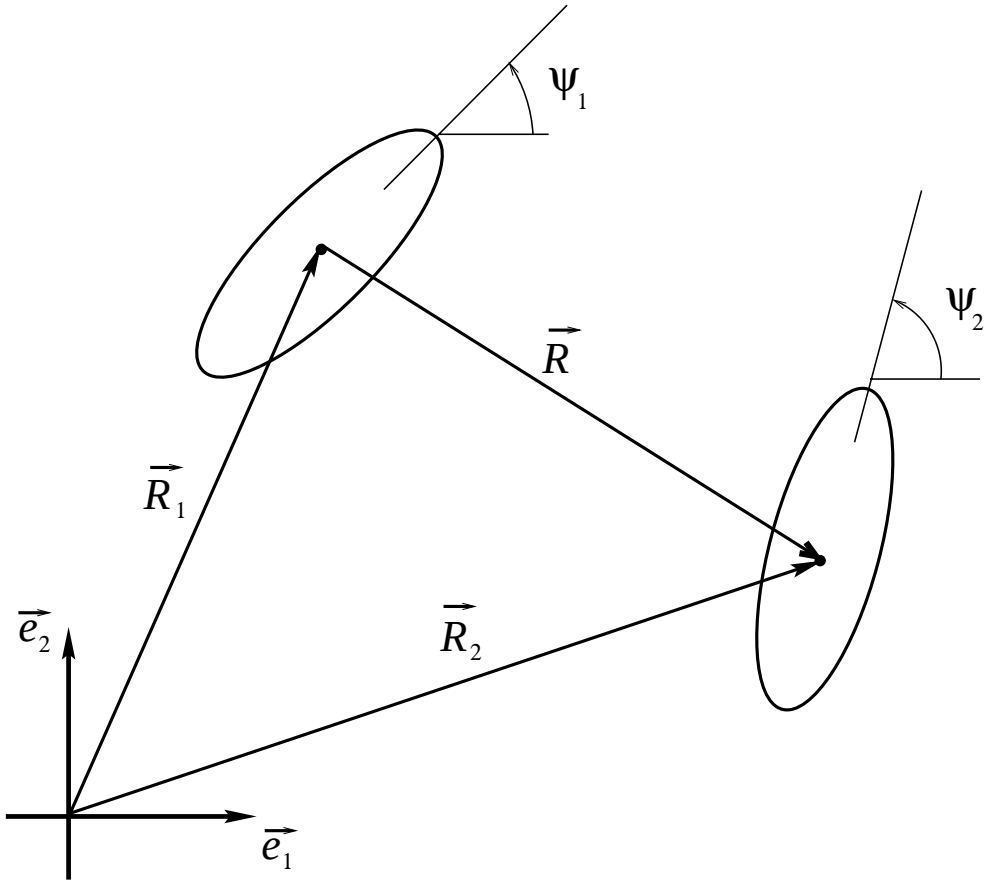


FIG. 5.1. *Diagram showing the coordinates used to describe the two-body system example of section* 5.3.

moment of inertia coefficients: I_{33}^{α} (body 1), I_{33}^{β} (body 2);
added mass matrices:

$$\begin{pmatrix} A_{11}^{\alpha} & 0 & 0 \\ 0 & A_{22}^{\alpha} & 0 \\ 0 & 0 & A_{66}^{\alpha} \end{pmatrix}, \quad \begin{pmatrix} A_{11}^{\beta} & 0 & 0 \\ 0 & A_{22}^{\beta} & 0 \\ 0 & 0 & A_{66}^{\beta} \end{pmatrix};$$

masses: m_{α} and m_{β};
inertia matrices (according to (4.2)):

$$b^{\alpha} = \begin{pmatrix} m_{\alpha} + A_{11}^{\alpha} & 0 \\ 0 & m_{\alpha} + A_{22}^{\alpha} \end{pmatrix}, \quad b^{\beta} = \begin{pmatrix} m_{\beta} + A_{11}^{\beta} & 0 \\ 0 & m_{\beta} + A_{22}^{\beta} \end{pmatrix},$$

and

$$\xi_{33}^{\alpha} = I_{33}^{\alpha} + A_{66}^{\alpha}, \quad \xi_{33}^{\beta} = I_{33}^{\beta} + A_{66}^{\beta};$$

centers of mass: $\vec{\tau}_\alpha \stackrel{\text{def}}{=} \tau_\alpha \vec{e}_{11}$ and $\vec{\tau}_\beta \stackrel{\text{def}}{=} \tau_\beta \vec{e}_{21}$;
volumes: η_α and η_β;
centers of volume: $\vec{\xi}_\alpha = \vec{0}$ and $\vec{\xi}_\beta = \vec{0}$.

For this system the form matrices λ_α and λ_β can be written as 2×2 matrices:

$$\lambda_\alpha = \begin{pmatrix} \eta_\alpha + \rho^{-1} A_{11}^\alpha & 0 \\ 0 & \eta_\alpha + \rho^{-1} A_{22}^\alpha \end{pmatrix}, \quad \lambda_\beta = \begin{pmatrix} \eta_\beta + \rho^{-1} A_{11}^\beta & 0 \\ 0 & \eta_\beta + \rho^{-1} A_{22}^\beta \end{pmatrix}.$$

The interaction matrix $F^{\alpha\beta}$ can be also written as a 2×2 matrix (since our symmetry assumptions imply $F_{16}^{\alpha\beta} = F_{26}^{\alpha\beta} = F_{61}^{\alpha\beta} = F_{62}^{\alpha\beta} = F_{66}^{\alpha\beta} = 0$):

$$\begin{pmatrix} F_{11}^{\alpha\beta} & F_{12}^{\alpha\beta} \\ F_{21}^{\alpha\beta} & F_{22}^{\alpha\beta} \end{pmatrix} = T_\alpha^\dagger G^{\alpha\beta} T_\beta.$$

Defining

$$\vec{R}_\beta - \vec{R}_\alpha \stackrel{\text{def}}{=} \vec{R} = (R_1 \vec{e}_1 + R_2 \vec{e}_2),$$

to simplify the notation, we get from (1.7)

$$G^{\alpha\beta} = \frac{1}{4\pi ||\vec{R}||^3} \left\{ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \frac{3}{||\vec{R}||^2} \begin{pmatrix} R_1 R_1 & R_1 R_2 \\ R_2 R_1 & R_2 R_2 \end{pmatrix} \right\}.$$

It is also convenient to define vectors $\vec{W}_\alpha$ and $\vec{W}_\beta$ in K given by

$$\vec{W}_\alpha \stackrel{\text{def}}{=} T_\alpha \lambda_\alpha \vec{v}_\alpha, \qquad \vec{W}_\beta \stackrel{\text{def}}{=} T_\beta \lambda_\beta \vec{v}_\beta.$$

Then from (4.17) we get

$$H^{\alpha\beta}(\vec{W}_\beta) = -\frac{3}{4\pi ||\vec{R}||^5} \cdot \left\{ \begin{pmatrix} 3W_{\beta 1} R_1 + W_{\beta 2} R_2 & W_{\beta 1} R_2 + W_{\beta 2} R_1 \\ W_{\beta 1} R_2 + W_{\beta 2} R_1 & 3W_{\beta 2} R_2 + W_{\beta 1} R_1 \end{pmatrix} - \frac{5(\vec{R} \cdot \vec{W}_\beta)}{||\vec{R}||^2} \begin{pmatrix} R_1 R_1 & R_1 R_2 \\ R_2 R_1 & R_2 R_2 \end{pmatrix} \right\}.$$

Now, from (4.30) and (4.32) we get the following equation for the velocities of body α written in the reference frame K_α:

$$\begin{aligned}
&b^\alpha \dot{\vec{v}}_\alpha + m_\alpha \tau_\alpha \ddot{\psi}_\alpha \vec{e}_{\alpha 2} - m_\alpha \tau_\alpha \dot{\psi}_\alpha^2 \vec{e}_{\alpha 1} + \dot{\psi}_\alpha \vec{e}_3 \times (b^\alpha \vec{v}_\alpha) \\
&+ (A_{11}^\alpha - A_{22}^\alpha) \dot{\psi}_\alpha \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} F^{\alpha\beta} \lambda_\beta \vec{v}_\beta \qquad (5.1) \\
&+ \rho \dot{\psi}_\beta \lambda_\alpha F^{\alpha\beta} [\vec{e}_3 \times (\lambda_\beta \vec{v}_\beta)] + \rho \lambda_\alpha F^{\alpha\beta} \lambda_\beta \dot{\vec{v}}_\beta \qquad (5.2) \\
&+ \rho \lambda_\alpha T_\alpha^\dagger H^{\alpha\beta}(\vec{W}_\beta) T_\beta \vec{v}_\beta \qquad (5.3) \\
&+ (A_{11}^\alpha - A_{22}^\alpha) \left[H_{12}^{\alpha\beta}(\vec{W}_\beta) \cos(2\psi_\alpha) + \frac{H_{22}^{\alpha\beta}(\vec{W}_\beta) - H_{11}^{\alpha\beta}(\vec{W}_\beta)}{2} \sin(2\psi_\alpha) \right] (\vec{e}_3 \times \vec{v}_\alpha) \\
&= \vec{f}_\alpha, \qquad (5.4)
\end{aligned}$$

where $\vec{f}_\alpha \in K_\alpha$ represents other forces that may act on body α. Notice that the terms in lines (5.1) and (5.2) are of order $\mathcal{O}(1/R^3)$ while those in lines (5.3) and (5.4) are of order $\mathcal{O}(1/R^4)$.

From (4.31) and (4.33) we get the following equations for the angular motion of body α:

$$\begin{aligned} &\xi^{\alpha}_{33}\ddot{\psi}_{\alpha} + m_{\alpha}\tau_{\alpha}\dot{v}_{\alpha 2} + m_{\alpha}\tau_{\alpha}\dot{\psi}_{\alpha}v_{\alpha 1} + (A^{\alpha}_{22} - A^{\alpha}_{11})v_{\alpha 1}v_{\alpha 2} \\ &\quad + (A^{\alpha}_{22} - A^{\alpha}_{11})\vec{v}_{\alpha}\cdot\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} F^{\alpha\beta}\lambda_{\beta}\vec{v}_{\beta} \\ &\quad = n_{\alpha}, \end{aligned} \tag{5.5}$$

where n_α represents other torques that may act in body α. Notice that the term in line (5.5) is of order $\mathcal{O}(1/R^3)$. To get the equations for body β it is enough to change α for β in the equations above.

Appendix. In this appendix we prove Lemma 2.2, namely we show that if δ is a solution of

$$\begin{cases} \Delta\delta(\vec{x}) = 0 & \text{for} \quad \vec{x} \in S, \\ \nabla\delta(\vec{x})\cdot\vec{n}(\vec{x}) = g(\vec{x}) = \mathcal{O}(1/R^k) & \text{for} \quad \vec{x} \in \partial S, \\ ||\nabla\delta(\vec{x})|| \to 0 & \text{as} \quad ||\vec{x}|| \to \infty, \end{cases}$$

then

$$||\delta||_{\partial S\infty} \stackrel{\text{def}}{=} \sup\{|\delta(\vec{x})| : \vec{x} \in \partial S\} = \mathcal{O}\left(\frac{1}{R^k}\right).$$

The proof requires several well-known results that can be found in partial differential equations textbooks. Here we will often refer to the book of Folland [18]. If function g is continuous, then δ is continuous in $\overline{S}$, it is differentiable in S, and its derivative in the direction of the normal to ∂S has a continuous extension to ∂S ([18, Chapter 3A]). Then for $\vec{x} \in S$, the following identity holds (Green's third identity; see the remark below Proposition 3.3 in [18]):

$$\delta(\vec{x}) = -\int_{\partial S}\delta(\vec{y})[\nabla_y\psi(\vec{x},\vec{y})\cdot\vec{n}(\vec{y})]d\sigma(\vec{y}) + \int_{\partial S}[\nabla_y\delta(\vec{y})\cdot\vec{n}(\vec{y})]\psi(\vec{x},\vec{y})d\sigma(\vec{y}), \tag{A.1}$$

where

$$\psi(\vec{x},\vec{y}) = -\frac{1}{4\pi}\frac{1}{||\vec{x}-\vec{y}||}$$

is the Newtonian potential and $d\sigma$ is the element of area on ∂S.

For any $\vec{x} \in \mathbb{R}^3$, there exists $C_1'(\alpha) > 0$, $\alpha = 1, \ldots, N$, such that

$$\int_{\partial B_\alpha}|\psi(\vec{x},\vec{y})|d\sigma(\vec{y}) < C_1'(\alpha).$$

So defining $C_1 = C_1'(1) + \cdots + C_1'(N)$ we obtain that for any $\vec{x} \in \mathbb{R}^3$:

$$\left|\int_{\partial S}[\nabla_y\delta(\vec{y})\cdot\vec{n}(\vec{y})]\psi(\vec{x},\vec{y})d\sigma(\vec{y})\right| = \left|\int_{\partial S} g(y)\psi(\vec{x},\vec{y})d\sigma(\vec{y})\right| \le C_1||g||_{\partial S\infty}. \tag{A.2}$$

Now we need the following result [18, Theorem 3.22].

THEOREM A.1. *Let ∂B be the twice continuously differentiable boundary of a bounded open set B (which may have several components) in R^3 and S be the open*

set of points in $\mathbb{R}^3$ *outside* $B \cup \partial B$. *Let* f *be a real valued continuous function in* ∂B *and* u *be a function in* $B \cup S$ *given by*

$$u(\vec{x}) = \int_S f(\vec{y})[\nabla_y \psi(\vec{x}, \vec{y}) \cdot \vec{n}(\vec{y})] d\sigma(\vec{y}),$$

where $\vec{n}$ *points outside* B. *Then* u *has continuous extensions* u_- *and* u_+ *to* $\overline{B}$ *and* $\overline{S}$, *respectively. Moreover, for* $\vec{x} \in \partial B$,

$$u_-(\vec{x}) = \frac{1}{2} f(\vec{x}) + \int_{\partial B} f(\vec{y})[\nabla_y \psi(\vec{x}, \vec{y}) \cdot \vec{n}(\vec{y})] d\sigma(\vec{y}),$$
$$u_+(\vec{x}) = -\frac{1}{2} f(\vec{x}) + \int_{\partial B} f(\vec{y})[\nabla_y \psi(\vec{x}, \vec{y}) \cdot \vec{n}(\vec{y})] d\sigma(\vec{y}).$$

Let $\vec{x}_*$ be a point in ∂B_α for a particular value of α, such that

$$|\delta(\vec{x}_*)| = ||\delta||_{\partial S \infty}.$$

Then taking $\vec{x} \to \vec{x}_*$, $\vec{x} \in S$, from Theorem A.1 we obtain

$$\delta(\vec{x}) + \int_{\partial S} \delta(\vec{y})[\nabla_y \psi(\vec{x}, \vec{y}) \cdot \vec{n}(\vec{y})] d\sigma(\vec{y}) \to \frac{1}{2}\delta(\vec{x}_*)$$
$$+ \int_{\partial B_\alpha} \delta(\vec{y})[\nabla_y \psi(\vec{x}_*, \vec{y}) \cdot \vec{n}(\vec{y})] d\sigma(\vec{y}) + \sum_{\substack{\beta \neq \alpha \\ \beta = 1}}^{N} \int_{\partial B_\beta} \delta(\vec{y})[\nabla_y \psi(\vec{x}_*, \vec{y}) \cdot \vec{n}(\vec{y})] d\sigma(\vec{y}).$$

(A.3)

If $\vec{y} \in \partial B_\beta$ and $\vec{x}_* \in \partial B_\alpha$, then $||\nabla_y \psi(\vec{x}_*, \vec{y})|| = \mathcal{O}(||\vec{x}_* - \vec{y}||^{-2})$. So there exists $\overline{R} > 0$ and $C_2'(\beta) > 0$ such that for $\beta \neq \alpha$ and $R > \overline{R}$ the following inequality holds:

$$\int_{\partial B_\beta} |[\nabla_y \psi(\vec{x}_*, \vec{y}) \cdot \vec{n}(\vec{y})]| d\sigma(\vec{y}) < \frac{C_2}{R^2}.$$

Defining $C_2 = \sum_{\beta \neq \alpha} C_2'(\beta)$ this implies that for $R > \overline{R}$,

$$\left| \sum_{\substack{\beta \neq \alpha \\ \beta = 1}}^{N} \int_{\partial B_\beta} \delta(\vec{y})[\nabla_y \psi(\vec{x}_*, \vec{y}) \cdot \vec{n}(\vec{y})] d\sigma(\vec{y}) \right| \leq ||\delta||_{\partial S \infty} \frac{C_2}{R^2}. \tag{A.4}$$

Now, taking the limit as $\vec{x} \to \vec{x}_*$, $\vec{x} \in S$, in (A.1) using (A.3) and inequalities (A.2) and (A.4), we get for $R > \overline{R}$

$$|\mathcal{F}(\delta)(\vec{x}_*)| \stackrel{\text{def}}{=} \left| \frac{1}{2}\delta(\vec{x}_*) + \int_{\partial B_\alpha} \delta(\vec{y})[\nabla_y \psi(\vec{x}_*, \vec{y}) \cdot \vec{n}(\vec{y})] d\sigma(\vec{y}) \right|$$
$$\leq ||\delta||_{\partial S \infty} \frac{C_2}{R^2} + C_1 ||g||_{\partial S \infty}. \tag{A.5}$$

From Theorem A.1 with $B = B_\alpha$ we notice that

$$\mathcal{F}(\delta)(\vec{x}_*) = \frac{1}{2}\delta(\vec{x}_*) + \int_{\partial B_\alpha} \delta(\vec{y})[\nabla_y \psi(\vec{x}_*, \vec{y}) \cdot \vec{n}(\vec{y})] d\sigma(\vec{y})$$

is the boundary value of a function δ_- that is harmonic inside the open set B_α. The mapping $\delta \to \mathcal{F}(\delta)$, from the Banach space of continuous functions on ∂B_α with the supremum norm to itself, is a bounded linear bijection (see, for instance, [18, Chapter 3, section E] or [19, section 9.2]). Thus by Banach's inverse theorem (see [20, Chap. 4, section 4, Theorem 3]) the inverse operator of $\mathcal{F}$ is also bounded, namely, there exists a constant C_3 which depends on ∂B_α such that

$$||\delta||_{\partial S\infty} = ||\delta||_{\partial B_\alpha \infty} \leq C_3 \sup_{\vec{x}\in\partial B_\alpha} \left| \frac{1}{2}\delta(\vec{x}) + \int_{\partial B_\alpha} \delta(\vec{y})[\nabla_y \psi(\vec{x},\vec{y})\cdot \vec{n}(\vec{y})]d\sigma(\vec{y}) \right|.$$

From this inequality and inequality (A.5) we obtain

$$||\delta||_{\partial S\infty} \leq ||\delta||_{\partial S\infty} \frac{C_3 C_2}{R^2} + C_1 C_3 ||g||_{\partial S\infty},$$

or using that $||g||_{\partial S\infty} = \mathcal{O}(1/R^k)$,

$$||\delta||_{\partial S\infty} \left(1 - \frac{C_3 C_2}{R^2}\right) \leq \mathcal{O}\left(\frac{1}{R^k}\right),$$

which implies

$$||\delta||_{\partial S\infty} = \mathcal{O}\left(\frac{1}{R^k}\right).$$

Acknowledgments. I am very grateful to J. A. P. Aranha for all his suggestions and for reading the original manuscript. I thank J. Koiller for having suggested some references. I am especially grateful to the late Prof. Daniel Henry to whom this paper is dedicated. He had always helped me in so many different subjects including some of those addressed in the present paper.

REFERENCES

[1] H. Lamb, *Hydrodynamics*, 6th ed., Dover, New York, 1932.
[2] G. Birkhoff, *Hydrodynamics*, Princeton University Press, Princeton, NJ, 1960.
[3] A. R. Galper and T. Miloh, *Motion stability of a deformable body in an ideal fluid with applications to the N spheres problem*, Phys. Fluids, 10 (1998), pp. 119–130.
[4] H. S. Kim and A. Prosperetti, *Numerical simulation of the motion of rigid spheres in potential flow*, SIAM J. Appl. Math., 52 (1992), pp. 1533–1562.
[5] L. van Wijngaarden, *Hydrodynamic interaction between gas bubbles in a liquid*, J. Fluid Mech., 77 (1976), p. 27.
[6] P. Smereka, *On the motion of bubbles in a periodic box*, J. Fluid Mech., 93 (1993), pp. 79–112.
[7] A. R. Galper and T. Miloh, *Hydrodynamics and stability of a deformable body moving in the proximity of interfaces*, Phys. Fluids, 11 (1999), pp. 795–806.
[8] N. E. Leonard and J. E. Marsden, *Stability and drift of underwater vehicle dynamics: Mechanical systems with rigid motion symmetry*, Phys. D, 105 (1997), p. 130.
[9] M. S. Howe, *On the force and moment on a body in an incompressible fluid, with application to rigid bodies and bubbles at high and low Reynolds number*, Quart. J. Mech. Appl. Math., 48 (1995), pp. 401–426.
[10] A. B. Basset, *A Treatise on Hydrodynamics with Numerous Examples*, Vol. 1, Dover, New York, 1961.
[11] A. N. Simos, E. A. Tannuri, C. P. Pesce, and J. A. P. Aranha, *A quasi-explicit hydrodynamic model for the dynamic analysis of a moored FPSO under current action*, J. Ship Research, 45 (2001), pp. 289–301.
[12] E. A. Tannuri, A. N. Simos, A. J. P. Leite, and J. A. P. Aranha, *Fishtailing instability of a moored FPSO: Theoretical prediction and experiments*, J. Ship Research, 45 (2001), pp. 302–314.

[13] M. J. Lighthill, *Fundamentals concerning wave loading on offshore structures*, J. Fluid Mech., 173 (1986), pp. 667–681.
[14] T. Sarpkaya, *On the force decompositions of Lighthill and Morrison*, J. Fluids Struct., 15 (2001), pp. 227–233.
[15] C. Grotta Ragazzo, *Dynamics of many bodies inside a liquid: Added-mass tensor of compounded bodies and systems with a fast oscillating body*, Phys. Fluids, 14 (2002), pp. 1590–1600.
[16] V. V. Kozlov, *Integrability and non-integrability in Hamiltonian mechanics*, Russian Math. Surveys, 38 (1983), pp. 1–76.
[17] I. Stakgold, *Boundary Value Problems of Mathematical Physics*, Vol. 2, MacMillan, New York, 1968.
[18] G. B. Folland, *Introduction to Partial Differential Equations*, Princeton University Press and University of Tokyo Press, Princeton, NJ, 1976.
[19] P. Garabedian, *Partial Differential Equations*, Chelsea, New York, 1964.
[20] A. N. Kolmogorov and S. V. Fomin, *Elements of The Theory of Functional Analysis*, 3rd ed., Mir, Moscow, 1978 (in Spanish).

SIAM J. APPL. MATH.
Vol. 63, No. 6, pp. 1998–2019

A CANARD MECHANISM FOR LOCALIZATION IN SYSTEMS OF GLOBALLY COUPLED OSCILLATORS*

HORACIO G. ROTSTEIN[†‡], NANCY KOPELL[†], ANATOL M. ZHABOTINSKY[‡], AND IRVING R. EPSTEIN[‡]

Abstract. Localization in a discrete system of oscillators refers to the partition of the population into a subset that oscillates at high amplitudes and another that oscillates at much lower amplitudes. Motivated by experimental results on the Belousov–Zhabotinsky reaction, which oscillates in the relaxation regime, we study a mechanism of localization in a discrete system of relaxation oscillators globally coupled via inhibition. The mechanism is based on the canard phenomenon for a single relaxation oscillator: a rapid explosion in the amplitude of the limit cycle as a parameter governing the relative position of the nullclines is varied. Starting from a parameter regime in which each uncoupled oscillator has a large amplitude and no other periodic or other stable solutions, we show that the canard phenomenon can be induced by increasing a global negative feedback parameter γ, with the network then partitioned into low and high amplitude oscillators. For the case in which the oscillators are synchronous within each of the two such populations, we can assign a canard-inducing critical value of γ separately to each of the two clusters; localization occurs when the value for the system is between the critical values of the two clusters. We show that the larger the cluster size, the smaller is the corresponding critical value of γ, implying that it is the smaller cluster that oscillates at large amplitude. The theory shows that the above results come from a kind of self-inhibition of each cluster induced by the local feedback. In the full system, there are also effects of interactions between the clusters, and we present simulations showing that these nonlocal interactions do not destroy the localization created by the self-inhibition.

Key words. canard phenomenon, globally coupled oscillators, relaxation oscillator, localization of oscillations

AMS subject classifications. 34C15, 34C26

DOI. 10.1137/S0036139902411843

1. Introduction.

The Belousov–Zhabotinsky (BZ) reaction is the prototype system in nonlinear chemical dynamics [1, 2, 3] (see references therein). In bulk, it is a relaxation oscillator. A wide variety of spatially extended patterns have been found in experiments on this reaction. Along with the experiments, chemically plausible mathematical models have been proposed and studied both analytically and numerically. The results obtained qualitatively reproduce experimental findings. Recently, new patterns have been found as nondiffusive couplings have been experimentally and numerically introduced.

In particular, the existence of localized oscillatory clusters has been reported in [4, 5] for the BZ reaction with global inhibitory feedback. Simulations performed on the Oregonator model [5] and a modified Oregonator model [6] of the BZ reaction, both with global inhibitor feedback, reproduce the experimental findings. However, the mechanism by which localized cluster formation occurs remains unclear from both the mathematical and chemical points of view.

*Received by the editors July 22, 2002; accepted for publication (in revised form) February 28, 2003; published electronically September 4, 2003.

http://www.siam.org/journals/siap/63-6/41184.html

[†]Department of Mathematics and Center for Biodynamics, Boston University, Boston, MA 02215 (horacio@bu.edu, nk@bu.edu). The work of the first author was partially supported by the Burroughs Wellcome Fund. The work of the second author was partially supported by NSF grant DMS-9706694.

[‡]Department of Chemistry and Volen Center for Complex Systems, Brandeis University, Waltham, MA 02454 (zhabotin@brandeis.edu, epstein@brandeis.edu). The work of the third and fourth authors was partially supported by NSF grant CHE-9988463.

We consider as a cluster a set of "cells" or "chemical points" of the reactor that oscillate synchronously with the same amplitude. In certain cases, depending on some parameter, only two different amplitude regimes of oscillations occur: large amplitude oscillations (LAO) and small amplitude oscillations (SAO). The LAO regime consists of limit cycles whose amplitudes are O(1), i.e., almost equal to the maximum amplitude of the limit cycle for a single uncoupled oscillator, whereas the SAO regime consists of limit cycles whose amplitudes are of order of magnitude $\epsilon \ll 1$. In such cases there is a range of amplitudes that is not observable because they occur in an exponentially small interval of the governing parameter. When the system is divided into two or more clusters and at least one oscillates in an LAO regime and one in an SAO regime, we say that the clusters are localized. Clusters that are in the same amplitude regime may oscillate with a small difference in their amplitudes (compared with the LAO), but we do not refer to that situation as a localization phenomenon. Note that our definition of clusters does not require oscillators in each cluster to be spatially grouped; we disregard spatial structure in this work.

The localized cluster patterns found in the experiments and simulations on the BZ reaction with global inhibitory feedback [4, 5, 6] present two main features that might seem counterintuitive:

1. Two different oscillatory regimes coexist in a system of identical coupled oscillators.
2. The cluster with the largest number of oscillators is always in the SAO regime whereas the smallest clusters are in an LAO regime; one might expect the largest cluster to be oscillating in an LAO regime and suppressing the smaller ones as occurs in other systems, e.g., neural systems with inhibitory synapses, with all-to-all identical coupling. In this latter case, if one cluster has a larger number of cells than the others, the former can suppress the latter.

In this paper we seek to explain the mechanism of localization for globally coupled relaxation oscillators of the FitzHugh–Nagumo (FHN) type, along with the two features mentioned above. The FHN-type models were chosen as simplifications of the Oregonator models. Although they are not precise as descriptions of chemical phenomena, they display the localization phenomenon and are easier to study analytically in order to give some insight into the dynamical mechanisms that produce localized solutions. In addition, they display some of the relevant qualitative features of the modified Oregonator model studied in [6], such as the shape of the nullclines, the fact that the limit cycle is created in a supercritical Hopf bifurcation, and the relaxation nature of the oscillator. In a forthcoming paper we will address the questions related to the mechanism of localization in the modified version of the Oregonator used in [6, 7].

We argue that the mechanism of localization is based on the canard phenomenon that occurs in single relaxation oscillators. The canard phenomenon is a very rapid change in the amplitude of the limit cycle of a relaxation oscillator as the inhibitor nullcline moves with respect to the activator nullcline [11, 12, 13, 14, 15, 16]. It arises in the context of experiments and simulations of nonlinear chemical dynamics [17, 18, 19] and in a two-pool model describing the mechanism of calcium-induced calcium release [20, 21].

The mechanism of localization in oscillatory systems, not of the relaxation type, has been studied in [8, 9, 10] for nonidentical diffusively coupled oscillators. To our knowledge, the mechanism of localization in relaxation-type oscillators has not been analyzed before.

In section 2 we present a general formulation for a class of models of globally coupled oscillators of FHN type that include the modified FHN (MFHN) models studied in this manuscript. For single FHN-type oscillators, the fast variable nullcline is cubic-like and intersects, on its middle branch, the slow variable nullcline, an increasing function. The parameters were chosen such that this intersection is an unstable fixed point. In section 3 we explain the reduction of dimensions strategy, a self-consistent argument that reduces the dimensionality of the mathematical problem by assuming the existence of M clusters, each with a different dynamical behavior (different amplitudes, phases or both). Within each cluster, the oscillators synchronize.

In section 4 we describe the canard phenomenon for a single FHN-type equation and review some results. Following [15], we present a mathematical expression for an asymptotic approximation to the "canard critical value" for the parameter λ, the parameter responsible for the displacement of the slow variable nullcline relative to the fast variable nullcline, as a function of the remaining parameters of the model. When, by increasing or decreasing λ through a critical value λ_c, there is a sudden change (of canard type) in the amplitude of the limit cycle, we say that the canard phenomenon has been induced by changes in λ, and we call λ_c the canard critical value of λ. Strictly speaking, the sudden change in the amplitude of the limit cycle takes place in an exponentially small interval of values of λ; the canard critical value is the limit of the interval as $\epsilon \to 0$.

In section 5 we show that, when there are synchronized (bulk) oscillations for the globally coupled system (only one cluster), the canard phenomenon may be induced by increasing the value of the global feedback parameter γ and keeping λ fixed. In the FHN-type models with global feedback presented here, as well as in the BZ model with global feedback used in [6], the intersection point between nullclines remains fixed as γ is increased. When $\gamma = 0$ (no global feedback) the uncoupled oscillators are in an LAO regime; localization for these models is a consequence of the global coupling. An asymptotic approximation to the critical global feedback value, γ_c, is also calculated as a function of λ and other parameters of the model. These results are the basis of our analysis of the localization phenomenon.

The localization phenomenon for a two-cluster system, in which one cluster is in an LAO regime and the other is in an SAO regime, is analyzed in section 6. The dynamics of the two-cluster globally coupled system is analyzed by studying each cluster separately and considering the other cluster as forcing it. Under specific assumptions, this dynamics is a combination of self-inhibition of each cluster, responsible for creating an interval of values of γ within which a localized solution may exist, and inhibition (forcing) exerted on each cluster by the remaining ones. We show that self-inhibition is stronger the larger the cluster size, which explains why in a localized solution the largest cluster is in an SAO regime. We show that, for the special case of the van der Pol (VDP) equations with global feedback, localization is produced by only the self-inhibition, and the forcing exerted on each cluster by the other does not affect localization. In this paper we analyze only the effect of self-inhibition; however, we present some simulations of other globally coupled FHN systems that support our claim that the localization phenomenon is present with the same features predicted theoretically. In section 7, we relate our results to experiments and simulations.

2. Models. In this paper we study models of the type

$$\begin{cases} v_k' = F(v_k, w_k) - \gamma\ (\langle w \rangle - \bar{w}), \\ w_k' = \epsilon\ G(v_k, w_k; \lambda) \end{cases} \tag{1}$$

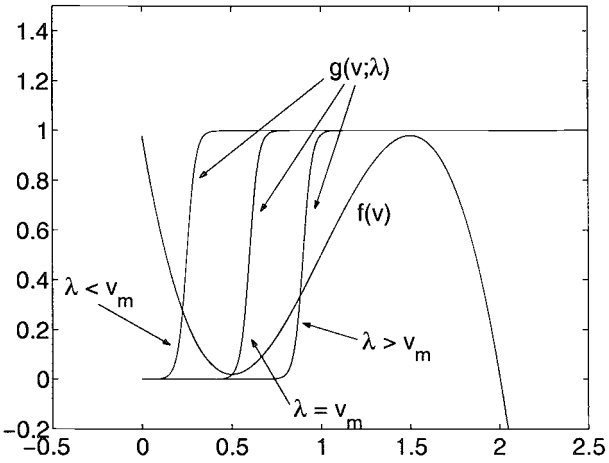


FIG. 1. *Nullclines for a sigmoid version of the FHN model for several values of* $\lambda = \bar{v} - v_{min}$ *(the intersection point between nullclines).*

for $k = 1, \ldots, N$ and $0 < \epsilon \ll 1$. In (1), $F(v, w)$ is such that the zero level curve $F(v, w) = 0$ can be expressed as $w = f(v)$ with $f(v)$ a cubic-like function having one local minimum at (v_m, w_m) and one local maximum at (v_M, w_M) with $v_m < v_M$ and $w_m < w_M$. The function G is a nonincreasing function of w such that the zero level curve $G(v, w; \lambda) = 0$ is an increasing function of v for every λ in a given neighborhood of $\lambda = 0$ and is also a decreasing function of λ for all v in a neighborhood of v_m. We further assume that $F = 0$ and $G = 0$ intersect at $(\bar{v}, \bar{w})$ with $\bar{v} = v_m$ when $\lambda = 0$ and that $(\bar{v}, \bar{w})$ is an unstable fixed point lying on the central branch of f when $\lambda > 0$. The constant γ is the global feedback parameter, and $\langle w \rangle$ is given by

$$\langle w \rangle = \frac{1}{N} \sum_{k=1}^{N} w_k. \tag{2}$$

Note that $(\bar{v}, \bar{w})$ does not depend on γ, as we can see by replacing $\langle w \rangle$ by $\bar{w}$ in (1). In all models considered here, the systems are assumed to be in a relaxation oscillatory regime in the absence of global coupling ($\gamma = 0$). For $\gamma = 0$, changes in the parameter λ alter the position of the w_k nullcline (see Figure 1). When this nullcline moves, the intersection point $(\bar{v}, \bar{w})$ changes. As we will explain in section 4, without loss of generality we can redefine λ such that $\lambda = \bar{v}$. In the literature v is usually referred to as the activator or the "potential" variable and w as the "inhibitor" or the recovery variable.

Some specific systems may be modeled by making simplifying assumptions on (1) and considering $F(v, w) = f(v) - w$ and $G(v, w; \lambda) = g(v; \lambda) - w$, where f is as described before and g is an increasing function of v for every λ in a given neighborhood of $\lambda = 0$ and a decreasing function of λ for all v in a neighborhood of v_m. Examples are

(i) VDP equations in Lienard form

$$f(v) = -v^3 + v^2, \qquad G(v, w; \lambda) = v - \lambda; \tag{3}$$

(ii) the classical FHN equations

$$f(v) = -h\ v^3 + a\ v^2 - b\ v + c, \qquad g(v; \lambda) = \beta\ v - \eta, \tag{4}$$

where h, a, b, c, β, and η are nonnegative constants;

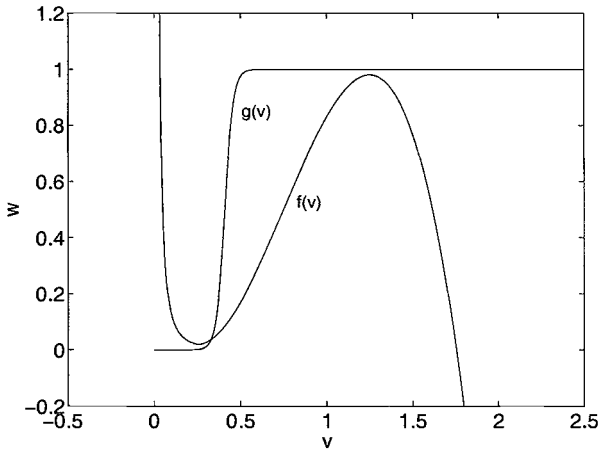


FIG. 2. *Nullclines for the MFHN model for a single oscillator (or equivalently for* $\gamma = 0$*). The values of the parameters used in our simulations are* $h = 1.92$*,* $a = 4.32$*,* $b = 1.8$*,* $c = 0.23$*,* $\beta = 0.41$*,* $\eta = 0.05$*,* $\epsilon = 0.05$*,* $v_m = 0.25$*,* $f(v_m) = 0.02$*,* $\bar{v} = 0.328$*,* $\bar{w} = 0.036092$*. The function* f *was constructed in such a way that the maximum and minimum are close to* 1 *and* 0*, respectively, in the following way:* (i) *we took a cubic function with minimum* 0 *at* $v = 0$ *and maximum* 0.98 *at* $v = 1$*,* (ii) *we shifted it up by* 0.02*,* (iii) *we shifted it to the right by* 0.25*. The function* g*, a sigmoid function, was built in such a way that it crosses* f *at a single point,* $\bar{v}$*, placed to the right of the Hopf bifurcation and such that* $\bar{v} > \lambda_c$ *(beyond the canard critical value for a single oscillator). Note that* g *is very steep and* $\lim_{v\to\pm\infty} = \pm 1$*.*

(iii) the sigmoid FHN equations

$$f(v) = -h\ v^3 + a\ v^2 - b\ v + c, \qquad g(v;\lambda) = \frac{1}{2}\ (\tanh((v-\beta)/\eta) + 1)\,, \tag{5}$$

where h, a, b, c, β, and η are nonnegative constants; and

(iv) the MFHN equations, which we use in our simulations,

$$f(v) = \begin{cases} f_{cub}(v), & v \geq v_m, \\ f_{cub}(v_m)\ v_m^2/v^2, & v \leq v_m, \end{cases}$$
$$g(v;\lambda) = \frac{1}{2}\ (\tanh((v-\beta)/\eta) + 1)\,, \tag{6}$$

and

$$f_{cub} = -h\ v^3 + a\ v^2 - b\ v + c. \tag{7}$$

Here v_m is the minimum of f_{cub}; a, b, c, h, β, η, and ϵ are nonnegative constants. In our simulations we use the following values for the parameters: $h = 1.92$, $a = 4.32$, $b = 1.8$, $c = 0.23$, $\beta = 0.41$, $\eta = 0.05$, and $\epsilon = 0.05$. With those parameters we get $(v_m, w_m) = (0.25, 0.02)$ and $(\bar{v}, \bar{w}) = (0.328, 0.036092)$. We can see the graph of the corresponding nullclines in Figure 2. In (4), (5), and (6) the parameter λ (which was defined as the v-coordinate of the intersection point between the two nullclines of the system) is implicitly defined by other parameters of the model.

The MFHN model is a simplification of the modified version of the Oregonator model used in [6]; it allows an easier qualitative dynamical understanding by reproducing important aspects of the BZ dynamics and keeping some of its features, including

the "N" shape of the nullcline corresponding to the first equation in (1), its asymptotic approach to the w axis, its qualitative behavior as a function of the global feedback parameter, and an inhibitor dynamics described by a sigmoid function rather than a line. The motivation for using the MFHN system instead of more classical versions of the FHN system is that, by changing the global feedback parameter γ, we can find small amplitude limit cycles with smaller amplitude in the v direction than for the FHN equations. This is due to the fact that the activator nullcline is asymptotic to the w axis.

3. Strategy: Reduction of dimension using clusters. We are interested in localized solutions to (1)–(2) for $\gamma \neq 0$, in which two different portions of the system display LAO and SAO, respectively. Toward this end we will analyze the existence, properties, and stability of solutions to models of type (1)–(2) with M different oscillatory behaviors ($M \leq N$). More specifically, we will look for solutions to (1)–(2) in which the system of N oscillators is divided into M different sets, each set containing a fraction α_k, $k = 1, \ldots, M$, of the N oscillators with $\sum_{k=1}^{M} \alpha_k = 1$, and such that all oscillators in a set synchronize and oscillate with the same amplitude.

Since all the oscillators in each set are equivalent, we can write

$$\langle w \rangle = \sum_{j=1}^{M} \alpha_j \, w_j. \tag{8}$$

Bulk oscillations correspond to $M = 1$. Two-phase (phase-locked) oscillations correspond to $M = 2$, as do localized oscillations in which a fraction of the system oscillates with large amplitude and the rest of the system oscillates with small amplitude. $M = 3$ includes three-phase (phase-locked) oscillations and localized oscillations in which a fraction of the system displays two-phase (phase-locked) LAO and the rest of the system oscillates with small amplitude.

In order to consider the influence of the rest of the system on the kth oscillator, we define

$$S_k = \sum_{j=1, j\neq k}^{M} \alpha_j \, w_j \tag{9}$$

for $k = 1, \ldots, M$. Using (9) and (1), we obtain

$$\begin{cases} v_k' = F(v_k, w_k) - \gamma \, \alpha_k \, w_k + \gamma \, \bar{w} - \gamma \, S_k, \\ w_k' = \epsilon \, G(v_k, w_k; \lambda) \end{cases} \tag{10}$$

for $k = 1, \ldots, M$. Note that the last term in the first equation of (10) is the only one depending on w_j, $j = 1, \ldots, M$, $j \neq k$. This term can be seen as a forcing exerted by the rest of the oscillators on the kth one. For $M = 1$ ($S_k = 0$, $\alpha_1 = 1$), (10) is an unforced oscillator with global coupling; it describes bulk oscillations of the whole system. For $M > 1$, the inhibitor nullsurfaces are not dependent on γ or S_k, while the activator nullsurfaces, which are solutions of

$$F(v_k, w_k) - \gamma \, \alpha_k \, w_k + \gamma \, \bar{w} - \gamma \, S_k = 0 \tag{11}$$

for $k = 1, \ldots, M$, vary depending on S_k and γ.

When $F(v, w) = f(v) - w$ in (10) we have FHN-type equations. In this case the activator nullsurfaces are given by

$$w_k = \frac{f(v_k) + \gamma \, \bar{w}}{1 + \gamma \, \alpha_k} - \frac{\gamma}{1 + \gamma \, \alpha_k} \, S_k \tag{12}$$

for $k = 1, \ldots, M$. For each k, the solutions (v_k, w_k) of the FHN-type equations can be considered as living in a three-dimensional space (v_k, w_k, S_k). The activator nullsurfaces vary in the S_k direction. As the system evolves, S_k changes in a periodic fashion. For each value of S_k we can consider the projection of (12) onto the (v_k, w_k) plane. This gives us the possibility of looking at the phase space of the FHN-type equations for each oscillator separately as if it were two-dimensional, with the activator nullcline moving up and down periodically according to S_k, i.e., according to the dynamics of the rest of the $M - 1$ oscillators. The intersection point in the (v_k, w_k) plane between the projections of the inhibitor and activator nullsurfaces becomes a periodic function of t that moves as the w_k nullcline moves. We call the v_k-coordinate of this time-dependent intersection point $\lambda_k = \lambda_k(t)$ for $k = 1, \ldots, M$. Thus, for systems of the form $f(v, w) = f(v) - w$, we can decompose the whole system into M forced subsystems of FHN type, one for each value of k. The forcing exerted on one oscillator depends on the remaining ones.

Stability of a solution to (10) does not automatically imply stability with respect to the full equations (1), since the solution may not be stable to perturbations that destroy the clustering into groups of equivalent oscillators. Hence once a solution has been numerically found for a specific model and value of M, it is desirable to check its stability in the N-array of globally coupled oscillators. We approach this problem numerically. In order to numerically solve system (1) we used the modified Euler method [22] for $N = 100$ with a step size $\Delta t = 0.01$. For $M = 2$ we divided the N oscillators into two sets, each with uniform initial conditions. Once each set of oscillators (with $\alpha_k N$ oscillators belonging to each set, $k = 1, 2$) settled down in a specific limit cycle, we applied a random perturbation of maximum amplitude 0.001 to each variable. We applied the following criterion for stability: if, after the perturbation, each oscillator returns to its original limit cycle and phase difference, then we say that the system is N-stable (numerically stable in an array of N oscillators). Otherwise we say that the system is N-unstable (numerically unstable in an array of N oscillators). We are aware that our definition of stability is not a rigorous one and can be affected by numerical instabilities. Still, it gives us valuable information about the stability of phase and localized clusters for the subset of values of γ for which they exist.

4. Canard phenomenon. In this section we review the canard phenomenon for relaxation oscillators. Consider system (1) for a single oscillator and $\gamma = 0$; i.e.,

$$\begin{cases} v' = F(v, w), \\ w' = \epsilon\, G(v, w; \lambda), \end{cases} \tag{13}$$

where $0 < \epsilon \ll 1$ and where F and G are as described in section 2. We first look at FHN-type models; i.e., $F(v, w) = f(v) - w$ with f as described in section 2. We assume that system (13) is in an oscillatory regime. The nullclines for a sigmoid-type FHN model and a limit cycle corresponding to a chosen set of parameters are shown in Figure 3.

The dynamics of system (13) depends on the value of λ, i.e., on the relative position of the w nullcline with respect to the v nullcline. For the FHN-type equations there exists a Hopf bifurcation point $\lambda_H(\epsilon) \geq v_m$ in a neighborhood of (v_m, w_m) which converges to (v_m, w_m) as $(\epsilon, \lambda) \to (0, 0)$ (see Appendix C and Figure 1). For values of $\lambda < v_H$ system (13) has a steady state as the only attractor, and the system is excitable [1, 3, 23, 24]; i.e., relatively small perturbations (but large enough to exceed a threshold, a curve in phase space, determined by the v nullcline and the parameters of the model) give rise to a large excursion that returns to the attractor. This trajectory

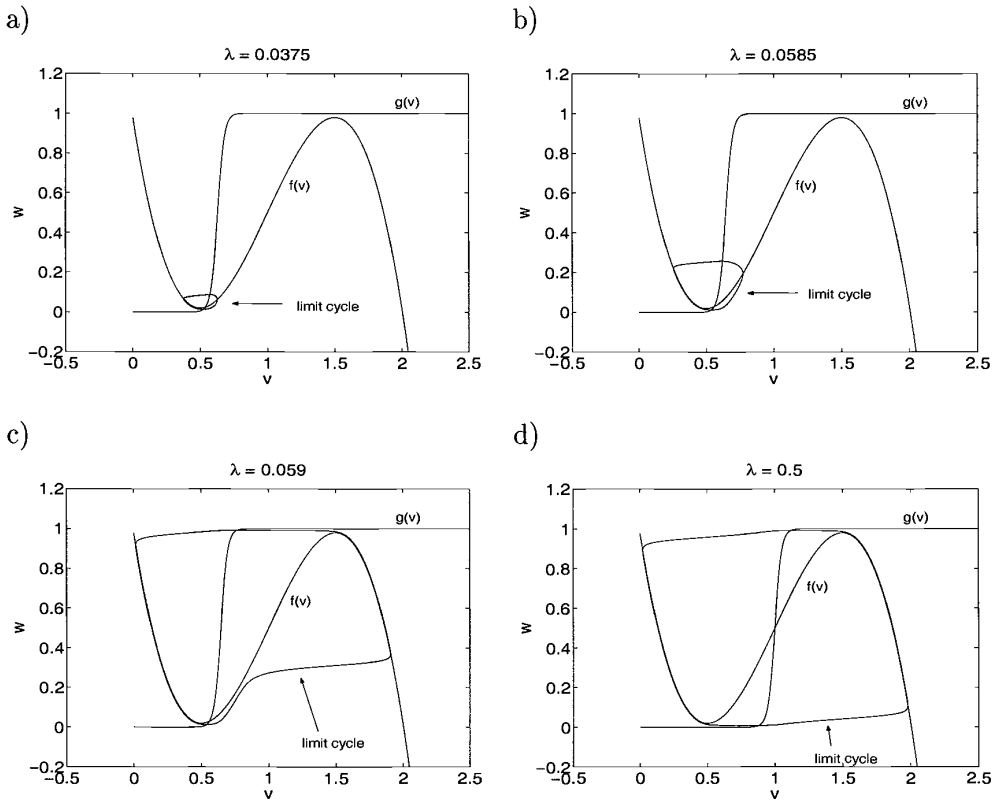


FIG. 3. *Nullclines and limit cycle for a sigmoid version of the FHN model where* $\lambda = \bar{v} - v_{min}$. *The values of the parameters are as in Figure* 2 *and* (a) $\lambda = 0.0375$, (b) $\lambda = 0.0585$, (c) $\lambda = 0.059$, (d) $\lambda = 0.5$.

is usually called an excitation loop, a pulse, or a spike. Subthreshold perturbations return to the stable fixed point with no large excursion. At $\lambda = \lambda_H(\epsilon)$, system (13) undergoes a supercritical Hopf bifurcation. As λ increases, the amplitude of the limit cycle increases slowly for small enough values of λ, part of the trajectory being very close to the unstable branch of the v nullcline for a while, then crossing the unstable branch and moving toward the left branch of the v nullcline, as illustrated in Figure 3(a) and 3(b). At some critical point $\lambda_c(\epsilon) > \lambda_H(\epsilon)$, the trajectory moves toward the right branch of the activator nullcline instead of moving toward the left branch, and the limit cycle expands rapidly (over an exponentially small interval in the parameter λ) becoming a relaxation oscillator [1, 3, 25, 26], as seen in the transition from Figure 3(b) to 3(c). After that, the amplitude of the limit cycle either increases slowly or remains constant as λ is increased, until the oscillator becomes like the one in Figure 3(d). By symmetry, when λ is near the maximum of $w = f(v)$, the same effect is seen in a small neighborhood of the Hopf bifurcation near $\lambda = v_M$. In Figure 4 we can see the amplitude of the limit cycle, given by the maximum and minimum values of v and w ($v_{\min}$, $v_{\max}$, $w_{\min}$, and $w_{\max}$), as a function of λ for the sigmoid version of the FHN equations (5). This rapid change from a "small" amplitude limit cycle to a "large" amplitude limit cycle is known as the canard phenomenon [11, 12, 14, 15, 16, 27]. In this case the canard phenomenon has been induced by changes in λ. Here we concentrate on the canard phenomenon near $v = v_m$.

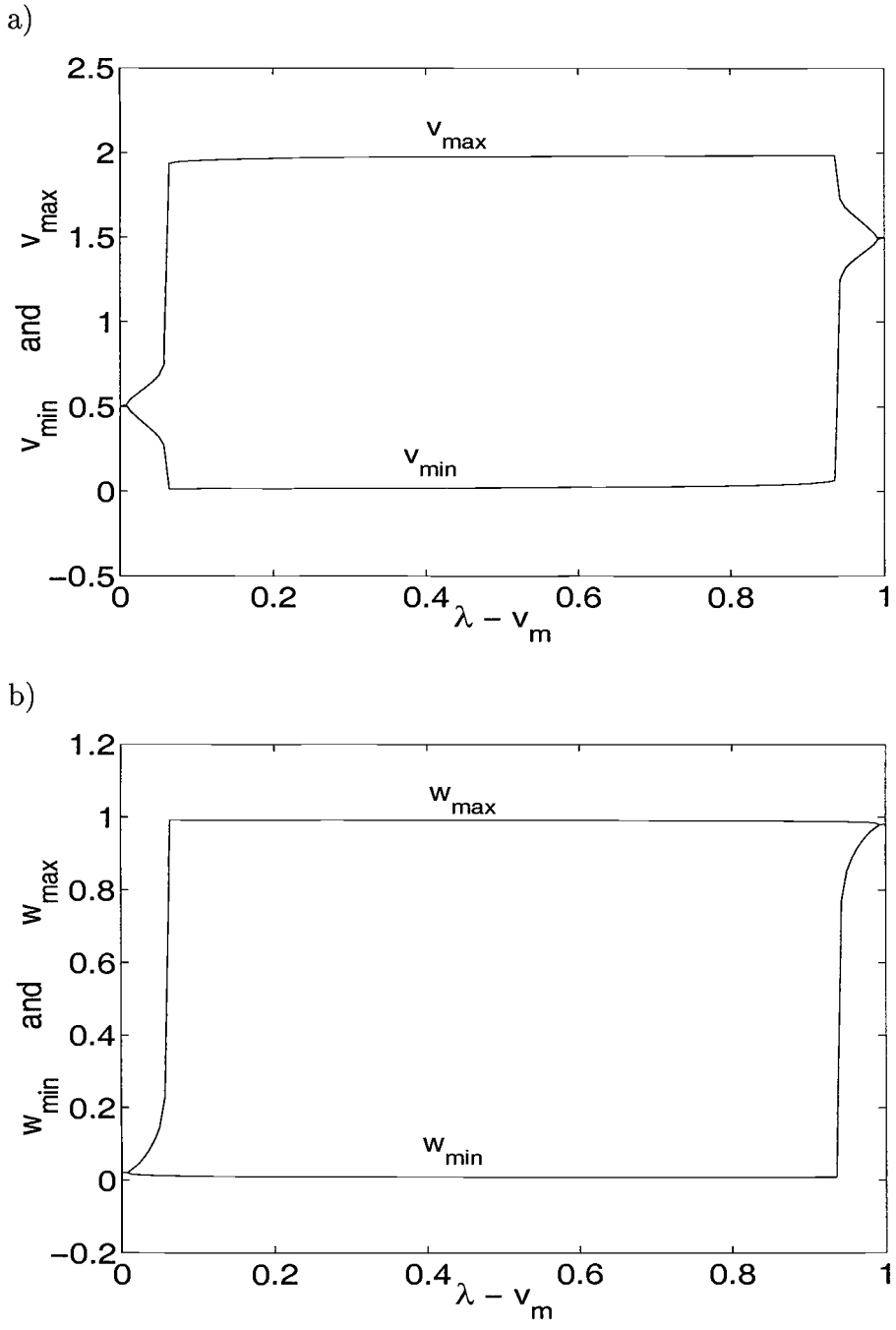


FIG. 4. *Amplitude of the limit cycle as a function of the crossing point between the activator and inhibitor nullclines λ for a single cell FHN oscillator. The values of the parameters are as in Figure* 2. (a) *v-amplitude (v_{min} and v_{max}),* (b) *w-amplitude (w_{min} and w_{max}).*

The canard phenomenon was discovered by Benoit et al. [14] for the VDP oscillator. In their work [14] they show that there exists a critical value $\lambda_c(\epsilon)$ of λ such that for λ in a small neighborhood of λ_c the limit cycle deforms into a curve similar to the one shown in Figure 3(c). While Benoit et al. [14] used nonstandard analysis techniques in their study, Eckhaus [12] and Baer and Emeux [13] used asymptotic techniques. In particular they found expressions for the canard critical value λ_c for VDP-type equations and for a generalization of system (13).

The canard phenomenon for (13) has been also studied by Dumortier and Roussarie [11] and by Krupa and Szmolyan [15, 16]. We follow the latter authors in our approach. In order to present their results, without loss of generality, we take $(v_m, w_m) = (0, 0)$. For $(v, w) = \mathbf{0}$ we assume that $F(\mathbf{0}) = \mathbf{0}$, $\partial F/\partial v(\mathbf{0}) = \mathbf{0}$, $\partial^2 F/\partial v^2(\mathbf{0}) \neq \mathbf{0}$, i.e., (v_m, w_m) is a nondegenerate local minimum (fold point) of the nullcline $F(v, w) = 0$ for λ in a suitable interval. Furthermore, $\partial F/\partial w(\mathbf{0}) \neq \mathbf{0}$. We also assume for $(v, w, \lambda) = \mathbf{0}$ that $G(\mathbf{0}) = \mathbf{0}$, $\partial G/\partial v(\mathbf{0}) \neq \mathbf{0}$, and $\partial G/\partial \lambda(\mathbf{0}) \neq \mathbf{0}$. These conditions defining a canard point (which will be referred to as *canard conditions*) mean that the nullcline $G(v, w, \lambda) = 0$ is transverse to the nullcline $F(v, w) = 0$, and it passes through the fold point with nonzero speed as λ varies. As pointed out above, $\bar{v}$ increases as λ increases (see Figure 1), allowing us to reparametrize λ such that $\lambda = \bar{v}$. For the VDP and FHN equations, these assumptions are satisfied with an appropriate change of variables.

We show in Appendix B that

$$\lambda_c = \Lambda\,\epsilon + |F_w|\,\Upsilon\,\epsilon + \mathcal{O}(\epsilon^{3/2}), \tag{14}$$

where

$$\Upsilon = \frac{G_v}{2\,F_{vv}\,|G_\lambda|}\left(\frac{G_v}{F_{vv}}\right)_v, \qquad \Lambda = -\frac{G_v}{2\,F_{vv}^3\,|G_\lambda|}\,(\,G_v\,F_{vw}\,F_{vv} + G_w\,F_{vv}^2\,), \tag{15}$$

and all the functions are calculated at $\mathbf{0}$. To obtain (14) we used an earlier result by Krupa and Szmolyan [15]. There, the cubic-like function was assumed to have its minimum at $(0, 0)$.

For the VDP equations (3), $\Lambda = 0$ and $\Upsilon = -f'''(0)\,/\,2\,(f''(0))^2 = 3/4$, coinciding with the expression found by Eckhaus [12]. For the classical FHN equations (4), $\Lambda = \beta\,/\,(2\,f''(0)\,|g_\lambda|)$ and $\Upsilon = -\beta\,f'''(0)\,/\,(2\,(f''(0))^3\,|g_\lambda|)$. The expression for the canard critical value becomes

$$\lambda_c = \frac{\beta}{2\,(f''(0))^3\,|g_\lambda(0)|}\,[\,(f''(0))^2 - \beta\,f'''(0)\,]\,\epsilon + \mathcal{O}(\epsilon^{3/2}) = \frac{\beta\,(\,2\,a^2 + 3\,\beta\,h\,)}{8\,a^3\,|g_\lambda(0)|}\,\epsilon + \mathcal{O}(\epsilon^{3/2}). \tag{16}$$

For the general FHN-type equations with $G(v, w; \lambda) = g(v; \lambda) - w$,

$$\lambda_c = \frac{g'(0)}{2\,(f''(0))^3\,|g_\lambda(0)|}\,[\,(f''(0))^2 - f'''(0)\,g'(0) + g''(0)\,f''(0)\,]\epsilon + \mathcal{O}(\epsilon^{3/2})$$

$$= \frac{g'(0)}{8\,a^3\,|g_\lambda(0)|}\,[\,2\,a^2 + 3\,h\,g'(0)\, + a\,g''(0)\,]\epsilon + \mathcal{O}(\epsilon^{3/2}). \tag{17}$$

Note that if the minimum of the activator nullcline $(v_m, f(v_m)) \neq (0, 0)$, a translation of coordinates may be performed without changing the values of the derivatives of f and g.

By construction, f in the MFHN model is a matching of two different functions. The result is continuous but not differentiable at the origin. In order for the theory described in this section to be applicable, F and G must be C^k-functions with $k \geq 3$ (continuous at least up to the third derivative) [16]. In the analysis presented here we consider functions f qualitatively similar to the MFHN function defined above, i.e., satisfying the canard conditions, but C^k with $k \geq 3$. Our numerical simulations with the MFHN function qualitatively agree with the analytical predictions.

5. Canard phenomenon induced by the global feedback parameter in synchronized (bulk) oscillatory systems. In this section we study the influence of the global feedback parameter γ on the amplitude regime (LAO or SAO) of the solution for $M = 1$ (bulk or synchronized oscillations). In what follows, all functions are calculated at $\mathbf{0}$. For $M = 1$, system (1) reads as

$$\begin{cases} v' = F(v,w) - \gamma\ w + \gamma\ \bar{w}, \\ w' = \epsilon\ G(v,w;\lambda). \end{cases} \tag{18}$$

We assume that at $\gamma = 0$, $\lambda_c = \mathcal{O}(\epsilon)$ such that $\lambda_c \neq \mathcal{O}(\epsilon^\nu)$, $\nu > 1$, $\lambda = \mathcal{O}(\epsilon)$ fixed, and $\lambda > \lambda_c$; i.e., the system is in an LAO regime for $\gamma = 0$.

First we explain how to apply the theory developed in section 4 to system (18). In the calculation of the canard critical value we use the fact that in a neighborhood of $(v_m, f(v_m)) = (0,0)$ the v nullcline can be described by a parabolic function (see (48) in Appendix A). Then, by our assumption $\lambda = \mathcal{O}(\epsilon)$ ($\bar{v} = \mathcal{O}(\epsilon)$), it follows that $\gamma\bar{w} = \mathcal{O}(\epsilon^2)$. We can rescale the last term in the first equation in (18) by defining $\bar{w} = \kappa\,\epsilon^2$, getting the following expression for the v nullcline: $\Phi(v,w,\epsilon) := F(v,w) - \gamma\ w + \gamma\ \kappa\,\epsilon^2$. Note that κ is independent of v and that $\Phi(0,0,0) = 0$ as required in [15].

Remark. When the activator nullcline $\Phi = 0$ is ϵ-dependent, the canonical equations (48) are augmented by a term $\epsilon\, h_6(v,w,\epsilon)$ and the expression for the canard critical value has an extra term proportional to $h_{6,v}$ [15].[1] For (18) $h_6 = \gamma\,\epsilon\,\kappa$, so $h_{6,v} = 0$ and the expression for the canard critical value is not affected. The only effect of γ on the canard critical value comes from the term $-\gamma\, w$ in the first equation in (18).

An expression for the canard critical value as a function of the global feedback parameter γ can be calculated as in the calculation for $\gamma = 0$ (see Appendix B) to obtain

$$\lambda_c(\gamma) = \Lambda\,\epsilon + (\,|F_w| + \gamma\,)\,\Upsilon\,\epsilon + \mathcal{O}(\epsilon^{3/2}) = \lambda_c(0) + \gamma\ \Upsilon\ \epsilon + \mathcal{O}(\epsilon^{3/2}). \tag{19}$$

Expressions (19) and (15) imply that, by increasing the value of the global feedback parameter, the value of the canard critical value is increased, provided

$$\left(\frac{G_v}{F_{vv}}\right)_v > 0, \tag{20}$$

since G_v and F_{vv} were assumed to be positive. So, if for $\gamma = 0$ we have $\lambda > \lambda_c(0)$ (the system is in an LAO regime), then the canard phenomenon can be induced by increasing the value of γ without changing the value of λ. The change from LAO to SAO takes place in an interval of values of γ exponentially small in ϵ.

We now compute γ_c, the amount that γ must be increased (in the limit as $\epsilon \to 0$) to induce the canard phenomenon, assuming that the system is in an LAO regime

[1] See the explanation after (53).

when $\gamma = 0$. Taking into account the assumptions made at the beginning of this section on λ and λ_c, the critical value $\gamma_c(\lambda)$ of γ may be calculated as the value of γ that brings $\lambda_c(\gamma)$ to λ, i.e., by replacing γ by γ_c and λ_c by λ, respectively, in (19):

$$\lambda = \lambda_c(0) + \gamma_c(\lambda)\ \Upsilon\ \epsilon + \mathcal{O}(\epsilon^{3/2}). \tag{21}$$

From (14)

$$|F_w|\,\Upsilon\,\epsilon = \lambda_c(0) - \Lambda\,\epsilon + \mathcal{O}(\epsilon^{3/2}). \tag{22}$$

Substituting (22) into (21), multiplied by $|F_w|$, we get

$$|F_w|\,[\,\lambda - \lambda_c(0)\,] = \gamma_c\,[\,\lambda_c(0) - \Lambda\,\epsilon\,] + \mathcal{O}(\epsilon^{3/2}). \tag{23}$$

Note that in the FHN models $|F_w| = \mathcal{O}(1)$, which has been used in the error term in (23). Rearranging terms and using (14), we get

$$\gamma_c(\lambda) = |F_w|\,\frac{\lambda - \lambda_c(0)}{\lambda_c(0) - \Lambda\,\epsilon} + \mathcal{O}\left(\frac{\epsilon^{3/2}}{\lambda_c(0) - \Lambda\,\epsilon}\right) = |F_w|\,\frac{\lambda - \lambda_c(0)}{\lambda_c(0) - \Lambda\,\epsilon} + \mathcal{O}(\epsilon^{1/2}). \tag{24}$$

For the VDP equations (3), $F_w = -1$ and G is independent of w so we have $\Lambda = 0$. Thus from (19)

$$\lambda_c(\gamma) = \lambda_c(0)\ (1 + \gamma) + \mathcal{O}(\epsilon^{3/2}), \tag{25}$$

and

$$\gamma_c(\lambda) = \frac{\lambda - \lambda_c(0)}{\lambda_c(0)} + \mathcal{O}(\epsilon^{1/2}). \tag{26}$$

Note that, since G is independent of w, the nullclines intersect at the same value λ for all $\gamma \geq 0$.

Using (19) and (24), the expressions for λ_c and γ_c for the FHN-type equations with $G(v, w; \lambda) = g(v; \lambda) - w$ are given by

$$\lambda_c(\gamma) = \lambda_c(0) + \frac{\gamma\ g'(0)}{2\ f''(0)\ |g(\lambda)|}\ \left[\frac{g''(0)\ f''(0) - g'(0)\ f'''(0)}{[f''(0)]^2}\right]\ \epsilon + \mathcal{O}(\epsilon^{3/2}) \tag{27}$$

and

$$\gamma_c(\lambda) = \frac{\lambda - \lambda_c(0)}{\lambda_c(0)}\ \left[1 + \frac{g'(0)}{2\ \lambda_c(0)\ f''(0)\ |g_\lambda|}\,\epsilon\right] + \mathcal{O}(\epsilon^{1/2}). \tag{28}$$

In both cases, as γ increases, the canard critical value moves to the right; then there exists a critical value of the global feedback parameter, γ_c, such that for values of γ below (above) γ_c, solutions display LAO (SAO).

As noted above, for the MFHN model used in our simulations, we do not have an expression for the canard critical value as a function of the parameters of the model and γ, but we conjecture on the basis of numerical simulations that the behavior is similar to the smooth case described before. The results of numerical simulations are shown in Figures 5 and 6. In Figure 5 we see the dependence of the amplitude of the limit cycle (represented by the minimum and maximum values of v and w) on γ for the MFHN model. We observe that for $\gamma = \gamma_c$ (in this case $\gamma_c \sim 0.429$), there is a sudden change in both the v- and w-amplitudes of the limit cycle.

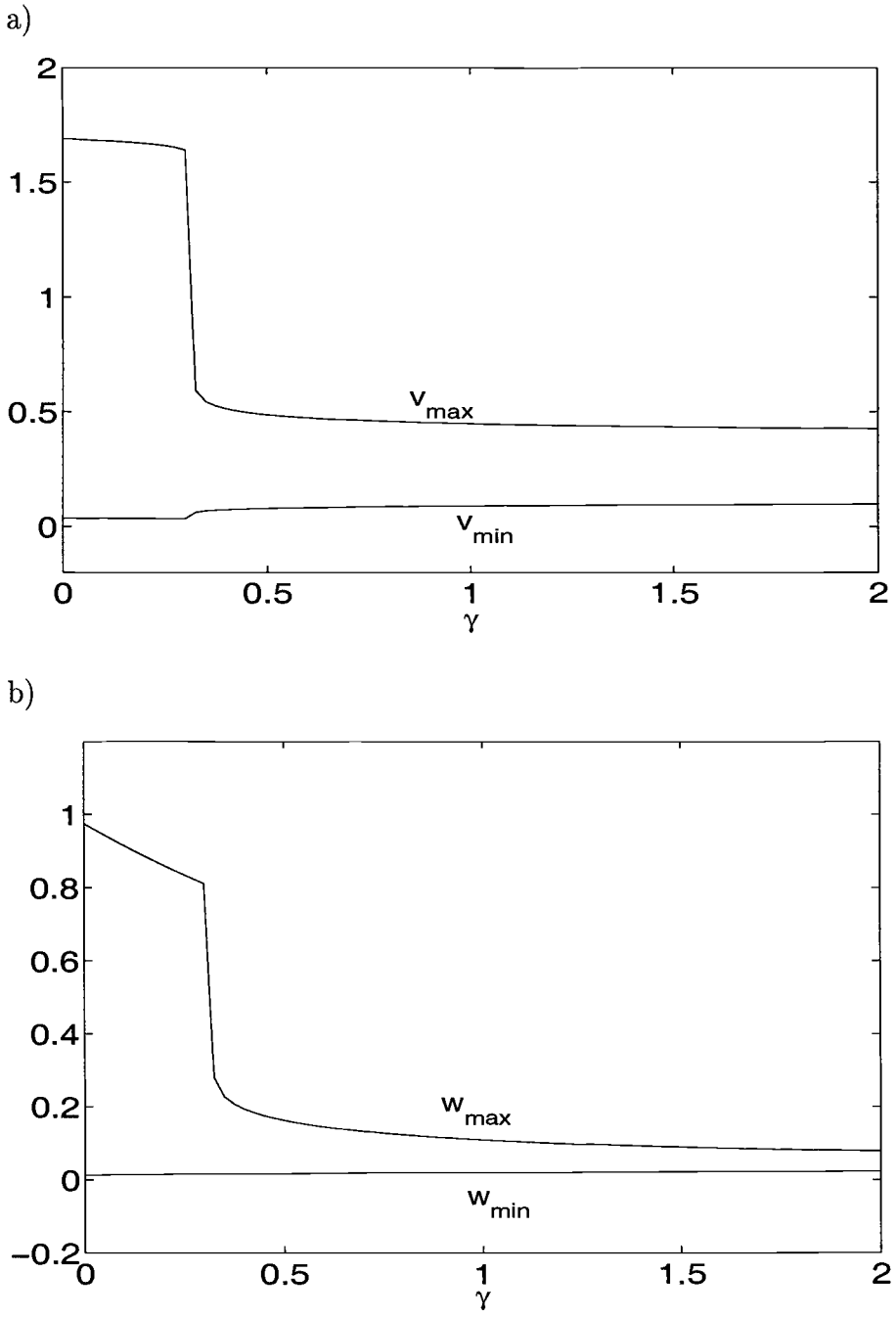


FIG. 5. *Amplitude of the limit cycle for the MFHN oscillator as a function of the global feedback parameter* γ. *The values of the parameters are as in Figure* 2.

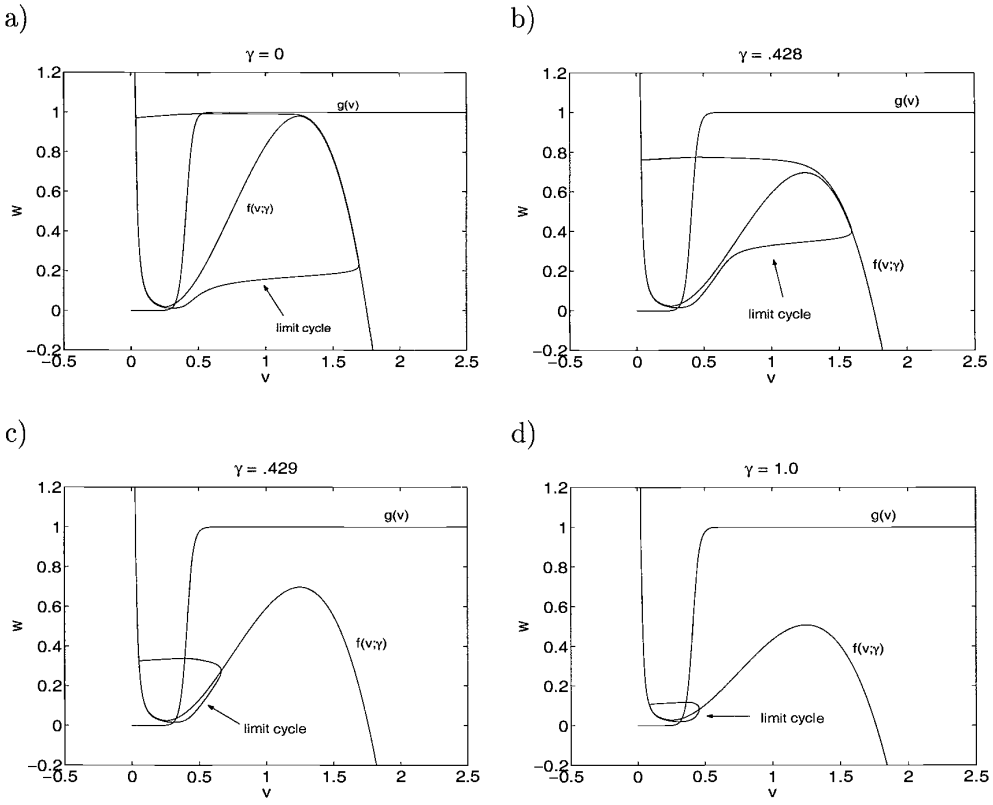


FIG. 6. *Nullclines and phase plane for the MFHN oscillator for various values of* γ. *The function* $f(v;\gamma) = (f(v) + \gamma\,\bar{w})/(1+\gamma)$. *The values of the parameters are as in Figure* 2 *and* (a) $\gamma = 0$, (b) $\gamma = 0.428$, (c) $\gamma = 0.429$, (d) $\gamma = 1$.

In Figure 6 we show the shape of the limit cycle for several values of γ above and below γ_c. For $\gamma = 0$ (Figure 6(a)) the system is in a relaxation oscillation regime. As γ increases, the limit cycle goes through the lower knee of the activator nullcline, comes up along the unstable branch for a while, and then moves rapidly to the right branch of the activator nullcline (Figure 6(b)) if $\gamma < \gamma_c$; if $\gamma > \gamma_c$, the trajectory crosses the unstable branch and moves rapidly to the left branch of the activator nullcline (Figure 6(c)).

Our numerical simulations show that bulk oscillations for the MFHN model are 100-stable for $\gamma \leq 0.39$ and $\gamma \geq 25.0$.

6. Localized solutions. In this section we analyze the existence of localized solutions for a system of globally coupled FHN-type equations, i.e., equations (1)–(2), where $F(v,w) = f(v) - w$. We deal here with the case $M = 2$; this can be easily generalized to larger values of M. In a two-cluster localized solution, some of the oscillators are in an SAO regime while the other oscillators are in an LAO regime.

By applying the reduction of dimensions described in section 3 we reduce the system of N oscillators to a system of two oscillators. The activator nullcline for each oscillator is given by (12) for $k = 1, 2$. The first term in (12) depends only on v_k and the second term is independent of (v_k, w_k) and is the only one depending on w_j, $j = 1, 2$, $j \neq k$. As explained in section 3, we can consider the second term in (12)

as moving the nullcline, whose shape is given by the first term in (12), up and down. We call $\lambda_{k,c}(\gamma)$ and $\gamma_{k,c}(\lambda)$ the canard critical value and the critical global feedback parameter value, respectively, for $k = 1, 2$. Looking at each of the two oscillators separately we can calculate the respective canard critical values as a function of γ and the fraction of oscillators in each cluster, α_k, following the same reasoning leading to (19) and (24) in section 5, where γ is replaced by $\alpha_k\,\gamma$. This yields

$$\lambda_{k,c}(\gamma) = \lambda_c(0) + \alpha_k\,\gamma\ \Upsilon\ \epsilon + \mathcal{O}(\epsilon^{3/2}) \tag{29}$$

and

$$\gamma_{k,c}(\lambda) = \frac{1}{\alpha_k}\,\frac{\lambda - \lambda_c(0)}{\lambda_c(0) - \Lambda\,\epsilon} + \mathcal{O}(\epsilon^{1/2}) \tag{30}$$

for $k = 1, 2$. Thus, for a given γ, the larger α_k the larger the canard critical point for the kth oscillator and the smaller the corresponding $\gamma_{k,c}$, i.e., the less the global feedback needed to get SAO. We can easily calculate

$$\lambda_{1,c}(\gamma) - \lambda_{2,c}(\gamma) = (\alpha_1 - \alpha_2)\,\gamma\ \Upsilon\ \epsilon + \mathcal{O}(\epsilon^{3/2}) \tag{31}$$

and

$$\gamma_{2,c} - \gamma_{1,c} = \frac{\alpha_1 - \alpha_2}{\alpha_1\ \alpha_2}\,\frac{\lambda - \lambda(0)}{\lambda_c(0) - \Lambda\,\epsilon} + \mathcal{O}(\epsilon^{1/2}). \tag{32}$$

For the VDP equations, the value of $\lambda_k(t)$ (see section 3 for the definition of this quantity) does not depend either on k or on t. Let us refer to it as λ. In this case, expression (31) implies that if $\alpha_1 \neq \alpha_2$, then we can find values of the global feedback parameter γ for which λ has a value between $\lambda_{1,c}$ and $\lambda_{2,c}$, thus producing a localized solution. As we can see from (31) and (32) the interval of values of λ and γ for which we can expect a localized solution increases with the difference between the fractions of oscillators in the two clusters. Since $\lambda_k(t)$ is independent of k and t, localization in the VDP model is a consequence only of nonsymmetric self-inhibition, i.e., not a consequence of the forcing that the oscillators exert on one another. Note that the cluster with the larger α_k is the one in the SAO regime, as seen in experiments and simulations on the BZ reaction with global feedback. (In the latter case the LAO regime consisted of two phase locked clusters.) The shape, frequency, and amplitude of each limit cycle (considered separately) in the localized solutions depend, in ways that are not yet fully understood, on γ, on the size of the other oscillator, and possibly on other quantities.

For the FHN-type equations self-inhibition creates intervals of critical values of λ and γ given by (31) and (32), respectively. In contrast to the VDP equations, when $\gamma > 0$, $\lambda_k(t)$ (see section 3 for the definition of this quantity) depends on both k and t. The forcing exerted on each oscillator by the other one changes the value of $\lambda_k(t)$. Thus there are two effects we must consider in understanding how localized solutions arise: self-inhibition and external inhibition or forcing. How external forcing interacts with localization is not yet understood. However, our simulations for the MFHN model show that localization is present as expected and that the numerically determined interval $\gamma_{2,c} - \gamma_{1,c}$ in which there is localization increases as $\alpha_1 - \alpha_2$ increases. In Figure 7 we show the amplitude of the solutions to the MFHN system with $M = 2$ as a function of the global feedback parameter γ for different values

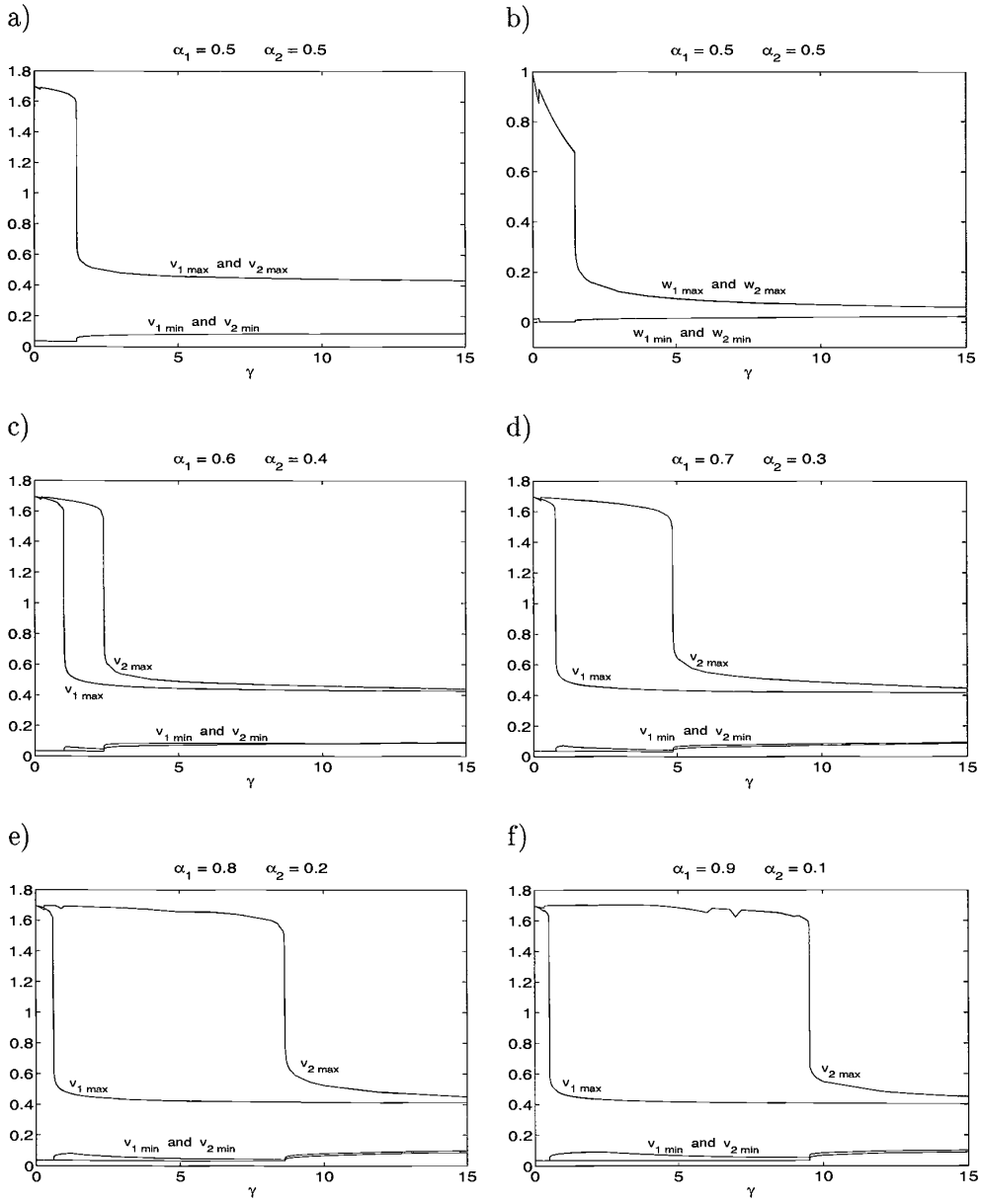


FIG. 7. *Amplitude of the limit cycle for* 2 *globally coupled* ($M = 2$) *MFHN oscillators as a function of the global feedback parameter* γ *and for different values of the fraction of oscillators in each cluster. The parameters are as in Figure* 2 *and* (a) $\alpha_1 = 0.5$, *v-amplitude,* (b) $\alpha_1 = 0.5$, *w-amplitude,* (c) $\alpha_1 = 0.6$, *v-amplitude,* (d) $\alpha_1 = 0.7$, *v-amplitude,* (e) $\alpha_1 = 0.8$, *v-amplitude,* (f) $\alpha_1 = 0.9$, *v-amplitude.*

of α_k, $k = 1, 2$. The amplitude of the oscillatory solutions for v_k is represented as the minimum and maximum values $v_{k,min}$ and $v_{k,max}$, respectively. In Table 1 we present numerical approximations of $\gamma_{1,c}$ and $\gamma_{2,c}$ for different values of α_1 and α_2. For $\alpha_1 = 0.6,\ 0.7$, and 0.8 we found that the localized solutions corresponding to a subset of values of γ included in $(\gamma_{1,c}, \gamma_{2,c})$ are 100-stable.

TABLE 1

Localized solution for the MFHN model. Values of the canard critical values $\gamma_{c,1}$ and $\gamma_{c,2}$ as a function of the fraction of oscillators in each cluster. The values of the parameters are as in Figure 2 with $M = 2$. The intervals of 100-stability, I, are (i) $I = 1.01$ *for* $\alpha_1 = 0.6$, (ii) $I = 0.6$ *for* $\alpha_1 = 0.7$ *and* $I = 0.05$ *for* $\alpha_1 = 0.8$. *For* $\alpha_1 = 0.9$ 100*-stable localized solutions were not found.*

α_1	α_2	$\gamma_{c,1}$	$\gamma_{c,2}$
0.5	0.5	1.47	1.47
0.6	0.4	1.01	2.40
0.7	0.3	0.77	4.85
0.8	0.2	0.6	8.64
0.9	0.1	0.5	9.53

The analysis presented in this section can be generalized for larger values of M, in which case we will have two regimes (LAO and SAO), but in each of the regimes we can have different amplitudes or phases for different clusters.

7. Discussion. In this paper we analyze the mechanism of localization of oscillations for a globally coupled system of relaxation oscillators of FHN type. In addition to localization, these models display the basic features of the modified Oregonator models studied in [6] and [5] to reproduce the experimental results: shape of the nullclines, a limit cycle created in a supercritical Hopf bifurcation, and display of canards among others.

Although the present study is motivated by the BZ reaction, a spatially extended system, experimental evidence suggests that the phenomena studied here, the mechanism of localization or creation of localized clusters, does not depend on diffusion [4, 28]. Based on the results of the simulations presented in [6] and simulations performed by the authors and not presented here, we conjecture that the diffusion plays an important role in spatially grouping together oscillators belonging to the same cluster.

We analyzed the canard phenomenon induced by the global feedback parameter γ for bulk oscillations ($M = 1$), obtaining an expression for the canard critical value λ_c and the critical global feedback parameter γ_c as functions of the parameters of the models considered. We showed that, by increasing the value of the global feedback parameter, the canard phenomenon is induced for a critical value γ_c; i.e., as γ passes γ_c the system rapidly changes from an LAO regime to an SAO regime due to self-inhibition. Our numerical stability calculations show that this limit cycle need not be 100-stable in a neighborhood of γ_c; e.g., for values of γ close enough to γ_c, bulk oscillations lose stability, generating other patterns, among them localized structures. The idea of induction of the canard phenomenon by changing γ is a key to the analysis of localization.

We used the idea of self-inhibition to partially explain the two-cluster localization phenomenon ($M = 2$) for a system of FHN-type equations. We applied the reduction of dimension via clusters, and we analyzed each of the two oscillators separately, considering each as a forcing exerted on the other. By writing the equations for the nullclines of each oscillator, we saw that their dynamics can be understood as a combination of two phenomena: self-inhibition of each oscillator and inhibition (forcing) exerted on each oscillator by the remaining ones. Self-inhibition creates intervals of critical values of λ and γ given by (31) and (32), respectively. The forcing exerted on each oscillator by the other one changes the values of $\lambda_k(t)$. We did not analyze the effect of the forcing exerted on each oscillator by the remaining one, but

we studied this effect numerically, showing that the main features of localization are present; i.e., the larger cluster is in an SAO regime, and the larger the size difference between two clusters the larger the interval of values of γ for which the system has a localized solution, which is 100-stable. Our analysis reveals that for the VDP equations, localization is produced by self-inhibition alone. For systems that are not of FHN type (e.g., the BZ equations [6]), the analysis becomes more complicated.

In experiments on the BZ reaction with global inhibitory feedback [4, 5] as well as in simulations using an Oregonator model [5] and another BZ model [6], localized structures consisted of three clusters, the largest cluster in an SAO regime and two smaller phase-locked clusters in an LAO regime. The mechanism we propose here for FHN-type models does not deal with the multiple clusters in LAO regimes but does explain the counterintuitive inverse relation between amplitude regime and cluster size and sheds light on the role of self-inhibition in the phenomenon of localization.

We conjecture that a similar mechanism is responsible for localization in a modified Oregonator model for the BZ reaction [6] that we study in a forthcoming paper, as well as in the Oregonator model [5]. The canard phenomenon for a single two-dimensional Oregonator model has been studied in [29], although in this case the Hopf bifurcation taking place in a neighborhood of the minimum of the activator nullcline may be subcritical instead of supercritical; then SAO are not possible for a single oscillator, though they might be possible in a globally coupled system. Our preliminary analysis shows that global feedback changes the stability type of the Hopf bifurcation point, thus allowing for SAO.

Appendix A. Calculation of the canonical form. The first step in calculating the canard critical value for system (13) is to transform it into its canonical form. We assume $(v_m, w_m) = 0$.

We first expand the right-hand sides in both equations in (13) in Taylor series:

$$\begin{cases} F(v,w) = -b\ w + a\ v^2 + H_1(v,w), \\ G(v,w,\lambda) = e\ v - c\ \lambda + d\ w + H_2(v,w,\lambda), \end{cases} \tag{33}$$

where

$$a = \frac{1}{2}\frac{\partial^2 F}{\partial v^2}(\mathbf{0}), \qquad b = \left|\frac{\partial F}{\partial w}(\mathbf{0})\right|, \tag{34}$$

$$c = \left|\frac{\partial G}{\partial \lambda}(\mathbf{0})\right|, \qquad d = \frac{\partial G}{\partial w}(\mathbf{0}), \qquad e = \frac{\partial G}{\partial v}(\mathbf{0}), \tag{35}$$

$$H_1(v,w) = \frac{\partial^2 F}{\partial v w}(\mathbf{0})\, v\, w + \frac{1}{6}\frac{\partial^3 F}{\partial v^3}(\mathbf{0})\, v^3 + \mathcal{O}(w^2, v^2 w, v w^2, w^3), \tag{36}$$

$$H_2(v,w,\lambda) = \frac{1}{2}\frac{\partial^2 G}{\partial v^2}(\mathbf{0})\, v^2 + \frac{\partial^2 G}{\partial v \lambda}(\mathbf{0})\, v\, \lambda + \mathcal{O}(w^2, \lambda^2, v\, w, w\lambda). \tag{37}$$

In (34) and (36) $\mathbf{0} = (0,0)$, whereas in (35) and (37) $\mathbf{0} = (0,0,0)$.

Next, we substitute (33) into (13), getting

$$\begin{cases} v' = -b\ w + a\ v^2 + H_1(v,w), \\ w' = \epsilon\,[\, e\ v - c\ \lambda + d\ w + H_2(v,w,\lambda)\,]. \end{cases} \tag{38}$$

Finally, we rescale system (38) by defining

$$V = e^{1/2} b^{1/2} a^{-1}, \quad W = e\, a^{-1}, \quad L = e^{3/2} b^{1/2} a^{-1} c^{-1}, \quad T = e^{-1/2} b^{-1/2}, \tag{39}$$

$$\hat{v} = \frac{v}{V}, \qquad \hat{w} = \frac{w}{W}, \qquad \hat{\lambda} = \frac{\lambda}{L}, \qquad \hat{t} = \frac{t}{T}, \tag{40}$$

$$\hat{H}_1(\hat{v}, \hat{w}) = \frac{T}{V} H_1(V\,\hat{v}, W\,\hat{w})$$

$$= T\,W\,\frac{\partial^2 F}{\partial \hat{v}\hat{w}}(\mathbf{0})\,\hat{v}\,\hat{w} + \frac{1}{6}\,T\,V^2\,\frac{\partial^3 F}{\partial \hat{v}^3}(\mathbf{0})\,\hat{v}^3 + \mathcal{O}(\hat{w}^2, \hat{v}^2\hat{w}, \hat{v}\hat{w}^2, \hat{w}^3)$$

$$= \hat{w}\,\left[T\,W\,\frac{\partial^2 F}{\partial \hat{v}\hat{w}}(\mathbf{0})\,\hat{v} + \mathcal{O}(\hat{w})\right] + \hat{v}^2\,\left[\frac{1}{6} T\,V^2\,\frac{\partial^3 F}{\partial \hat{v}^3}(\mathbf{0})\,\hat{v} + \mathcal{O}(\hat{w})\right], \tag{41}$$

and

$$\hat{H}_2(\hat{v}, \hat{w}, \hat{\lambda}) = \frac{T}{W} H_2(V\,\hat{v}, W\,\hat{w}, L\,\hat{\lambda})$$

$$= \frac{T\,V^2}{2W}\,\frac{\partial^2 G}{\partial \hat{v}^2}(\mathbf{0})\,\hat{v}^2 + \frac{T\,V\,L}{W}\,\frac{\partial^2 G}{\partial \hat{v}\hat{\lambda}}(\mathbf{0})\,\hat{v}\,\hat{\lambda} + \mathcal{O}(\hat{w}^2, \hat{\lambda}^2, \hat{v}\,\hat{w}, \hat{w}\hat{\lambda})$$

$$= \hat{v}\,\left[\frac{T\,V^2}{2\,W}\frac{\partial^2 G}{\partial \hat{v}^2}(\mathbf{0})\,\hat{v} + \mathcal{O}(\hat{w})\right] + \hat{\lambda}\,\left[\frac{T\,V\,L}{W}\frac{\partial^2 G}{\partial \hat{v}\hat{\lambda}}(\mathbf{0})\,\hat{v} + \mathcal{O}(\hat{w}, \hat{\lambda})\right] \tag{42}$$

and substituting (39)–(42) into (38). Calling

$$h_1(\hat{v}, \hat{w}) = -T\,W\frac{\partial^2 F}{\partial \hat{v}\hat{w}}(\mathbf{0})\,\hat{v} + \mathcal{O}(\hat{w}) = b^{-1/2}\,e^{1/2}\,a^{-1}\,\frac{\partial^2 F}{\partial \hat{v}\,\hat{w}}(\mathbf{0})\,\hat{v} + \mathcal{O}(\hat{w}), \tag{43}$$

$$h_2(\hat{v}, \hat{w}) = \frac{1}{6} T\,V^2\,\frac{\partial^3 F}{\partial \hat{v}^3}(\mathbf{0})\,\hat{v} + \mathcal{O}(\hat{w}) = \frac{1}{6}\,e^{1/2}\,b^{1/2}\,a^{-2}\,\frac{\partial^3 F}{\partial \hat{v}^3}(\mathbf{0})\,\hat{v} + \mathcal{O}(\hat{w}), \tag{44}$$

$$h_3(\hat{v}, \hat{w}, \hat{\lambda}) = \frac{T\,V^2}{2\,W}\,\frac{\partial^2 G}{\partial \hat{v}^2}(\mathbf{0})\,\hat{v} + \mathcal{O}(\hat{w}) = \frac{1}{2}\,e^{-1/2}\,b^{1/2}\,a^{-1}\,\frac{\partial^2 G}{\partial \hat{v}^2}(\mathbf{0})\,\hat{v} + \mathcal{O}(\hat{w}), \tag{45}$$

$$\begin{aligned} h_4(\hat{v}, \hat{w}, \hat{\lambda}) &= -\frac{T\,V\,L}{W}\,\frac{\partial^2 G}{\partial \hat{v}\hat{\lambda}}(\mathbf{0})\,\hat{v} + \mathcal{O}(\hat{w}, \hat{\lambda}) \\ &= -e^{1/2}\,a^{-1}\,b^{1/2}\,c^{-1}\,\frac{\partial^2 G}{\partial \hat{v}\hat{\lambda}}(\mathbf{0})\,\hat{v} + \mathcal{O}(\hat{w}, \hat{\lambda}), \end{aligned} \tag{46}$$

$$h_5(\hat{v}, \hat{w}, \hat{\lambda}) = d\,T + \mathcal{O}(\hat{v}, \hat{w}, \hat{\lambda}) = d\,e^{-1/2}\,b^{-1/2} + \mathcal{O}(\hat{v}, \hat{w}, \hat{\lambda}), \tag{47}$$

rearranging terms we get the canonical form

$$\begin{cases} \hat{v}' = -\hat{w} + \hat{v}^2 - \hat{w}\ h_1(\hat{v},\hat{w}) + \hat{v}^2\ h_2(\hat{v},\hat{w}), \\ \hat{w}' = \epsilon\ [\,\hat{v} - \hat{\lambda} + \hat{v}\ h_3(\hat{v},\hat{w},\hat{\lambda}) - \hat{\lambda}\ h_4(\hat{v},\hat{w},\hat{\lambda}) + \hat{w}\ h_5(\hat{v},\hat{w},\hat{\lambda})\,]. \end{cases} \tag{48}$$

Note that in (48) the sign $'$ represents $d/d\,\hat{t}$.

Appendix B. Calculation of the canard critical value. In [16] an expression for $\hat{\lambda_c}$ was found:

$$\hat{\lambda_c} = \frac{-a_1 + 3\ a_2 - 2\ a_3 + 2\ a_5}{8}\ \epsilon + \mathcal{O}(\epsilon^{3/2}), \tag{49}$$

where

$$a_1 = \frac{\partial h_1}{\partial \hat{v}}, \qquad a_2 = \frac{\partial h_2}{\partial \hat{v}}, \qquad a_3 = \frac{\partial h_4}{\partial \hat{v}}, \qquad a_5 = h_5. \tag{50}$$

Substituting (34), (35), and (43)–(47) into (50), we get

$$a_1 = -\frac{2\ G_v^{1/2}\ F_{vw}}{|F_w|^{1/2}\ F_{vv}}, \qquad a_2 = \frac{2\ G_v^{1/2}\ |F_w|^{1/2}\ F_{vvv}}{3\ F_{vv}^2}, \tag{51}$$

$$a_3 = \frac{|F_w|^{1/2}\ G_{vv}}{G_v^{1/2}\ F_{vv}}, \qquad a_5 = \frac{G_w}{G_v^{1/2}\ |F_w|^{1/2}}, \tag{52}$$

where all the functions are calculated at $\mathbf{0}$. The corresponding expression for $\lambda_c = L\,\hat{\lambda_c}$ is

$$\begin{aligned} \lambda_c(\sqrt{\epsilon}) &= -\frac{g_v^{3/2}\ |F_w|^{1/2}}{4\ F_{vv}\ |G_\lambda|}\ [-a_1 + 3\ a_2 - 2\ a_3 + 2\ a_5]\ \epsilon + \mathcal{O}(\epsilon^{3/2}) \\ &= -\frac{G_v}{2\,F_{vv}^3\,|G_\lambda|}[\,G_v\,F_{vw}\,F_{vv} + G_v\,|F_w|\,F_{vvv} \\ &\qquad\qquad - |F_w|\,G_{vv}\,F_{vv} + G_w\,F_{vv}^2\,]\ \epsilon + \mathcal{O}(\epsilon^{3/2}), \end{aligned} \tag{53}$$

where all the functions are calculated at $\mathbf{0}$.

If F were not independent of ϵ, then we would need to add a term $\epsilon\,h_6$ in (48). This would produce an additional $\mathcal{O}(\epsilon)$ term, proportional to dh_6/dv, in the expression for λ_c [15].

Appendix C. Equilibrium point and Hopf bifurcation. Here we present a result by Krupa and Szmolyan [15]. Based on the calculations from appendices A and B, we apply it to system (1) and the examples presented in section 2.

Consider system (1) with $\gamma = 0$ and (v_k, w_k) replaced by (v, w). Call

$$A = -a_1 + 3\,a_2 - 2\,a_3 - 2\,a_5. \tag{54}$$

Assume the following:

(i) The critical manifold $\{(v,w) : F(v,w) = 0\}$ can be written in the form $w = f(v)$, and the function f is cubic-like, i.e., it has precisely two critical points, one nondegenerate minimum and one nondegenerate maximum, each of which satisfies $\partial^2 F/\partial v^2(\mathbf{p}) \neq \mathbf{0}$ and $\partial F/\partial w(\mathbf{p}) \neq \mathbf{0}$. Without loss of generality, the minimum of f can be taken as $(0,0)$.

(ii) For $\epsilon = 0$ the left and right branches of the critical manifold $F(v, w) = 0$ are attracting and the central branch is repelling.

(iii) For $\lambda = 0$ the fold point $(0, 0)$ is a nondegenerate canard point; i.e., it satisfies $\partial G/\partial v(\mathbf{0}) \neq \mathbf{0}$ and $\partial G/\partial \lambda(\mathbf{0}) \neq \mathbf{0}$.

(iv) When $\lambda = 0$, $v' < 0$ for the slow flow on the right branch of f and $v' > 0$ for the slow flow on the central and left branches of f, including the point $(0, 0)$.

Then there exist $\epsilon_0 > 0$ and $\lambda_0 > 0$ such that, for each $0 < \epsilon < \epsilon_0$, $|\lambda| < \lambda_0$, system (1) with $\gamma = 0$ and (v_k, w_k) replaced by (v, w) has precisely one equilibrium point p_e in a neighborhood of the origin which converges to the canard point as $(\epsilon, \lambda) \to (0, 0)$. Moreover, there exists a curve

$$\lambda_H(\sqrt{\epsilon}) = -\frac{a_5}{2}\epsilon + \mathcal{O}(\epsilon^{3/2}) \tag{55}$$

such that p_e is stable (unstable) for $\lambda < \lambda_H$ ($\lambda > \lambda_H$). The equilibrium point p_e loses stability through a supercritical (subcritical) Hopf bifurcation if $A > 0$ ($A < 0$).

The proof is given in [15].

By substituting (51) and (52) into (54) we get

$$A = \frac{2}{|F_w|^{1/2}\,(G_v)^{1/2}\,(F_{vv})^2}\left[G_v\,F_{vv}\,F_{vw} + |F_w|\,G_v\,F_{vvv} - |F_w|\,G_{vv}\,F_{vv} - G_w\,(F_{vv})^2\right]. \tag{56}$$

Note that for the FHN equations (examples (ii) and (iii) in section 2) the condition for a supercritical Hopf bifurcation is equivalent to $3\,h\,g_v + a\,g_{vv} > 2\,a^2$. In particular, for the classical FHN equations (example (ii) in section 2) this becomes $3\,h\,\beta > 2\,a^2$.

By substituting the second equation in (52) into (55) we get

$$\lambda_H(\sqrt{\epsilon}) = -\frac{G_w}{2\,(G_v)^{1/2}\,|F_w|^{1/2}}\epsilon + \mathcal{O}(\epsilon^{3/2}). \tag{57}$$

Note that for the FHN equations $\lambda_H(\sqrt{\epsilon}) = 1\,/\,(G_v)^{1/2}\epsilon + \mathcal{O}(\epsilon^{3/2})$, and for the classical FHN equations $\lambda_H(\sqrt{\epsilon}) = 1\,/\beta^{1/2}\epsilon + \mathcal{O}(\epsilon^{3/2})$.

Acknowledgments. The authors thank Tasso Kaper and Kresimir Josic for helpful comments on the mathematical treatment of the canard phenomenon, Milos Dolnik for useful comments on the BZ reaction, and the canard group at Boston University for the collective learning experience on the mathematics of canards.

REFERENCES

[1] J. D. Murray, *Mathematical Biology*, Springer-Verlag, Berlin, 1989.
[2] I. R. Epstein and K. Showalter, *Nonlinear chemical dynamics: Oscillations, patterns and chaos*, J. Phys. Chem., 100 (1996), pp. 13132–13147.
[3] S. H. Strogatz, *Nonlinear Dynamics and Chaos*, Addison–Wesley, Reading, MA, 1994.
[4] V. K. Vanag, L. Yang, M. Dolnik, A. M. Zhabotinsky, and I. R. Epstein, *Oscillatory cluster patterns in a homogeneous chemical system with global feedback*, Nature, 406 (2000), pp. 389–391.
[5] V. K. Vanag, A. M. Zhabotinsky, and I. R. Epstein, *Pattern formation in the Belousov-Zhabotinsky reaction with photochemical global feedback*, J. Phys. Chem., 104A (2000), pp. 11566–11577.
[6] L. Yang, M. Dolnik, A. M. Zhabotinsky, and I. R. Epstein, *Oscillatory clusters in a model of the photosensitive Belousov-Zhabotinsky reaction system with global feedback*, Phys. Rev. E, 62 (2000), pp. 6414–6420.

[7] A. M. Zhabotinsky, F. Buchholtz, A. B. Kiyatkin, and I. R. Epstein, *Oscillations and waves in metal-ion-catalyzed bromate oscillating reactions in highly oxidized states*, J. Phys. Chem., 97 (1993), pp. 7578–7584.
[8] R. Kuske and T. Erneux, *Localized synchronization of two coupled solid state lasers*, Optics Communications, 139 (1997), pp. 125–131.
[9] R. Kuske and T. Erneux, *Bifurcation to localized oscillations*, European J. Appl. Math., 8 (1997), pp. 389–402.
[10] M. Boukalouch, J. Elezgaray, A. Arneodo, J. Boissonade, and P. De Kepper, *Oscillatory instability induced by mass interchange between two coupled steady-state reactors*, J. Phys. Chem., 91 (1987), pp. 5843–5845.
[11] F. Dumortier and R. Roussarie, *Canard Cycles and Center Manifolds*, Mem. Amer. Math. Soc., 121 (1996).
[12] W. Eckhaus, *Relaxation oscillations including a standard chase on French ducks*, in Asymptotic Analysis, II, Lecture Notes in Math. 985, Springer-Verlag, Berlin, 1983, pp. 449–497.
[13] S. M. Baer and T. Erneux, *Singular Hopf bifurcation to relaxation oscillations*, SIAM J. Appl. Math., 46 (1986), pp. 721–739.
[14] E. Benoit, J. L. Callot, F. Dienner, and M. Dienner, *Chasse au Canard*, IRMA, Strasbourg, France, 1980.
[15] M. Krupa and P. Szmolyan, *Relaxation oscillation and canard explosion*, J. Differential Equations, 174 (2001), pp. 312–368.
[16] M. Krupa and P. Szmolyan, *Extending geometric singular perturbation theory to nonhyperbolic points—fold and canard points in two dimensions*, SIAM J. Math. Anal., 33 (2001), pp. 286–314.
[17] M. Brons and K. Bar-Eli, *Canard explosion and excitation in a model of the BZ reaction*, J. Phys. Chem., 95 (1991), pp. 8706–8713.
[18] F. Buchholtz, M. Dolnik, and I. R. Epstein, *Diffusion-induced instabilities near a canard*, J. Phys. Chem., 99 (1995), pp. 15093–15101.
[19] H. G. Rotstein, N. Kopell, A. Zhabotinsky, and I. R. Epstein, *Canard phenomenon and localization of oscillations in the Belousov-Zhabotinsky reaction with global feedback*, submitted.
[20] H. G. Rotstein and R. Kuske, *personal communication*.
[21] H. G. Rotstein, R. Kuske, and N. Kopell, *Localized Oscillations in a Coupled Two-Pool Model. A Canard Mechanism*, in preparation.
[22] R. L. Burden and J. D. Faires, *Numerical Analysis*, PWS Publishing, Boston, 1980.
[23] R. FitzHugh, *Impulses and physiological states in models of nerve membrane*, Biophys. J., 1 (1961), pp. 445–466.
[24] F. C. Hoppensteadt, *An Introduction to the Mathematics of Neurons*, Cambridge University Press, New York, 1996.
[25] E. F. Mishchenko and N. Kh. Rozov, *Differential Equations with Small Parameters and Relaxation Oscillations*, Plenum Press, New York, London, 1980.
[26] J. Grasman, *Asymptotic Methods for Relaxation Oscillations and Applications*, Springer-Verlag, New York, 1986.
[27] F. Dumortier, *Techniques in the theory of local bifurcations: Blow-up, normal forms, nilpotent bifurcations, singular perturbations*, in Bifurcations and Periodic Orbits of Vector Fields, D. Schlomiuk, ed., Kluwer Academic Press, Dordrecht, The Netherlands, 1993, pp. 19–73.
[28] V. K. Vanag, A. M. Zhabotinsky, and I. R. Epstein, *Oscillatory clusters in the periodically illuminated, spatially extended Belousov-Zhabotinsky reaction*, Phys. Rev. Lett., 86 (2001), pp. 552–555.
[29] M. Krupa, W. F. Langford, and J. P. Voroney, *Canard Explosion in the Oregonator*, in preparation.

SIAM J. APPL. MATH.
Vol. 63, No. 6, pp. 2020–2041

LOW MACH NUMBER FLOWS IN TIME-DEPENDENT DOMAINS*

G. ALÌ†

Abstract. We perform a multiple time scale, single space scale analysis of a compressible fluid in a time-dependent domain, when the time variations of the boundary are small with respect to the acoustic velocity. We introduce an average operator with respect to the fast time. The averaged leading order variables satisfy modified incompressible equations, which are coupled to linear acoustic equations with respect to the fast time. We discuss possible initial-boundary data for the asymptotic equations inherited from the initial-boundary data for the compressible equations.

Key words. asymptotic analysis, Euler equations, low Mach number

AMS subject classifications. 35L65, 35C20, 35B40

DOI. 10.1137/S0036139902400738

1. Introduction.

A most remarkable property of the compressible fluid equations is, roughly speaking, their ability to force a solution to remain approximately incompressible when the initial data are approximately incompressible. This is a singular limit result, since the compressible and the incompressible equations are related through a passage to infinite of a characteristic speed. The relationship between compressible and incompressible flows was rigorously stated and proved for isentropic flows in unbounded domains by Klainerman and Majda at the beginning of the eighties [13, 14] and later extended to nonisentropic flows in bounded domains by Schochet [27]. In all these papers, the initial data are "prepared," that is, they are compatible with the limiting incompressible equations. When the data are not prepared, the limit is still valid but it is not uniform for times close to zero, since the initial time derivative fails to be uniformly bounded (cf. [1, 7, 8, 9, 33] and, recently, [25]).

Despite its mathematical subtleties, the physical mechanism of this singular limit is very simple. The smallness of the Mach number constrains the spatial variation of the pressure to stay small at order zero for consistency with the conservation of momentum. Then, the energy equation, written in terms of the pressure, forces the divergence of the fluid velocity to stay close to zero. Finally, if no compression or expansion over the boundary of the domain takes place, the initial density configuration is advected by the flow.

The low Mach number theory plays an important role in many applications, such as in combustion, both in confined and unbounded domains [18, 21], and in astrophysics, in the framework of magnetohydrodynamics, in order to justify the small density fluctuations, called pseudosound, in the solar wind incompressible model [22, 34, 35]. In the context of these physical applications, it is very important to develop numerical schemes that are uniformly valid as the Mach number goes to zero. In the one-dimensional case, by using a multiple scale asymptotic analysis, Klein was able to construct a numerical scheme for compressible flows which retains its validity also at low Mach number regimes [15] (see also the extensive report [16], where, in particular, the issue of asymptotic adaptivity of this scheme with respect to the Mach number is addressed). These results were later extended to the multidimensional case

*Received by the editors January 9, 2002; accepted for publication (in revised form) January 29, 2003; published electronically September 4, 2003.
http://www.siam.org/journals/siap/63-6/40073.html
†Istituto per le Applicazioni del Calcolo "M. Picone," Consiglio Nazionale delle Ricerche, Napoli I-80131, Italy (g.ali@iac.cnr.it).

in [31]. The main point of Klein's asymptotic analysis, later assessed within a systematic mathematical framework in [23] (for unbounded domains), is the introduction of two different length scales corresponding to the large scale motion of the fluid and to small acoustic perturbations. In fact, a simple asymptotic analysis shows that highly oscillating acoustic perturbations show up at the same order as pressure variations become significant [18].

In this paper, we perform a similar analysis, when only a length scale is relevant but two time scales need to be considered. A typical example of this occurrence is a fluid confined in a bounded time-dependent domain, when the time variation of the boundary (the "boundary velocity") is small compared to the acoustic velocity of the fluid. We introduce a slow time t, related to the boundary velocity and therefore to the large scale motion of the fluid, and a fast time $\tau = t/\epsilon$, related to the acoustic speed. The parameter ϵ is proportional to the Mach number. This multiple scale analysis shows that the limiting incompressible variables can be interpreted as averages of the compressible variables with respect to τ. Moreover, they are coupled to linear acoustic equations with respect to the fast time τ. More precisely, we show that at leading order the motion is adiabatic, that is, the leading order pressure $p(t)$ is related to the volume $V(t)$ of the domain by the law $p = cV^{-\gamma}$. Also, the leading order density ρ does not depend on the fast time, and the leading order velocity $\mathbf{u}$ can be split into the averaged part with respect to the fast time and the average-free part, $\mathbf{u} = \bar{\mathbf{u}} + \delta\mathbf{u}$. These variables satisfy the following equations:

$$
\begin{aligned}
&\frac{1}{\rho}(\partial_t + \bar{\mathbf{u}}\cdot\nabla)\rho = -\frac{1}{V}\frac{dV}{dt},\\
&\rho(\partial_t + \bar{\mathbf{u}}\cdot\nabla)\bar{\mathbf{u}} + \nabla\pi = -\nabla\cdot\left(\rho\left\langle\delta\mathbf{u}\otimes\delta\mathbf{u}\right\rangle\right),\\
&\nabla\cdot\bar{\mathbf{u}} = \frac{1}{V}\frac{dV}{dt},
\end{aligned}
\tag{1.1}
$$

where the curled brackets denote average with respect to τ. The "incompressible" pressure π coincides with the averaged second order pressure. Finally, the average-free leading order velocity is coupled to the first order average-free pressure $\delta p^{(1)}$ by the linear acoustic system

$$
\begin{aligned}
&\partial_\tau\delta\mathbf{u} + \frac{1}{\rho}\nabla\delta p^{(1)} = 0,\\
&\partial_\tau\delta p^{(1)} + \gamma p\nabla\cdot\delta\mathbf{u} = 0.
\end{aligned}
\tag{1.2}
$$

The limit equations (1.1) and the acoustic equations (1.2) decouple if no pressure fluctuations appear at order ϵ. In this case, $\delta p^{(1)}$ vanishes as well as $\delta\mathbf{u}$, and (1.1) reduces to the nearly incompressible equations (using the terminology of [22, 34, 35]). The same theory applies if the domain is not moving for compressible flows with typical speed small compared to the acoustic velocity. In this case, we have $dV/dt = 0$ and the system (1.1) reduces to the incompressible system describing a stratified fluid, as found in [27]. An analogous theory can be developed for unbounded domains, obtaining the same set of equations (1.1), (1.2), with $dV/dt = 0$.

The acoustic system (1.2) deserves further comments. Thus far, the problem of high-frequency acoustics has only been addressed for unbounded domains [2, 3, 4, 5, 6, 10, 11, 12, 19, 20, 26, 28, 29], while the moving boundary problem was considered satisfactorily only in a single space dimension (the so-called piston problem) [17, 30, 32]. Here, we address the problem of high-frequency acoustics in more than one space dimension *and* in a closed, variable boundary domain.

The plan of the paper is the following. In section 2, we set the problem and introduce the fundamental equations and initial-boundary conditions. In section 3, we perform a preliminary single scale analysis, following [23]. The details of our asymptotic multiscale analysis are expounded in section 4. In the following two sections, we discuss the initial-boundary conditions for the asymptotic equations inherited from the initial-boundary data for the original compressible equations. In general, it is not possible to assign uniquely initial data for the acoustic equations with respect to the fast time τ. Anyway, at the end of section 6, we are able to produce a simple class of motion for which this problem can be solved. Finally, in section 7, we give an outlook of our results and draw some conclusions. The paper is supplemented by an appendix on weakly nonisentropic flows, that is, flows with constant entropy at order zero, which are produced by constant initial density distributions.

2. The compressible fluid equations. The Euler equations for a compressible fluid are

$$\begin{aligned} &\partial_t \rho + \nabla \cdot (\rho \mathbf{u}) = 0, \\ &\partial_t (\rho \mathbf{u}) + \nabla \cdot (\rho \mathbf{u} \otimes \mathbf{u} + pI) = 0, \\ &\partial_t E + \nabla \cdot [(E+p)\mathbf{u}] = 0, \end{aligned} \tag{2.1}$$

where $E = \rho\left(e + \frac{1}{2}|\mathbf{u}|^2\right)$ is the total energy. The internal energy, e, the pressure, p, and the density, ρ, are related to the entropy, s, and the temperature, T, by

$$de + pd\left(\frac{1}{\rho}\right) = Tds. \tag{2.2}$$

It is useful to write equations (2.1) in nonconservative form:

$$\begin{aligned} &(\partial_t + \mathbf{u} \cdot \nabla)\rho = -\rho \nabla \cdot \mathbf{u}, \\ &\rho(\partial_t + \mathbf{u} \cdot \nabla)\mathbf{u} = -\nabla p, \\ &\rho(\partial_t + \mathbf{u} \cdot \nabla)e = -p \nabla \cdot \mathbf{u}. \end{aligned} \tag{2.3}$$

Equations (2.3) and (2.2) immediately yield the entropy equation

$$(\partial_t + \mathbf{u} \cdot \nabla)s = 0 \tag{2.4}$$

and the identity

$$(\partial_t + \mathbf{u} \cdot \nabla)p = -\rho \left(\left.\frac{\partial p}{\partial \rho}\right|_e + \frac{p}{\rho^2} \left.\frac{\partial p}{\partial e}\right|_\rho \right) \nabla \cdot \mathbf{u}. \equiv -\rho \left.\frac{\partial p}{\partial \rho}\right|_s \nabla \cdot \mathbf{u}. \tag{2.5}$$

In the following, we will assume the existence of an equation of state of the form $s = s(\rho, p)$ and regard (2.1) as a hyperbolic system of partial differential equations for the variables $(\rho, \mathbf{u}, p)$. Using (2.5), we can write the pressure equation

$$\partial_t p + \mathbf{u} \cdot \nabla p + \gamma p \nabla \cdot \mathbf{u} = 0, \tag{2.6}$$

where the adiabatic exponent γ is defined by [24]

$$\gamma = \frac{\rho}{p} \left.\frac{\partial p}{\partial \rho}\right|_s \equiv \frac{\rho}{p} \frac{\partial s / \partial \rho}{\partial s / \partial p}. \tag{2.7}$$

For a perfect fluid satisfying the relation $p = R\rho T$, the adiabatic coefficient is a function of temperature only, that is, γ depends on the ratio $\gamma(p/\rho)$ only. In particular, γ is constant for a polytropic gas. In this case, the equation of state is explicitly given by

$$s = \log\left(p\rho^{-\gamma}\right). \tag{2.8}$$

We consider the equations (2.1), (2.2) on a bounded, time-dependent domain $\Omega_t \in \mathbb{R}^n$. We denote by Ω_0 the domain at the initial time $t = 0$ and describe its evolution by a family of invertible maps $\Phi_t : \mathbb{R}^n \to \mathbb{R}^n$, depending continuously on the time t such that

$$\Omega_t = \Phi_t(\Omega_0) \qquad \text{for all } t.$$

This severe assumption on the domain Ω_t is, nevertheless, general enough to include a moving rigid domain, or a cylinder cut by a fixed surface and a moving surface (piston problem), or a contracting-expanding sphere (star).

In the case of a moving rigid domain, we can choose Φ_t as

$$\Phi_t(\mathbf{x}) = \mathbf{x} + \mathbf{c}(t). \tag{2.9}$$

In the piston model, using the notation $\mathbf{y} = (x^1, x^2, \dots, x^{n-1})$ and $z = x^n$, the time-dependent domain can be written as

$$\Omega_t = \{(\mathbf{y}, z) \in \mathbb{R}^n \,:\, \mathbf{y} \in \Sigma,\ \alpha(\mathbf{y}) \le z \le \beta_t(\mathbf{y})\}, \tag{2.10}$$

where Σ is a bounded set of $\mathbb{R}^{n-1}$, α, β_t are functions defined in Σ, with $\min_\Sigma\{\alpha - \beta_t\} < 0$ for all $t \ge 0$. The piston deformation can be described by

$$\Phi_t(\mathbf{y}, z) = \left(\mathbf{y}, \alpha(\mathbf{y}) + \frac{\beta_t(\mathbf{y}) - \alpha(\mathbf{y})}{\beta_0(\mathbf{y}) - \alpha(\mathbf{y})}(z - \alpha(\mathbf{y}))\right). \tag{2.11}$$

In the star model, we have

$$\Omega_t = \{\mathbf{x} \in \mathbb{R}^3 \,:\, |\mathbf{x}| \le \eta(\hat{\mathbf{x}}, t)\}, \tag{2.12}$$

with $\hat{\mathbf{x}} = \frac{\mathbf{x}}{|\mathbf{x}|}$ and $\eta(\hat{\mathbf{x}}, t) > 0$, $\eta(\hat{\mathbf{x}}, 0) = \eta_0(\hat{\mathbf{x}})$. We can choose

$$\Phi_t(\mathbf{x}) = \begin{cases} \dfrac{\eta(\hat{\mathbf{x}}, t)}{\eta_0(\hat{\mathbf{x}})}\mathbf{x} & \text{if } \mathbf{x} \neq 0, \\ 0 & \text{if } \mathbf{x} = 0. \end{cases} \tag{2.13}$$

The map Φ_t has a geometric meaning and is related neither to the fluid motion nor to the Lagrangian variables. Moreover, Φ_t does not need to be globally unique, since only its restriction to a neighborhood of the boundary $\partial\Omega_0$ characterizes the motion of the domain's boundary $\partial\Omega_t$. In particular, we can use the map Φ_t to define the velocity $\mathbf{u}_\Omega$ of the boundary by

$$\mathbf{u}_\Omega(\mathbf{x}) = \frac{\partial \Phi_t}{\partial t}(\Phi_t^{-1}(\mathbf{x})) \qquad \text{for all } \mathbf{x} \in \partial\Omega_t. \tag{2.14}$$

The aim of this paper is to describe the fluid motion when the velocity of the boundary is small compared to the sound speed. In particular, this requirement is always valid when the boundary is fixed.

To nondimensionalize the equations (2.1), (2.6), we consider reference values for the density, the pressure, the length scale, and the time scale, ρ_{ref}, p_{ref}, l_{ref}, and t_{ref}, respectively. We introduce the rescaled variables

$$\begin{aligned} &\rho' = \frac{\rho}{\rho_{\text{ref}}}, \quad \mathbf{u}' = \frac{\mathbf{u}}{u_{\text{ref}}}, \quad p' = \frac{p}{p_{\text{ref}}}, \\ &\mathbf{x}' = \frac{\mathbf{x}}{l_{\text{ref}}}, \quad t' = \frac{t}{t_{\text{ref}}}, \quad u_{\text{ref}} = \frac{l_{\text{ref}}}{t_{\text{ref}}}. \end{aligned} \tag{2.15}$$

We choose u_{ref} as a typical value of the boundary speed. Some care is required in selecting the "correct" typical velocity. In [32], several choices are checked numerically for a unidimensional piston model. For this simple model, the conclusion is that the most appropriate typical speed is the maximum speed of the piston.

Substituting (2.15) into (2.1), (2.6), we obtain

$$\begin{aligned} &\partial_{t'}\rho' + \nabla' \cdot (\rho' \mathbf{u}') = 0, \\ &\partial_{t'}(\rho' \mathbf{u}') + \nabla' \cdot (\rho' \mathbf{u}' \otimes \mathbf{u}') + \frac{1}{\epsilon^2}\nabla' p' = 0, \\ &\partial_{t'} p' + (\mathbf{u}' \cdot \nabla') p' + \gamma' p' \nabla' \cdot \mathbf{u}' = 0, \end{aligned} \tag{2.16}$$

with $\gamma'(\rho', p') = \gamma(\rho_{\text{ref}}\rho', p_{\text{ref}}p')$, and

$$\epsilon^2 = \frac{u_{\text{ref}}^2}{p_{\text{ref}}/\rho_{\text{ref}}}. \tag{2.17}$$

The parameter ϵ is proportional to the square of the Mach number $M = u_{\text{ref}}/c_{\text{ref}}$, where c_{ref} is the typical sound speed:

$$c_{\text{ref}}^2 = \left.\frac{\partial p}{\partial \rho}\right|_s (\rho_{\text{ref}}, p_{\text{ref}}) \equiv \gamma(\rho_{\text{ref}}, p_{\text{ref}}) \frac{p_{\text{ref}}}{\rho_{\text{ref}}}.$$

Equation (2.16) is defined in the domain $\Omega'_{t'} = \frac{1}{l_{\text{ref}}}\Omega_{t_{\text{ref}}t'}$. We rescale the map Φ_t accordingly as

$$\Phi'_{t'}(\mathbf{x}') = \frac{1}{l_{\text{ref}}} \Phi_{t_{\text{ref}}t'}(l_{\text{ref}}\mathbf{x}'). \tag{2.18}$$

Then we have $\Omega'_{t'} = \Phi'_{t'}(\Omega'_0)$ and

$$\mathbf{u}'_\Omega = \frac{\mathbf{u}_\Omega}{u_{\text{ref}}}. \tag{2.19}$$

It is convenient to drop the primes and rewrite (2.16) as

$$\begin{aligned} &\partial_t \rho + \nabla \cdot \mathbf{m} = 0, \\ &\partial_t \mathbf{m} + \nabla \cdot (\mathbf{m} \otimes \mathbf{u}) + \frac{1}{\epsilon^2}\nabla p = 0, \\ &\frac{1}{\gamma p}(\partial_t + \mathbf{u} \cdot \nabla) p + \nabla \cdot \mathbf{u} = 0, \end{aligned} \tag{2.20}$$

with $\mathbf{m} = \rho\mathbf{u}$.

In conclusion, we consider the system (2.20) on the domain Ω_t with initial data

$$\rho(\mathbf{x}, 0) = \rho_0(\mathbf{x}), \quad \mathbf{u}(\mathbf{x}, 0) = \mathbf{u}_0(\mathbf{x}), \quad p(\mathbf{x}, 0) = p_0(\mathbf{x}) \qquad \text{in} \quad \Omega_0, \tag{2.21}$$

and boundary conditions

$$\mathbf{u}\cdot\mathbf{n}_t = \mathbf{u}_\Omega\cdot\mathbf{n}_t \qquad \text{on} \quad \partial\Omega_t. \tag{2.22}$$

Here, $\mathbf{n}_t$ denotes the outward normal on the boundary $\partial\Omega_t$.

In principle, it is possible to consider more general boundary conditions, by expanding (2.21) as a power series of ϵ. Also, it is possible to consider more general boundary conditions by assuming a given flux rate through the boundary. For simplicity, we will not consider such generalizations in this paper. Moreover, we will restrict our study to a polytropic fluid so that γ is a constant and the equation of state is given by (2.8).

The limit of (2.20) as ϵ tends to zero is a singular limit. There are many ways to approach the study of the limiting behavior of the compressible solutions. Here, we will follow a multiple scale asymptotic approach. Anyway, it is instructive to perform a preliminary single scale asymptotic analysis of this limit. In the following section, we deduce some general properties, which can be recovered just from the boundedness of Ω_t and from the boundary condition (2.22) using a single scale expansion. A detailed account of the appropriate multiple scale asymptotics will be given in the subsequent section 4.

3. Single scale asymptotics. In order to study the basic mechanism of the singular limit in (2.20), we assume that for a polytropic fluid a solution of (2.20) admits a single scale expansion of the form

$$\begin{aligned}
\rho(\mathbf{x},t) &= \rho^{(0)}(\mathbf{x},t) + \epsilon\rho^{(1)}(\mathbf{x},t) + O(\epsilon^2),\\
\mathbf{u}(\mathbf{x},t) &= \mathbf{u}^{(0)}(\mathbf{x},t) + \epsilon\mathbf{u}^{(1)}(\mathbf{x},t) + O(\epsilon^2),\\
p(\mathbf{x},t) &= p^{(0)}(\mathbf{x},t) + \epsilon p^{(1)}(\mathbf{x},t) + \epsilon^2 p^{(2)}(\mathbf{x},t) + O(\epsilon^3).
\end{aligned} \tag{3.1}$$

By using this expansion in (2.20) and equating to zero the coefficient of each power of ϵ, we obtain asymptotic equations for the functions of the expansion at each order. Explicitly, up to order ϵ, we obtain

$$\partial_t\rho^{(0)} + \nabla\cdot\mathbf{m}^{(0)} = 0, \tag{3.2}$$

$$\partial_t\rho^{(1)} + \nabla\cdot\mathbf{m}^{(1)} = 0, \tag{3.3}$$

$$\nabla p^{(0)} = 0, \tag{3.4}$$

$$\nabla p^{(1)} = 0, \tag{3.5}$$

$$\partial_t\mathbf{m}^{(0)} + \nabla\cdot\left(\mathbf{m}^{(0)}\otimes\mathbf{u}^{(0)}\right) + \nabla p^{(2)} = 0, \tag{3.6}$$

$$\frac{1}{\gamma p^{(0)}}\partial_t p^{(0)} + \nabla\cdot\mathbf{u}^{(0)} = 0, \tag{3.7}$$

$$\partial_t\left(\frac{p^{(1)}}{\gamma p^{(0)}}\right) + \nabla\cdot\mathbf{u}^{(1)} = 0. \tag{3.8}$$

Here, we have $\mathbf{m}^{(0)} = \rho^{(0)}\mathbf{u}^{(0)}$, $\mathbf{m}^{(1)} = \rho^{(0)}\mathbf{u}^{(1)} + \rho^{(1)}\mathbf{u}^{(0)}$.

Using (3.4) and (3.5), we can state the following result.

THEOREM 3.1. *If $(\rho,\mathbf{u},p)$ is a solution to* (2.20) *which admits the asymptotic expansion* (3.1)*, then, at the first two orders in ϵ, the pressure p is a function of time:*

$$p = p^{(0)}(t) + \epsilon p^{(1)}(t) + O(\epsilon^2). \tag{3.9}$$

Using the boundary condition (2.22), we can recover the functions $p^{(0)}(t)$ and $p^{(1)}(t)$ from (3.7) and (3.8).

THEOREM 3.2. *If $(\rho, \mathbf{u}, p)$ is a solution to* (2.20), (2.22) *which admits the asymptotic expansion* (3.1), *then $p^{(0)}(t)$ and $p^{(1)}(t)$ are given by*

$$p^{(0)} = C_0 |\Omega_t|^{-\gamma}, \tag{3.10}$$

$$p^{(1)} = C_1 p^{(0)}, \tag{3.11}$$

where $|\Omega_t|$ is the measure of the domain Ω_t and C_0, C_1 are constants.

Proof. We integrate (3.7) and (3.8) over the domain Ω_t. Using Green's formula and the boundary condition (2.22), we find

$$\frac{1}{\gamma p^{(0)}} \frac{dp^{(0)}}{dt} = -\frac{\int_{\partial\Omega_t} \mathbf{u}^{(0)} \cdot \mathbf{n}_t \, dS}{\int_{\Omega_t} dV} = -\frac{1}{|\Omega_t|} \int_{\partial\Omega_t} \mathbf{u}_\Omega \cdot \mathbf{n}_t \, dS, \tag{3.12}$$

$$\frac{d}{dt}\left(\frac{p^{(1)}}{\gamma p^{(0)}}\right) = -\frac{\int_{\partial\Omega_t} \mathbf{u}^{(1)} \cdot \mathbf{n}_t \, dS}{\int_{\Omega_t} dV} = 0. \tag{3.13}$$

The second equation immediately gives (3.11). To evaluate the right-hand side of the first equation, we observe that $\mathbf{u}_\Omega$ can be extended to a differentiable function defined in Ω_t by

$$\mathbf{u}_\Omega(\mathbf{x}) = \partial_t \Phi_t(\Phi_t^{-1}(\mathbf{x})) \qquad \text{for all } \mathbf{x} \in \Omega_t. \tag{3.14}$$

Then we have

$$\begin{aligned}\int_{\partial\Omega_t} \mathbf{u}_\Omega \cdot \mathbf{n}_t \, dS &= \int_{\Omega_t} \nabla \cdot \mathbf{u}_\Omega \, dV = \int_{\Omega_0} \nabla \cdot \frac{\partial \Phi_t}{\partial t} \left|\frac{\partial \Phi_t}{\partial \mathbf{x}_0}\right| dV_0 \\ &= \int_{\Omega_0} \frac{\partial}{\partial t} \left|\frac{\partial \Phi_t}{\partial \mathbf{x}_0}\right| dV_0 = \frac{d}{dt} \int_{\Omega_0} \left|\frac{\partial \Phi_t}{\partial \mathbf{x}_0}\right| dV_0 = \frac{d|\Omega_t|}{dt}.\end{aligned}$$

Here we have used the well-known identity

$$\frac{\partial}{\partial t}\left|\frac{\partial \Phi_t}{\partial \mathbf{x}_0}\right| = \nabla \cdot \frac{\partial \Phi_t}{\partial t}\left|\frac{\partial \Phi_t}{\partial \mathbf{x}_0}\right|.$$

In conclusion, using this result in (3.12), we arrive at the equation

$$\frac{1}{\gamma p^{(0)}} \frac{dp^{(0)}}{dt} = -\frac{1}{|\Omega_t|} \frac{d|\Omega_t|}{dt}, \tag{3.15}$$

which immediately implies (3.10). ∎

As a consequence of Theorem 3.2, since the function $p^{(0)}$ is known in terms of the measure of the domain, (3.7) can be regarded as a constraint for the leading order velocity. In particular, if the domain is fixed, (3.7) reduces to the incompressibility condition.

The dynamics of $\rho^{(0)}$ and $\mathbf{u}^{(0)}$ is ruled by (3.2) and (3.6). Writing them in terms of density, pressure, and velocity, and using Theorem 3.2, we obtain the following result.

THEOREM 3.3. *Under the same assumptions of Theorem* 3.2*, the leading order density and velocity* $(\rho^{(0)}, \mathbf{u}^{(0)})$ *satisfy the system*

$$\begin{aligned} &\frac{1}{\rho^{(0)}}\frac{D\rho^{(0)}}{Dt} = -\frac{1}{|\Omega_t|}\frac{d|\Omega_t|}{dt}, \\ &\rho^{(0)}\frac{D\mathbf{u}^{(0)}}{Dt} + \nabla p^{(2)} = 0, \\ &\nabla\cdot\mathbf{u}^{(0)} = \frac{1}{|\Omega_t|}\frac{d|\Omega_t|}{dt}, \end{aligned} \tag{3.16}$$

where

$$\frac{D}{Dt} = \partial_t + \mathbf{u}^{(0)}\cdot\nabla.$$

From Theorem 3.3, obvious restrictions on the initial data in Ω_0 are

$$\mathbf{u}_0(\mathbf{x}) = \mathbf{u}_0^{(0)}(\mathbf{x}) + O(\epsilon), \qquad p_0(\mathbf{x}) = \text{constant} + O(\epsilon^2), \tag{3.17}$$

with the constraint

$$\nabla\cdot\mathbf{u}_0^{(0)} = \text{constant}. \tag{3.18}$$

Therefore, we expect new effects, which are not described by the single scale theory, to arise when the restriction (3.18) does not hold. In the following section we develop a multiscale analysis which includes an additional time scale related to the acoustic speed.

4. A multiscale analysis. Two time scales arise naturally in the zero Mach number limit of a compressible flow in a bounded domain. They are related to the typical speed of the flow under consideration, u_{ref}, and to the typical acoustic speed, c_{ref}. More precisely, since the domain is bounded, the reference length scale, l_{ref}, can be chosen as the typical size of the domain. Then, we can define the time scales $t_{\text{ref}} = l_{\text{ref}}/u_{\text{ref}}$ and $t_{\text{s}} = l_{\text{ref}}/c_{\text{ref}}$. We recall that the small parameter ϵ is proportional to the ratio between u_{ref} and c_{ref}, that is, between t_{s} and t_{ref}.

In (2.15), we have used t_{ref} to rescale the time according to the relation $t' = t/t_{\text{ref}}$ (here, t is the unscaled time). Since u_{ref} is related to the motion of the domain's boundary (for instance, the piston speed), the rescaled time, t_{ref}, is related to the large scale motion of the fluid (for instance, to flows with speed sufficiently close to the domain's boundary velocity). To describe smaller fluctuations of the flow, due to the propagation of acoustic wave we need to introduce a dependence also on the other time which appears naturally, that is, $t/t_{\text{s}} \sim t'/\epsilon$.

With this motivation we want to express the solution $(\rho, \mathbf{u}, p)$ of (2.20), with initial data (2.21) and boundary condition (2.22), as a series in the powers of ϵ:

$$\begin{aligned} \rho(\mathbf{x},t;\epsilon) &= \rho^{(0)}(\mathbf{x},t,\tau) + \epsilon\rho^{(1)}(\mathbf{x},t,\tau) + \epsilon^2\rho^{(2)}(\mathbf{x},t,\tau) + \cdots, \\ \mathbf{u}(\mathbf{x},t;\epsilon) &= \mathbf{u}^{(0)}(\mathbf{x},t,\tau) + \epsilon\mathbf{u}^{(1)}(\mathbf{x},t,\tau) + \epsilon^2\mathbf{u}^{(2)}(\mathbf{x},t,\tau) + \cdots, \\ p(\mathbf{x},t;\epsilon) &= p^{(0)}(\mathbf{x},t,\tau) + \epsilon p^{(1)}(\mathbf{x},t,\tau) + \epsilon^2 p^{(2)}(\mathbf{x},t,\tau) + \cdots \end{aligned} \tag{4.1}$$

with $\tau = \frac{t}{\epsilon}$. We assume that the expansion (4.1) is uniformly valid in time.

THEOREM 4.1. *The uniform validity of the series* (4.1) *with respect to time for all positive times is equivalent to the sublinearity conditions*

$$\lim_{\tau\to\infty}\frac{1}{\tau}(\rho^{(i)}, \mathbf{u}^{(i)}, p^{(i)})(\mathbf{x},t,\tau) = 0, \qquad i = 0,1,2,\ldots. \tag{4.2}$$

Proof. It suffices to observe that

$$\lim_{\epsilon\to 0}\epsilon w(\mathbf{x},t,\tau)=t\lim_{\epsilon\to 0}\frac{1}{t/\epsilon}w(\mathbf{x},t,t/\epsilon)=t\lim_{\tau\to\infty}\frac{1}{\tau}w(\mathbf{x},t,\tau)$$

for any appropriate function $w(\mathbf{x},t,\tau)$ for all $t>0$. ∎

Recalling the sublinearity condition (4.2), it is natural to introduce for any function $w(\mathbf{x},t,\tau)$ the fast time average

$$\langle w\rangle(\mathbf{x},t)=\lim_{\tau\to\infty}\frac{1}{\tau}\int_0^\tau w(\mathbf{x},t,s)ds.$$

If this limit exists and it is finite, we can write $w(\mathbf{x},t,\tau)=\bar{w}(\mathbf{x},t)+\delta w(\mathbf{x},t,\tau)$, with $\bar{w}=\langle w\rangle$ and $\langle\delta w\rangle=0$. Plugging (4.1) into (2.20) and equating to zero the coefficient of each power of ϵ, we find

$$\partial_\tau\rho^{(0)}=0, \tag{4.3}$$

$$\partial_\tau\rho^{(1)}+\partial_t\rho^{(0)}+\nabla\cdot\mathbf{m}^{(0)}=0, \tag{4.4}$$

$$\partial_\tau\rho^{(2)}+\partial_t\rho^{(1)}+\nabla\cdot\mathbf{m}^{(1)}=0, \tag{4.5}$$

$$\nabla p^{(0)}=0, \tag{4.6}$$

$$\partial_\tau\mathbf{m}^{(0)}+\nabla p^{(1)}=0, \tag{4.7}$$

$$\partial_\tau\mathbf{m}^{(1)}+\partial_t\mathbf{m}^{(0)}+\nabla\cdot(\mathbf{m}^{(0)}\otimes\mathbf{u}^{(0)})+\nabla p^{(2)}=0, \tag{4.8}$$

$$\partial_\tau\mathbf{m}^{(2)}+\partial_t\mathbf{m}^{(1)}+\nabla\cdot(\mathbf{m}^{(0)}\otimes\mathbf{u}^{(1)}+\mathbf{m}^{(1)}\otimes\mathbf{u}^{(0)})+\nabla p^{(3)}=0, \tag{4.9}$$

$$\partial_\tau p^{(0)}=0, \tag{4.10}$$

$$\frac{1}{\gamma p^{(0)}}\partial_\tau p^{(1)}+\frac{1}{\gamma p^{(0)}}\partial_t dp^{(0)}+\nabla\cdot\mathbf{u}^{(0)}=0, \tag{4.11}$$

$$\frac{1}{\gamma p^{(0)}}\partial_\tau p^{(2)}+(\partial_t+\mathbf{u}^{(0)}\cdot\nabla)\left(\frac{p^{(1)}}{\gamma p^{(0)}}\right)+\nabla\cdot\mathbf{u}^{(1)}=0. \tag{4.12}$$

Averaging (4.7) with respect to τ and using the sublinearity condition (4.2), we obtain

$$\nabla\bar{p}^{(1)}=0. \tag{4.13}$$

The following result is a simple consequence of (4.3), (4.6), (4.10), and (4.13).

THEOREM 4.2. *If $(\rho,\mathbf{u},p)$ is a solution to* (2.20) *which admits the asymptotic expansion* (4.1)*, then the density and the pressure have the expansions*

$$\rho=\rho^{(0)}(\mathbf{x},t)+O(\epsilon), \tag{4.14}$$

$$p=p^{(0)}(t)+\epsilon\left[\bar{p}^{(1)}(t)+\delta p^{(1)}(\mathbf{x},t,\tau)\right]+O(\epsilon^2). \tag{4.15}$$

Equations (4.11) and (4.12) convey further information about the leading order and the first order pressure. Averaging (4.11) and (4.12) with respect to τ and using the sublinearity condition (4.2), we obtain

$$\frac{1}{\gamma p^{(0)}}\frac{dp^{(0)}}{dt}+\nabla\cdot\bar{\mathbf{u}}^{(0)}=0, \tag{4.16}$$

$$\frac{d}{dt}\left(\frac{\bar{p}^{(1)}}{\gamma p^{(0)}}\right)+\nabla\cdot\bar{\mathbf{u}}^{(1)}=-\frac{1}{\gamma p^{(0)}}\left\langle\mathbf{u}^{(0)}\cdot\nabla p^{(1)}\right\rangle. \tag{4.17}$$

The next result follows immediately by subtracting (4.13) from (4.7) and (4.16) from (4.11) and recalling (4.14).

THEOREM 4.3. *If $(\rho, \mathbf{u}, p)$ is a solution to* (2.20) *which admits the asymptotic expansion* (4.1)*, then the average-free parts of the first order pressure and of the leading order velocity, $\delta p^{(1)}$ and $\delta \mathbf{u}^{(0)}$, respectively, satisfy the linear acoustic system*

$$\begin{aligned} &\partial_\tau \delta \mathbf{u}^{(0)} + \frac{1}{\rho^{(0)}} \nabla \delta p^{(1)} = 0, \\ &\partial_\tau \delta p^{(1)} + \gamma p^{(0)} \nabla \cdot \delta \mathbf{u}^{(0)} = 0. \end{aligned} \tag{4.18}$$

We can use the acoustic system (4.18) to simplify (4.17). Using the sublinearity condition (4.2), we immediately find

$$\left\langle \mathbf{u}^{(0)} \cdot \nabla p^{(1)} \right\rangle = \left\langle \delta \mathbf{u}^{(0)} \cdot \nabla \delta p^{(1)} \right\rangle = -\frac{\rho^{(0)}}{2} \left\langle \partial_\tau \left| \delta \mathbf{u}^{(0)} \right|^2 \right\rangle = 0, \tag{4.19}$$

and hence the right-hand side of (4.17) vanishes. Now we are ready to prove the analogue of Theorem 3.2.

THEOREM 4.4. *If $(\rho, \mathbf{u}, p)$ is a solution to* (2.20)*,* (2.22) *which admits the asymptotic expansion* (4.1)*, then $p^{(0)}(t)$ and $\bar{p}^{(1)}(t)$ are given by*

$$p^{(0)} = C_0 |\Omega_t|^{-\gamma}, \tag{4.20}$$

$$\bar{p}^{(1)} = C_1 p^{(0)}, \tag{4.21}$$

where $|\Omega_t|$ is the measure of the domain Ω_t and C_0, C_1 are constants.

Proof. The relations (4.20) and (4.21) follow from (4.16), (4.17), and (4.19) as in Theorem 3.2, once we note that

$$\bar{\mathbf{u}}^{(0)} \cdot \mathbf{n}_t = \mathbf{u}_\Omega \cdot \mathbf{n}_t, \qquad \bar{\mathbf{u}}^{(1)} \cdot \mathbf{n}_t = 0 \qquad \text{on } \partial\Omega_t. \tag{4.22}$$

The boundary conditions (4.22) follow from (2.22) after using the expansion (4.1) and averaging with respect to τ. ∎

In the following section we will show that the constants C_0 and C_1 can be recovered by the initial data (2.21) (see Theorem 5.2). As in the single scale analysis, (4.16) can be regarded as a constraint for $\bar{\mathbf{u}}^{(0)}$, which reduces to the incompressibility condition when the domain is not moving. To derive evolution equations for $(\rho^{(0)}, \bar{\mathbf{u}}^{(0)})$, we average (4.4) and (4.8) with respect to τ. Then using the condition (4.2) and recalling that the leading order density is independent of τ, we obtain

$$\partial_t \rho^{(0)} + \nabla \cdot \left(\rho^{(0)} \bar{\mathbf{u}}^{(0)} \right) = 0, \tag{4.23}$$

$$\partial_t \left(\rho^{(0)} \bar{\mathbf{u}}^{(0)} \right) + \nabla \cdot \left(\rho^{(0)} \left\langle \mathbf{u}^{(0)} \otimes \mathbf{u}^{(0)} \right\rangle \right) + \nabla \bar{p}^{(2)} = 0. \tag{4.24}$$

We can rewrite equations (4.23) and (4.24) so that the leading order material derivative $\frac{\bar{D}}{Dt} = \partial_t + \bar{\mathbf{u}}^{(0)} \cdot \nabla$ appears. Also, we can express $\nabla \cdot \bar{\mathbf{u}}^{(0)}$ in terms of $|\Omega_t|$ by using (4.16) and (4.20). The resulting equations are reported in the following statement.

THEOREM 4.5. *Under the same assumptions of Theorem* 4.4*, the averaged leading order density and velocity $(\rho^{(0)}, \bar{\mathbf{u}}^{(0)})$ satisfy the system*

$$\begin{aligned} &\frac{1}{\rho^{(0)}} \frac{\bar{D}\rho^{(0)}}{Dt} = -\frac{1}{|\Omega_t|} \frac{d|\Omega_t|}{dt}, \\ &\rho^{(0)} \frac{\bar{D}\bar{\mathbf{u}}^{(0)}}{Dt} + \nabla \bar{p}^{(2)} = -\nabla \cdot \left(\rho^{(0)} \left\langle \delta \mathbf{u}^{(0)} \otimes \delta \mathbf{u}^{(0)} \right\rangle \right), \\ &\nabla \cdot \bar{\mathbf{u}}^{(0)} = \frac{1}{|\Omega_t|} \frac{d|\Omega_t|}{dt}. \end{aligned} \tag{4.25}$$

Equation (4.25) is the analogue of (3.16). The main difference resides in the coupling of $(\rho^{(0)}, \bar{\mathbf{u}}^{(0)})$ with $\delta\mathbf{u}^{(0)}$ in $(4.25)_2$. In principle, (4.18), (4.20), and (4.25) determine completely $\rho^{(0)}$, $\mathbf{u}^{(0)} = \bar{\mathbf{u}}^{(0)} + \delta\mathbf{u}^{(0)}$, $p^{(1)} = \delta p^{(1)}$, and the average $\bar{p}^{(2)}$. We will discuss the appropriate initial-boundary conditions in the next section.

The next theorem concerns the higher order variables: $\bar{\rho}^{(1)} + \delta\rho^{(1)}$, $\bar{\mathbf{u}}^{(1)} + \delta\mathbf{u}^{(1)}$, and $\delta p^{(2)}$.

THEOREM 4.6. *If $(\rho, \mathbf{u}, p)$ is a solution to* (2.20), (2.21), (2.22) *which admits the asymptotic expansion* (4.1), *then the average-free functions $\delta\rho^{(1)}$, $\delta\mathbf{u}^{(1)}$, and $\delta p^{(2)}$ satisfy the linear acoustic system*

$$
\begin{aligned}
&\partial_\tau \delta\rho^{(1)} = -\nabla\cdot\left(\rho^{(0)}\delta\mathbf{u}^{(0)}\right),\\
&\rho^{(0)}\partial_\tau\delta\mathbf{u}^{(1)} + \nabla\delta p^{(2)} = -\rho^{(1)}\partial_\tau\delta\mathbf{u}^{(0)} - \rho^{(0)}(\partial_t + \mathbf{u}^{(0)}\cdot\nabla)\mathbf{u}^{(0)} - \nabla\bar{p}^{(2)},\\
&\partial_\tau\delta p^{(2)} + \gamma p^{(0)}\nabla\cdot\delta\mathbf{u}^{(1)} = -\gamma p^{(0)}(\partial_t + \mathbf{u}^{(0)}\cdot\nabla)\left(\frac{\delta p^{(1)}}{\gamma p^{(0)}}\right).
\end{aligned}
\tag{4.26}
$$

Moreover, the averaged perturbations $\bar{\rho}^{(1)}$, $\bar{\mathbf{u}}^{(1)}$ satisfy the equations

$$
\partial_t\bar{\rho}^{(1)} + \nabla\cdot\left(\rho^{(0)}\bar{\mathbf{u}}^{(1)} + \bar{\rho}^{(1)}\bar{\mathbf{u}}^{(0)}\right) = 0, \tag{4.27}
$$

$$
\begin{aligned}
\partial_t\left(\rho^{(0)}\bar{\mathbf{u}}^{(1)} + \bar{\rho}^{(1)}\bar{\mathbf{u}}^{(0)}\right) &+ \nabla\cdot\Big\{\bar{\rho}^{(1)}\bar{\mathbf{u}}^{(0)}\otimes\bar{\mathbf{u}}^{(0)}\\
&+ \rho^{(0)}\left(\bar{\mathbf{u}}^{(0)}\otimes\bar{\mathbf{u}}^{(1)} + \bar{\mathbf{u}}^{(1)}\otimes\bar{\mathbf{u}}^{(0)}\right)\Big\} + \nabla\bar{p}^{(3)}\\
= -\nabla\cdot\Big\{\Big\langle\rho^{(1)}\delta\mathbf{u}^{(0)}\otimes\delta\mathbf{u}^{(0)}\Big\rangle &+ \rho^{(0)}\Big\langle\delta\mathbf{u}^{(0)}\otimes\delta\mathbf{u}^{(1)} + \delta\mathbf{u}^{(1)}\otimes\delta\mathbf{u}^{(0)}\Big\rangle\Big\},
\end{aligned}
\tag{4.28}
$$

$$
\nabla\cdot\bar{\mathbf{u}}^{(1)} = 0. \tag{4.29}
$$

Proof. The system (4.26) comes from (4.4), (4.8), and (4.12) after some simple algebra by using (4.14) and (4.23). For the second part of the thesis, we average (4.5) and (4.9) with respect to τ and use the sublinearity condition. The result is

$$
\partial_t\bar{\rho}^{(1)} + \nabla\cdot\bar{\mathbf{m}}^{(1)} = 0, \tag{4.30}
$$

$$
\partial_t\bar{\mathbf{m}}^{(1)} + \nabla\cdot\left\langle\mathbf{m}^{(0)}\otimes\mathbf{u}^{(1)} + \mathbf{m}^{(1)}\otimes\mathbf{u}^{(0)}\right\rangle + \nabla\bar{p}^{(3)} = 0, \tag{4.31}
$$

with $\mathbf{m}^{(0)} = \rho^{(0)}\mathbf{u}^{(0)}$, $\mathbf{m}^{(1)} = \rho^{(0)}\mathbf{u}^{(1)} + \rho^{(1)}\mathbf{u}^{(0)}$. Recalling (4.14), we have

$$
\bar{\mathbf{m}}^{(1)} = \rho^{(0)}\bar{\mathbf{u}}^{(1)} + \bar{\rho}^{(1)}\bar{\mathbf{u}}^{(0)} + \left\langle\delta\rho^{(1)}\delta\mathbf{u}^{(0)}\right\rangle.
$$

We will show in section 6 that

$$
\left\langle\delta\rho^{(1)}\delta\mathbf{u}^{(0)}\right\rangle = 0. \tag{4.32}
$$

After some algebra, (4.30) and (4.31) yield (4.27) and (4.28). Finally, recalling (4.19), equation (4.17) reduces to (4.29) provided $\bar{p}^{(1)} = 0$. The proof of this equality depends on the boundary conditions and will be postponed to the following section (see Theorem 5.2). ∎

The average-free components $(\delta\rho^{(1)}, \delta\mathbf{u}^{(1)}, \delta p^{(2)})$ are coupled with $(\bar{\rho}^{(1)}, \bar{\mathbf{u}}^{(1)})$ through the averaged first order density $\bar{\rho}^{(1)}$ appearing in the right-hand side of $(4.26)_2$. There are two important cases when the two groups of equations (4.26)

and (4.27)–(4.29) decouple. The first case is when the average-free functions $\delta\mathbf{u}^{(0)}$ and $\delta p^{(1)}$ are identically zero. Then, the system (4.25) reduces to (3.16), the right-hand sides of the equations (4.26), (4.27), and (4.28) vanish, and $\delta\rho^{(1)} = 0$. The second case leading to decoupling is induced by a constant initial density distribution, $\rho(\mathbf{x}, 0) = \rho_0$. In this case, the leading order density $\rho^{(0)}$ is inversely proportional to the volume $|\Omega_t|$, the averaged first order density $\bar{\rho}^{(1)}$ vanishes, and the leading order entropy is constant. Since entropy variations appear only at order ϵ, the resulting flow will be called weakly nonisentropic. In the appendix at the end of the paper, we report the main results regarding weakly nonisentropic flows.

5. Initial-boundary conditions. In the previous section, we have shown that a solution of (2.20), which admits the expansion (4.1), takes the form

$$\begin{aligned}
\rho(\mathbf{x}, t; \epsilon) &= \rho^{(0)}(\mathbf{x}, t) + \epsilon \left[\bar{\rho}^{(1)}(\mathbf{x}, t) + \delta\rho^{(1)}(\mathbf{x}, t, \tau)\right] + O(\epsilon^2),\\
\mathbf{u}(\mathbf{x}, t; \epsilon) &= \left[\bar{\mathbf{u}}^{(0)}(\mathbf{x}, t) + \delta\mathbf{u}^{(0)}(\mathbf{x}, t, \tau)\right]\\
&\quad + \epsilon \left[\bar{\mathbf{u}}^{(1)}(\mathbf{x}, t) + \delta\mathbf{u}^{(1)}(\mathbf{x}, t, \tau)\right] + O(\epsilon^2),\\
p(\mathbf{x}, t; \epsilon) &= p^{(0)}(t) + \epsilon \left[\bar{p}^{(1)}(t) + \delta p^{(1)}(\mathbf{x}, t, \tau)\right]\\
&\quad + \epsilon^2 \left[\bar{p}^{(2)}(\mathbf{x}, t) + \delta p^{(2)}(\mathbf{x}, t, \tau)\right] + O(\epsilon^3),
\end{aligned} \tag{5.1}$$

where $p^{(0)}$ and $\bar{p}^{(1)}$ satisfy (4.20) and (4.21), $(\delta\mathbf{u}^{(0)}, \delta p^{(1)})$ satisfy (4.18), and $(\rho^{(0)}, \bar{\mathbf{u}}^{(0)})$ satisfy (4.25). Moreover, we have

$$\left\langle \delta\rho^{(1)} \right\rangle = 0, \qquad \left\langle \delta\mathbf{u}^{(0)} \right\rangle = \left\langle \delta\mathbf{u}^{(1)} \right\rangle = 0, \qquad \left\langle \delta p^{(1)} \right\rangle = \left\langle \delta p^{(2)} \right\rangle = 0. \tag{5.2}$$

In order to determine completely the asymptotic expansion (5.1), we need to assign at each order appropriate initial-boundary conditions which are induced by the initial-boundary conditions (2.21), (2.22) for the original system (2.20), that is,

$$\rho(\mathbf{x}, 0) = \rho_0(\mathbf{x}), \quad \mathbf{u}(\mathbf{x}, 0) = \mathbf{u}_0(\mathbf{x}), \quad p(\mathbf{x}, 0) = p_0(\mathbf{x}) \qquad \text{in} \quad \Omega_0, \tag{5.3}$$

$$\mathbf{u} \cdot \mathbf{n}_t = \mathbf{u}_\Omega \cdot \mathbf{n}_t \qquad \text{on} \quad \partial\Omega_t. \tag{5.4}$$

The theoretical analysis performed in [1, 7, 8, 9, 25, 33] shows that the low Mach number limit for generic data is not uniform in a time interval containing the zero, unless the initial data is compatible with the zero Mach number equations. Therefore, in general we must expect a sort of instantaneous jump from the initial data (5.3) to possibly different initial data adjusted to the asymptotic equations.

In this section, we collect some results which are related to the asymptotic initial-boundary conditions and can be derived immediately from the asymptotic expansion (4.1). First, we discuss the implications of the boundary condition (5.4). Using (5.1) in (5.4) and averaging with respect to τ, we find

$$\bar{\mathbf{u}}^{(0)} \cdot \mathbf{n}_t = \mathbf{u}_\Omega \cdot \mathbf{n}_t, \qquad \bar{\mathbf{u}}^{(1)} \cdot \mathbf{n}_t = 0 \qquad \text{on } \partial\Omega_t. \tag{5.5}$$

In turn, the previous condition implies

$$\delta\mathbf{u}^{(0)} \cdot \mathbf{n}_t = \delta\mathbf{u}^{(1)} \cdot \mathbf{n}_t = 0 \qquad \text{on } \partial\Omega_t. \tag{5.6}$$

Next, we discuss the relationship between the initial data (5.3) for the original system and the initial data for the functions appearing in the asymptotic expansion (5.1). We write $\tau = t/\epsilon$ in the expansion (5.1) and evaluate it at the time $t = 0$. Comparing the result with (5.3), we obtain the following conditions for the density:

$$\rho^{(0)}(\mathbf{x}, 0) = \rho_0(\mathbf{x}), \tag{5.7}$$

$$\bar{\rho}^{(i)}(\mathbf{x}, 0) + \delta\rho^{(i)}(\mathbf{x}, 0, 0) = 0 \qquad \text{for } i = 1, 2, \dots, \tag{5.8}$$

with $\mathbf{x} \in \Omega_0$. The condition (5.7) provides initial data for $\rho^{(0)}$. In the previous section, we have shown that

$$\partial_\tau \rho^{(i)} + \partial_t \rho^{(i-1)} + \nabla \cdot \mathbf{m}^{(i-1)} = 0, \qquad i = 1, 2.$$

It is simple to see that this equation is valid for all integers i. Then, subtracting the averaged equation and using the sublinearity condition (4.2), we obtain

$$\partial_\tau \delta\rho^{(i)} + \nabla \cdot \delta\mathbf{m}^{(i-1)} = 0, \qquad i = 1, 2, \dots. \tag{5.9}$$

Since $\mathbf{m}^{(i-1)}$ is known at this stage, we can integrate (5.9) and recover $\delta\rho^{(i)}$ using the condition $\langle \delta\rho^{(i)} \rangle = 0$. The result is

$$\delta\rho^{(i)}(\mathbf{x}, t, \tau) = \left\langle \nabla \cdot \int_0^\tau \delta\mathbf{m}^{(i-1)}(\mathbf{x}, t, \tau')\, d\tau \right\rangle - \nabla \cdot \int_0^\tau \delta\mathbf{m}^{(i-1)}(\mathbf{x}, t, \tau')\, d\tau, \tag{5.10}$$

which, using (5.8), gives

$$\bar{\rho}^{(i)}(\mathbf{x}, 0) = -\delta\rho^{(i)}(\mathbf{x}, 0, 0) = -\left\langle \nabla \cdot \int_0^\tau \delta\mathbf{m}^{(i-1)}(\mathbf{x}, 0, \tau')\, d\tau \right\rangle \quad \text{for } \mathbf{x} \in \Omega_0. \tag{5.11}$$

In particular, for $i = 1$ we have $\delta\mathbf{m}^{(0)} = \rho^{(0)}\delta\mathbf{u}^{(0)}$, which implies

$$\bar{\rho}^{(1)}(\mathbf{x}, 0) = -\delta\rho^{(1)}(\mathbf{x}, 0, 0) = 0 \qquad \text{for } \mathbf{x} \in \Omega_0. \tag{5.12}$$

Next, comparing the expansion (4.1) for the velocity, written at $t = 0$, and the initial data (5.3), we find

$$\bar{\mathbf{u}}^{(0)}(\mathbf{x}, 0) + \delta\mathbf{u}^{(0)}(\mathbf{x}, 0, 0) = \mathbf{u}_0(\mathbf{x}), \tag{5.13}$$

$$\bar{\mathbf{u}}^{(i)}(\mathbf{x}, 0) + \delta\mathbf{u}^{(i)}(\mathbf{x}, 0, 0) = 0 \qquad \text{for } i = 1, 2, \dots, \tag{5.14}$$

with $\mathbf{x} \in \Omega_0$. Some care is required to determine appropriate initial data for $\bar{\mathbf{u}}^{(i)}$, $i = 0, 1$, since the functions $\bar{\mathbf{u}}^{(0)}(\mathbf{x}, 0)$ and $\bar{\mathbf{u}}^{(1)}(\mathbf{x}, 0)$ must be compatible with $(4.25)_3$ and (4.29). Accordingly, we decompose the initial velocity $\mathbf{u}_0$ as

$$\mathbf{u}_0 = \mathbf{u}_0^* + \nabla\omega^* + \nabla\omega_0. \tag{5.15}$$

In (5.15), $\mathbf{u}_0^*$ is the divergence-free part of $\mathbf{u}_0$ in Ω_0, and ω^* is given by the system

$$\Delta\omega^* = \left(\frac{1}{|\Omega_t|} \frac{d|\Omega_t|}{dt} \right)_{t=0} \qquad \text{in } \Omega_0, \tag{5.16}$$

$$\frac{\partial\omega^*}{\partial\mathbf{n}_0} = 0 \qquad \text{on } \partial\Omega_0. \tag{5.17}$$

Using the decomposition (5.15), the condition (5.13), and the constraint $(4.25)_3$ leads to the reasonable assumption

$$\bar{\mathbf{u}}^{(0)}(\mathbf{x},0) = \mathbf{u}_0^*(\mathbf{x}) + \nabla\omega^*, \qquad \delta\mathbf{u}^{(0)}(\mathbf{x},0,0) = \nabla\omega_0(\mathbf{x}) \qquad \text{in } \Omega_0. \tag{5.18}$$

In the same way, using (4.29), it is simple to see that the condition (5.14) yields

$$\bar{\mathbf{u}}^{(1)}(\mathbf{x},0) = 0, \qquad \delta\mathbf{u}^{(1)}(\mathbf{x},0,0) = 0 \qquad \text{in } \Omega_0. \tag{5.19}$$

Finally, comparing the expansion (4.1) for the pressure, written at $t = 0$, and the initial data (5.3), we find

$$p^{(0)}(0) = p_0(\mathbf{x}), \tag{5.20}$$

$$\bar{p}^{(i)}(\mathbf{x},0) + \delta p^{(i)}(\mathbf{x},0,0) = 0 \qquad \text{for } i = 1,2,\dots, \tag{5.21}$$

with $\mathbf{x} \in \Omega_0$. It is immediately seen that (5.20) is compatible with a constant initial pressure p_0. However, as we will see, general initial data are admissible for p. Then, the condition (5.20) implies that the convergence as ϵ tends to zero cannot be uniform in a neighborhood of $t = 0$ if the initial pressure is not constant in Ω_0.

The functions $p^{(0)}$ and $\bar{p}^{(1)}$ will be determined in terms of $p_0(\mathbf{x})$ by simple physical considerations. We need the following lemma.

LEMMA 5.1. *For any function $a(\mathbf{x},t)$ defined for $\mathbf{x} \in \Omega_t$, $t \geq 0$, we have*

$$\frac{d}{dt}\int_{\Omega_t} a\,dV = \int_{\Omega_t} \partial_t a + \nabla\cdot(a\mathbf{u}^*)\,dV \tag{5.22}$$

for any vector-valued function $\mathbf{u}^*(\mathbf{x},t)$ such that

$$\mathbf{n}_t\cdot\mathbf{u}^* = \mathbf{n}_t\cdot\mathbf{u}_\Omega \qquad \textit{for } \mathbf{x} \in \partial\Omega_t. \tag{5.23}$$

In particular, if

$$\begin{aligned} \partial_t a + \nabla\cdot(a\mathbf{u}^*) &= \nabla\cdot\mathbf{v} \qquad \textit{for } \mathbf{x} \in \Omega_t,\\ \mathbf{n}_t\cdot\mathbf{v} &= 0 \qquad \textit{for } \mathbf{x} \in \partial\Omega_t, \end{aligned}$$

we have

$$\int_{\Omega_t} a\,dV = \int_{\Omega_0} a_0\,dV, \tag{5.24}$$

with $a_0(\mathbf{x}) = a(\mathbf{x},0)$.

Proof. We evaluate the left-hand side of (5.22) as

$$\begin{aligned} \frac{d}{dt}\int_{\Omega_t} a\,dV &= \frac{d}{dt}\int_{\Omega_0} a(\Phi_t(\mathbf{x}_0),t)\left|\frac{\partial\Phi_t}{\partial\mathbf{x}_0}\right| dV_0\\ &= \int_{\Omega_0}\left\{\left(\partial_t + \frac{\partial\Phi_t}{\partial t}\cdot\nabla\right)a + a\nabla\cdot\frac{\partial\Phi_t}{\partial t}\right\}\left|\frac{\partial\Phi_t}{\partial\mathbf{x}_0}\right| dV_0\\ &= \int_{\Omega_t}\left\{\partial_t a + \nabla\cdot(a\mathbf{u}^*) + \nabla\cdot\left[a\left(\frac{\partial\Phi_t}{\partial t} - \mathbf{u}^*\right)\right]\right\} dV\\ &= \int_{\Omega_t}\partial_t a + \nabla\cdot(a\mathbf{u}^*)\,dV + \int_{\partial\Omega_t} a\,\mathbf{n}_t\cdot\left(\frac{\partial\Phi_t}{\partial t} - \mathbf{u}^*\right) dS. \end{aligned}$$

The thesis follows immediately from (2.14) and (5.23). ∎

The following result completes Theorem 4.4.

THEOREM 5.2. *If* $(\rho, \mathbf{u}, p)$ *is a solution to* (2.20), (5.3), (5.4) *which admits the asymptotic expansion* (4.1), *then* $p^{(0)}(t)$ *and* $\bar{p}^{(1)}(t)$ *are given by*

$$p^{(0)}(t) = ||p_0||_{L^{1/\gamma}(\Omega_0)} |\Omega_t|^{-\gamma}, \tag{5.25}$$

$$\bar{p}^{(1)}(t) = 0, \tag{5.26}$$

where $p_0(\mathbf{x})$ *is the initial pressure. Moreover,* $\delta p^{(1)}$ *satisfies the condition*

$$\int_{\Omega_t} \delta p^{(1)} \, dV = 0 \qquad \text{for all } t, \tau \geq 0. \tag{5.27}$$

Proof. Using (2.4), it is possible to show that

$$\partial_t(\rho f(s)) + \nabla \cdot (\rho f(s) \mathbf{u}) = 0$$

for any smooth function f of the entropy s. Then, we can apply Lemma 5.1 with $a = \rho e^{s/\gamma}$, $\mathbf{u}^* = \mathbf{u}$, and $\mathbf{v} = 0$. Recalling (2.8), we obtain

$$\int_{\Omega_t} p^{1/\gamma} \, dV = \int_{\Omega_0} p_0^{1/\gamma} \, dV. \tag{5.28}$$

This equation is valid also when we use the power series expansion of p. At the first two orders in ϵ, we find

$$\int_{\Omega_t} \left[p^{(0)}\right]^{1/\gamma} \, dV = \int_{\Omega_0} p_0^{1/\gamma} \, dV, \tag{5.29}$$

$$\int_{\Omega_t} \frac{1}{\gamma} \left[p^{(0)}\right]^{1/\gamma - 1} \left(\bar{p}^{(1)} + \delta p^{(1)}\right) \, dV = 0. \tag{5.30}$$

Since $p^{(0)}$ and $\bar{p}^{(1)}$ are functions of time only, (5.29) immediately gives (5.25). Equation (5.30) implies

$$\bar{p}^{(1)}(t) = -\frac{1}{|\Omega_t|} \int_{\Omega_t} \delta p^{(1)} \, dV. \tag{5.31}$$

Averaging (5.31) with respect to τ, we obtain (5.26), and (5.31) becomes (5.27). ∎

As a consequence of (5.26), the condition (5.20) implies

$$\delta p^{(1)}(\mathbf{x}, 0, 0) = 0. \tag{5.32}$$

In general, once $\bar{p}^{(i)}$, $i = 2, 3, \ldots$, is given as a result of the evolutionary equations for $\bar{\mathbf{u}}^{(i-2)}$, the condition (5.21) gives the value of $\delta p^{(i)}$ at $t = \tau = 0$ in Ω_0.

In conclusion, we have shown how to assign initial-boundary conditions for all the averaged quantities appearing in the expansion (4.1), namely (5.5), (5.7), (5.12), (5.18), and (5.19). In particular, at the first two orders the pressure is explicitly given by (5.25) and (5.26). We still need to assign appropriate conditions for the average-free functions $\delta \mathbf{u}^{(0)}$, $\delta \mathbf{u}^{(1)}$, $\delta p^{(0)}$, and $\delta p^{(1)}$, depending on the fast time τ and satisfying the acoustic equations (4.18) and (4.26). This problem will be addressed in the following section.

6. The fast acoustic equation. This section is entirely devoted to determining the average-free functions appearing in the expansion (4.1). For simplicity of exposition, we will refer only to $\delta\mathbf{u}^{(0)}$ and $\delta p^{(1)}$, satisfying the system (4.18) and the conditions (5.6), (5.18), and (5.32). The same analysis can be extended to the next orders, since all the systems for $\delta\mathbf{u}^{(i)}$ and $\delta p^{(i+1)}$ share the same differential structure.

The slow time t and the fast time τ are linearly related by the small parameter ϵ. Since the parameter ϵ is arbitrary, it might be meaningful to regard the fast time τ as an independent variable and to assign, for $\tau = 0$, "initial" data depending on t. As a most unfortunate consequence of this approach, arbitrary "initial" data would be compatible with the initial data (5.3). Nevertheless, it is useful to derive a formal representation of the general solution of (4.18). This will be the object of the first part of this section. Subsequently, we will show that it is possible to resolve completely the average-free functions for a relevant class of motion of the domain Ω_t.

We consider the acoustic equation (4.18), supplemented with the boundary condition

$$\delta\mathbf{u}^{(0)} \cdot \mathbf{n}_t = 0 \qquad \text{on } \partial\Omega_t \tag{6.1}$$

and with initial data

$$\delta\mathbf{u}^{(0)}(\mathbf{x}, t, 0) = \nabla\omega_t(\mathbf{x}), \qquad \delta p^{(1)}(\mathbf{x}, t, 0) = \pi_t(\mathbf{x}) \qquad \text{in } \Omega_t, \tag{6.2}$$

consistently with (5.18) and (5.32), that is,

$$\delta\mathbf{u}^{(0)}(\mathbf{x}, 0, 0) = \nabla\omega_0(\mathbf{x}), \qquad \delta p^{(1)}(\mathbf{x}, 0, 0) = \pi_0(\mathbf{x}) = 0 \qquad \text{in } \Omega_0. \tag{6.3}$$

The acoustic system (4.18), (6.1), (6.2) can be replaced by the wave equation

$$\partial_\tau^2 \delta p^{(1)} - c^2 \nabla \cdot \left(r \nabla \delta p^{(1)} \right) = 0, \tag{6.4}$$

where

$$c^2(t) = \frac{\gamma p^{(0)}(t)}{\tilde{\rho}^{(0)}(t)}, \qquad r(\mathbf{x}, t) = \frac{\tilde{\rho}^{(0)}(t)}{\rho^{(0)}(\mathbf{x}, t)}, \qquad \tilde{\rho}^{(0)} = \frac{1}{|\Omega_t|} \int_{\Omega_t} \rho^{(0)}\, dV, \tag{6.5}$$

with boundary condition

$$\frac{\partial \delta p^{(1)}}{\partial \mathbf{n}_t} = 0 \qquad \text{on } \partial\Omega_t \tag{6.6}$$

and initial data

$$\delta p^{(1)}(\mathbf{x}, t, 0) = \pi_t(\mathbf{x}), \qquad \partial_\tau \delta p^{(1)}(\mathbf{x}, t, 0) = -\gamma p^{(0)} \Delta\omega_t(\mathbf{x}) \qquad \text{in } \Omega_t. \tag{6.7}$$

Once we know $\delta p^{(1)}$, the average-free velocity can be recovered from the equation $(4.18)_2$, supported with the condition $\langle \delta\mathbf{u}^{(0)} \rangle = 0$.

The parameter r in (6.4) characterizes the degree of "isentropicity" of the flow, in the sense that when $r = 1$ the flow is weakly nonisentropic. Some properties of weakly nonisentropic flows are reported in the appendix.

The operator $\nabla \cdot r\nabla$, which appears in (6.4), admits a family of eigenfunctions $\{w_j^t\}$ which forms an orthonormal basis of $L^2(\Omega_t)$. They are defined by

$$\begin{aligned} &\nabla \cdot \left(r \nabla w_j^t \right) = -\lambda_j^t w_j^t \qquad \text{in } \Omega_t, \\ &\frac{\partial w_j^t}{\partial \mathbf{n}_t} = 0 \qquad \text{on } \partial\Omega_t, \\ &\left|\left| w_j^t \right|\right|_{L^2(\Omega_t)} = 1. \end{aligned} \tag{6.8}$$

Then, we can express $\delta p^{(1)}$ as

$$\delta p^{(1)}(\mathbf{x},t,\tau)=\sum_j \sigma_j^t(\tau)w_j^t(\mathbf{x}). \tag{6.9}$$

The function σ_j^t for all j and t satisfies the system

$$\left(\frac{d^2}{d\tau^2}+c^2(t)\lambda_j^t\right)\sigma_j^t=0, \tag{6.10}$$

$$\sigma_j^t(0)=\left(\pi_t,w_j^t\right), \qquad \frac{d\sigma_j^t}{d\tau}(0)=-\gamma p^{(0)}\left(\Delta\omega_t,w_j^t\right). \tag{6.11}$$

The eigenvalues λ_j^t are nonnegative for all j and monotonically increasing to infinity as j increases. In particular, the eigenfunction w_0^t associated with the eigenvalue $\lambda_0^t=0$ is

$$w_0^t(\mathbf{x})=|\Omega|^{-1/2}. \tag{6.12}$$

Using (5.27) in (6.9), we get

$$\sigma_0^t(\tau)=0 \qquad \text{for all } t,\tau\geq 0, \tag{6.13}$$

and hence π_t and $\Delta\omega_t$ necessarily satisfy the condition

$$\int_{\Omega_t}\pi_t\,dV=0, \qquad \int_{\Omega_t}\Delta\omega_t\,dV=0. \tag{6.14}$$

The solution of (6.10) for $j>0$ is

$$\sigma_j^t(\tau)=\left(\pi_t,w_j^t\right)\cos\left[c(\lambda_j^t)^{1/2}\tau\right]-\frac{\tilde\rho^{(0)}c\left(\Delta\omega_t,w_j^t\right)}{(\lambda_j^t)^{1/2}}\sin\left[c(\lambda_j^t)^{1/2}\tau\right]. \tag{6.15}$$

Equations (6.9) and (6.15) give a full representation of $\delta p^{(1)}$ in terms of the eigenfunctions w_j^t and the eigenvectors λ_j^t of the operator $\nabla\cdot r\nabla$ in Ω_t. Moreover, using (4.18), (6.2), and (6.9), we find

$$\delta\mathbf{u}^{(0)}(\mathbf{x},t,\tau)=\sum_j S_j^t(\tau)r(\mathbf{x},t)\nabla w_j^t(\mathbf{x}), \tag{6.16}$$

$$S_j^t(\tau)=-\frac{\left(\pi_t,w_j^t\right)}{\tilde\rho^{(0)}c(\lambda_j^t)^{1/2}}\sin\left[c(\lambda_j^t)^{1/2}\tau\right]-\frac{\left(\Delta\omega_t,w_j^t\right)}{\lambda_j^t}\cos\left[c(\lambda_j^t)^{1/2}\tau\right], \tag{6.17}$$

together with the necessary constraint

$$\nabla\omega_t=-\sum_j\left(\Delta\omega_t,w_j^t\right)\frac{r}{\lambda_j^t}\nabla w_j^t. \tag{6.18}$$

Using (5.10), (6.16), and (6.17), it is possible to assess the validity of (4.32).

We conclude this section by showing that the fast acoustics can be completely resolved at least for a class of motion of the boundary $\partial\Omega_t$.

We consider any invertible map $\Phi_t:\Omega_0\to\Omega_t$ which describes the motion of $\partial\Omega_t$. For simplicity, we consider weakly isentropic flows so that $\rho^{(0)}=\rho^{(0)}(t)$. The

following transformation is compatible with the asymptotic analysis presented in this paper:

$$(\mathbf{x}, t, \tau) \to (\mathbf{y}, s, \sigma) = \left(\Psi_t(\mathbf{x}), s(t), \frac{s(t)}{\epsilon} \right) \qquad \text{with } s'(t) \neq 0. \tag{6.19}$$

Here, $\Psi_t : \Omega_t \to \Omega_0$ is the inverse of the map Φ_t, and $s(t)$ is a function to be determined (not the entropy!). After this change of variables the acoustic system (4.18) becomes

$$\begin{aligned} &\partial_\sigma \delta \mathbf{u}^{(0)} + \frac{1}{s' \rho^{(0)}} M \nabla_{\mathbf{y}} \delta p^{(1)} = 0, \\ &s' \partial_\sigma \delta p^{(1)} + \gamma p^{(0)} (M \nabla_{\mathbf{y}}) \cdot \delta \mathbf{u}^{(0)} = 0, \qquad \mathbf{y} \in \Omega_0, \end{aligned} \tag{6.20}$$

where $\nabla_{\mathbf{y}}$ is the gradient operator with respect to $\mathbf{y}$, $M(\mathbf{y}, t)$ is the inverse of the matrix $\partial \Phi_t(\mathbf{y}) / \partial \mathbf{y}$, and t is expressed everywhere as a function of s. The wave equation (6.4) is replaced by

$$\partial_\sigma^2 \delta p^{(1)} - \frac{\gamma p^{(0)}}{(s')^2 \rho^{(0)}} \left| M \nabla_{\mathbf{y}} \right|^2 \delta p^{(1)} = 0. \tag{6.21}$$

We wish to choose $s(t)$ so that the wave speed for (6.21) does not depend on t. This is not possible in general, since the matrix M depends on time as well as on the space variables. Anyway, there is a simple class of motion for which this problem has a solution corresponding to $\Phi_t(\mathbf{y}) = a(t)\mathbf{y} + \mathbf{b}(t)$. In this case we have $M = (1/a)I$. Then, for weakly nonisentropic flows, the choice

$$s'(t) = \text{constant} \cdot \left(\frac{\gamma p^{(0)}(t)}{\rho^{(0)}(t) a^2(t)} \right)^{\frac{1}{2}} \tag{6.22}$$

leads to a wave speed independent of the slow time t. Using (5.25) and (A.2), and noting that $|\Omega_t| / |\Omega_0| = a^n$, it is possible to rewrite the relation (6.22) in terms of $p^{(0)}$:

$$s'(t) = \text{constant} \cdot \left(p^{(0)}(t) \right)^{\frac{\gamma - 1 + 2/n}{2\gamma}}. \tag{6.23}$$

The wave equation (6.21) can be solved by following the method outlined in the first part of this section with initial data (6.3). For the one-dimensional piston problem, this approach leads to equivalent results to the ones obtained in [17].

7. Conclusions. In this paper we have derived the main implications of the ansatz (4.1) in the framework of the low Mach number limit for a compressible flow in a domain with variable boundary. We have shown that the limit incompressible variables are coupled with high-frequency oscillations produced by the motion of the boundary through the equations (4.18), (4.25). Understanding the role of the interplay with fast acoustics has an enormous relevance for the development of numerical schemes capable of dealing with low Mach number phenomena in a bounded domains.

The analysis performed here is not conclusive since the theory presented is not capable of providing a full resolution of high-frequency acoustics, as we have shown in the previous sections. Nevertheless, the representation given by (6.9) and (6.15) gives a first hint for a theoretical comprehension of the acoustic modes generated by the motion of the boundary. In our opinion, the key point is that one fast variable

is not sufficient to describe the sequence of modes produced by a generic motion of the boundary. Thus, we need to extend the theory by including a family of fast variables nonlinearly related to the slow time and to the space variables. The number of independent fast variables for each term of the expansion should increase with the order of the term. This extension, which has a theoretical interest in itself and is a necessary step for the development of efficient numerical schemes for low Mach number flows in bounded domains with moving boundary, is still a work in progress.

Appendix. Weakly nonisentropic flows. In this appendix we collect some results regarding weakly nonisentropic flows resulting from the condition

$$\rho(\mathbf{x},0)=\rho_0=\text{constant}. \tag{A.1}$$

THEOREM A.1. *Let* $(\rho,\mathbf{u},p)$ *be a solution of* (2.20) *which admits the asymptotic expansion* (4.1). *If* (A.1) *holds, the leading order density is a function of time only,* $\rho^{(0)}=\rho^{(0)}(t)$, *given by*

$$\rho^{(0)}(t)=\frac{|\Omega_0|}{|\Omega_t|}\rho_0. \tag{A.2}$$

Moreover,

$$s^{(0)}=\text{constant}, \tag{A.3}$$

where $s^{(0)}$ *is the leading order entropy.*

Proof. Using $(4.25)_1$, we obtain

$$\frac{\bar{D}}{Dt}\left(\rho^{(0)}|\Omega_t|\right)=0.$$

Then, applying Lemma 5.1 with $a=\rho^{(0)}(\rho^{(0)}|\Omega_t|)^{q-1}$, where q is an integer, $\mathbf{u}^*=\bar{\mathbf{u}}^{(0)}$, and $\mathbf{v}=0$, we find

$$\frac{1}{|\Omega_t|}\int_{\Omega_t}\left(\rho^{(0)}|\Omega_t|\right)^q\,dV=(\rho_0|\Omega_0|)^q\qquad\text{for all integers } q,$$

which implies

$$\left|\left|\rho^{(0)}|\Omega_t|-\rho_0|\Omega_0|\right|\right|_{L^q(\Omega_t)}=0\qquad\text{for all integers } q.$$

Thus, we have

$$\left|\left|\rho^{(0)}|\Omega_t|-\rho_0|\Omega_0|\right|\right|_{L^\infty(\Omega_t)}=0,$$

and hence (A.2). Finally, using (4.20) and (A.2) and recalling the equation of state (2.8), we have also (A.3). ∎

THEOREM A.2. *For a weakly nonisentropic flow, the leading order averaged velocity* $\bar{\mathbf{u}}^{(0)}$ *satisfies the system*

$$\begin{aligned}&\frac{\bar{D}\bar{\mathbf{u}}^{(0)}}{Dt}+\nabla\frac{\bar{p}^{(2)}}{\rho^{(0)}}=-\nabla\cdot\left\langle\delta\mathbf{u}^{(0)}\otimes\delta\mathbf{u}^{(0)}\right\rangle,\\&\nabla\cdot\bar{\mathbf{u}}^{(0)}=\frac{1}{|\Omega_t|}\frac{d|\Omega_t|}{dt}.\end{aligned} \tag{A.4}$$

Moreover, the leading order average-free velocity $\delta\mathbf{u}^{(0)}$ *and the first order average-free pressure* $\delta p^{(1)}$ *satisfy system* (4.18)*. In particular, we have*

$$\partial_\tau^2 \delta p^{(1)} - c^2 \Delta \delta p^{(1)} = 0, \tag{A.5}$$

with $c^2 = \gamma p^{(0)}/\rho^{(0)}$.

Proof. Equation (A.4) follows from (4.25) after using Theorem A.1. Equation (A.5) follows from (6.4) after using (A.2). ∎

THEOREM A.3. *For a weakly nonisentropic flow* $(\rho, \mathbf{u}, p)$*, the first order density* $\rho^{(1)} = \bar{\rho}^{(1)} + \delta\rho^{(1)}$ *is given by*

$$\bar{\rho}^{(1)} = 0, \tag{A.6}$$

$$\delta\rho^{(1)} = \frac{\rho^{(0)}}{\gamma p^{(0)}} \delta p^{(1)}. \tag{A.7}$$

Proof. Using (A.2) and (4.18), equation $(4.26)_1$ becomes

$$\partial_\tau \delta p^{(1)} = -\rho^{(0)} \nabla \cdot \delta\mathbf{u}^{(0)} = \partial_\tau \left(\frac{\rho^{(0)}}{\gamma p^{(0)}} \delta p^{(1)} \right).$$

After integrating with respect to τ, we obtain (A.7). Then, recalling the condition $(4.31)_3$, equation $(4.31)_1$ becomes

$$\partial_t \bar{\rho}^{(1)} + \nabla \cdot \left(\bar{\rho}^{(1)} \bar{\mathbf{u}}^{(0)} \right) = -\nabla \cdot \left\langle \delta\rho^{(1)} \delta\mathbf{u}^{(0)} \right\rangle = -\frac{\rho^{(0)}}{\gamma p^{(0)}} \nabla \cdot \left\langle \delta p^{(1)} \delta\mathbf{u}^{(0)} \right\rangle.$$

Using (4.18), we obtain

$$\nabla \cdot \left\langle \delta p^{(1)} \delta\mathbf{u}^{(0)} \right\rangle = -\rho^{(0)} \left\langle \delta\mathbf{u}^{(0)} \cdot \partial_\tau \delta\mathbf{u}^{(0)} \right\rangle - \frac{1}{\gamma p^{(0)}} \left\langle \delta p^{(1)} \partial_\tau \delta p^{(1)} \right\rangle = 0,$$

and hence

$$\partial_t \bar{\rho}^{(1)} + \nabla \cdot \left(\bar{\rho}^{(1)} \bar{\mathbf{u}}^{(0)} \right) = 0.$$

Using $(A.4)_2$, we find

$$\frac{1}{\bar{\rho}^{(1)}} \frac{\bar{D}\bar{\rho}^{(1)}}{Dt} = -\frac{1}{|\Omega_t|} \frac{d|\Omega_t|}{dt}.$$

Finally, proceeding as in Theorem A.1, we arrive at (A.6). ∎

REFERENCES

[1] K. ASANO, *On the incompressible limit of the compressible Euler equation*, Japan J. Appl. Math., 4 (1987), pp. 455–488.

[2] C. CHEVERRY, *The modulation equations of nonlinear geometric optics*, Comm. Partial Differential Equations, 21 (1996), pp. 1119–1140.

[3] C. CHEVERRY, *Justification de l'optique geometrique non lineaire pour un systeme de lois de conservation*, Duke Math. J., 87 (1997), pp. 213–263.

[4] R. DIPERNA AND A. MAJDA, *The validity of nonlinear geometric optics for weak solutions of conservation laws*, Comm. Math. Phys., 98 (1985), pp. 313–347.

[5] J. K. HUNTER, A. MAJDA, AND R. R. ROSALES, *Resonantly interacting weakly nonlinear hyperbolic waves* II. *Several space variables*, Stud. Appl. Math., 75 (1986), pp. 187–226.

[6] J. K. Hunter and J. B. Keller, *Weakly nonlinear high frequency waves*, Comm. Pure Appl. Math., 36 (1983), pp. 547–569.
[7] H. Isozaki, *Wave operators and the incompressible limit of the incompressible Euler equation in* $\mathbb{R}^n_+$, Comm. Math. Phys., 110 (1987), pp. 519–524.
[8] H. Isozaki, *Singular limits for the compressible Euler equation in an exterior domain*, J. Reine Angew. Math., 381 (1987), pp. 1–36.
[9] H. Isozaki, *Singular limits for the compressible Euler equation in an exterior domain.* II. *Bodies in a uniform flow*, Osaka J. Math., 26 (1989), pp. 399–410.
[10] J. L. Joly, G. Métivier, and J. Rauch, *Resonant one-dimensional nonlinear geometric optics*, J. Funct. Anal., 114 (1993), pp. 106–231.
[11] J. L. Joly, G. Métivier, and J. Rauch, *Coherent and focusing multidimensional nonlinear geometric optics*, Ann. Sci. École Norm. Sup. (4), 28 (1995), pp. 51–113.
[12] J. L. Joly and J. Rauch, *Nonlinear resonance can create dense oscillations*, in Microlocal Analysis and Nonlinear Waves, IMA Vol. Math. Appl. 30, M. Beals, R. Melrose, and J. Rauch, eds., Springer-Verlag, New York, 1991, pp. 113–123.
[13] S. Klainerman and A. Majda, *Singular perturbation of quasilinear hyperbolic systems with large parameters and the incompressible limit of compressible fluids*, Comm. Pure Appl. Math., 34 (1981), pp. 481–524.
[14] S. Klainerman and A. Majda, *Compressible and incompressible fluids*, Comm. Pure Appl. Math., 35 (1982), pp. 629–651.
[15] R. Klein, *Semi-implicit extension of a Godunov-type scheme based on low Mach number asymptotics* I: *One-dimensional flow*, J. Comput. Phys., 121 (1995), pp. 213–237.
[16] R. Klein, N. Botta, L. Hofmann, A. Meister, C. D. Munz, S. Roller, and T. Sonar, *Asymptotic adaptive methods for multiscale problems in fluid mechanics*, J. Engrg. Math., 39 (2001), pp. 261–343.
[17] R. Klein and N. Peters, *Cumulative effects of weak pressure waves during the induction period of a thermal explosion in a closed cylinder*, J. Fluid Mech., 187 (1988), pp. 197–230.
[18] A. Majda, *Compressible Fluid Flow and Systems of Conservation Laws in Several Space Variables*, Springer-Verlag, New York, Berlin, Heidelberg, Tokyo, 1984.
[19] A. Majda, *Nonlinear geometric optics for hyperbolic systems of conservation laws*, in Oscillation Theory, Computation and Methods of Compensated Compactness, IMA Vol. Math. Appl. 2, C. Dafermos, J. Erikson, D. Kinderlehrer, and M. Slemrod, eds., Springer-Verlag, New York, 1986, pp. 115–165.
[20] A. Majda and R. R. Rosales, *Resonantly interacting weakly nonlinear hyperbolic waves* I. *A single space variable*, Stud. Appl. Math., 71 (1983), pp. 149.
[21] A. Majda and J. Sethian, *The derivation and numerical solution of the equations for zero Mach number combustion*, Combustion Science and Technology, 42 (1985), pp. 185–205.
[22] W. H. Matthaeus and M. R. Brown, *Nearly incompressible magnetohydrodynamics at low Mach number*, Phys. Fluids, 31 (1988), pp. 3634–3644.
[23] A. Meister, *Asymptotic single and multiple scale expansions in the low Mach number limit*, SIAM J. Appl. Math., 60 (1999), pp. 256–271.
[24] R. Menikoff and B. J. Plohr, *Riemann problem for fluid flow of real materials*, Rev. Modern Phys., 61 (1989), pp. 75–130.
[25] G. Métivier and S. Schochet, *The incompressible limit of the non-isentropic Euler equations*, Arch. Ration. Mech. Anal., 158 (2001), pp. 61–90.
[26] R. Pego, *Some explicit resonating waves in weakly nonlinear gas dynamics*, Stud. Appl. Math., 79 (1988), pp. 263–270.
[27] S. Schochet, *The compressible Euler equations in a bounded domain: Existence of solutions and the incompressible limit*, Comm. Math. Phys., 104 (1986), pp. 49–75.
[28] S. Schochet, *Resonant nonlinear geometric optics for weak solutions of conservation laws*, J. Differential Equations, 113 (1994), pp. 473–504.
[29] S. Schochet, *Fast singular limits of hyperbolic PDEs*, J. Differential Equations, 114 (1994), pp. 476–512.
[30] W. Schneider, *Mathematische Methoden in der Strömungsmechanik*, Vieweg, Braunschweig, 1978.
[31] Th. Schneider, N. Botta, K.-J. Geratz, and R. Klein, *Extension of finite volume compressible flow solvers to multi-dimensional, variable density, zero Mach number flow*, J. Comput. Phys., 155 (1999), pp. 248–286.
[32] F. Tiberi Timperi, *Simulazione numerica della compressione di un pistone e confronto con l'approssimazione adiabatica*, Tesi di laurea, Università dell'Aquila, L'Aquila, Italy, 1997.

[33] S. Ukai, *The incompressible limit and the initial layer of the compressible Euler equation*, J. Math. Kyoto Univ., 26 (1986), pp. 323–331.
[34] G. P. Zank and W. H. Matthaeus, *The equations of nearly incompressible fluids.* I. *Hydrodynamics, turbulence and waves*, Phys. Fluids A, 3 (1991), pp. 69–82.
[35] G. P. Zank and W. H. Matthaeus, *Nearly incompressible fluids.* II. *Magnetohydrodynamics, turbulence and waves*, Phys. Fluids A, 5 (1993), pp. 257–273.

SIAM J. APPL. MATH.
Vol. 63, No. 6, pp. 2042–2062

FROZEN PATH APPROXIMATION FOR TURBULENT DIFFUSION AND FRACTIONAL BROWNIAN MOTION IN RANDOM FLOWS*

ALBERT FANNJIANG† AND TOMASZ KOMOROWSKI‡

Abstract. We establish the conditions for the frozen path approximation for turbulent transport in a class of nonmixing Gaussian flows with long-range correlation. We identify the regimes of fractional Brownian motion limit as well as the Brownian motion limit.

Key words. turbulent transport, fractional Brownian motion, Taylor–Kubo formula

AMS subject classifications. Primary, 60F05, 76F05, 76R50; Secondary, 58F25

DOI. 10.1137/S0036139998335293

1. Introduction. The study of turbulent transport is fundamental to understanding of temperature fields as well as pollutant or tracer particles movement in the atmosphere and oceans and solute transport in groundwater flows [1]. For a long time, the Brownian motion (BM) and the heat equation have been the paradigm for describing large-scale turbulent transport since Taylor's works in the 1920s. The wide applicability of the Brownian motion and the related Gaussian processes have much to do with the central limit theorem which is often assumed to be valid over large scales if there is no memory or intermittency effect.

To account for the memory or intermittency effect, anomalous diffusions have been introduced in recent years as phenomenological models within the framework of fractional kinetic equations or continuous-time random walks (see [21], [12], [19], and the references therein). The mechanisms for anomalous behaviors are generally attributed to long waiting times (subdiffusion) or long flights (superdiffusion) or both. The former results in fractional-in-time (hence non-Markovian) differential operators while the latter results in fractional-in-space differential operators. In both cases the underlying processes are non-Gaussian.

In this paper we derive rigorously the fractional Brownian motions (FBMs) as limiting processes of large-scale motions of particles being advected by a family of random flows that are decorrelated both in space and time but in a manner depending on the wave modes of the velocity. This dependence is described in terms of two crucial parameters (α and β) of the flows. Our limit theorem also characterizes the multiple-particle motions in the FBM regime. FBMs are Gaussian but non-Markovian processes and are different from the phenomenological models mentioned above. The FBMs we find in this paper are invariably superdiffusive due to the positive memory effect, while the FBMs we found elsewhere [7] for a different type of flows can be subdiffusive as well as superdiffusive. By varying the parameters we see that the

*Received by the editors March 9, 1998; accepted for publication (in revised form) August 27, 2002; published electronically September 4, 2003.
http://www.siam.org/journals/siap/63-6/33529.html

†Department of Mathematics, UC Davis, One Shields Ave., Davis, CA 95616-8633 (cafannjiang@ucdavis.edu). The research of this author was supported by National Science Foundation grant DMS-9971322 and the Centennial Fellowship from the American Mathematical Society.

‡Institute of Mathematics, University of Marii Curie-Skłodowskiej, pl. Marii Curie-Skłodowskiej 1, 20-031 Lublin, Poland and Institute of Mathematics of the Polish Academy of Sciences, ul. Sniadeckich 8, 00-950 Warzawa, Poland (komorow@golem.umcs.lublin.pl). The research of this author was supported by KBN grant 2PO3A 031 23.

limiting processes can switch from FBMs to the Brownian motion, and we characterize the boundary of transition precisely.

The FBM regime indicates the breakdown of the central limit theorem, but the Gaussianity persists in the limit and is inherited from that of the velocity field. It is an open problem if one would obtain non-Gaussian limits for non-Gaussian velocity fields, which are beyond the methodology of the paper.

For the particle displacement $\boldsymbol{x}(t)$ in a given random velocity $\mathbf{V}(t,\mathbf{x})$ we consider the general large-scale limit

$$\boldsymbol{x}^{\varepsilon}(t) = \varepsilon \boldsymbol{x}\left(\frac{t}{\varepsilon^{2q}}\right)$$

satisfying the equation

$$d\boldsymbol{x}^{\varepsilon}(t) = \varepsilon^{1-2q}\mathbf{V}(\varepsilon^{-2q}t, \varepsilon^{-p}\boldsymbol{x}^{\varepsilon}(t))dt + \varepsilon^{1-q}\sqrt{2\kappa}d\mathbf{B}(t), \quad p \geq 0, \tag{1.1}$$

for some $q > 0$ (to be determined) as ϵ tends to zero. Here $\mathbf{B}(t)$ is the Brownian motion and κ is the molecular diffusivity. The special case of $p = 0$ and $q = 1$ is the white-noise-in-time limit. The scaling limit with $p = 1$ is the homogenization limit.

We assume that, in addition to incompressibility, the velocity $\mathbf{V}(t,\mathbf{x}), (t,\mathbf{x}) \in \mathbb{R} \times \mathbb{R}^d$, is a zero mean, time-stationary, space-homogeneous, isotropic, Ornstein–Uhlenbeck (thus, Gaussian and Markovian) process with *long-range correlations* (see below). Here the scaling exponent q depends on the correlation functions of the velocity. The scaling limit (1.1) has been studied by Kesten and Papanicolaou [11] in the case of $p = 0$ and Komorowki [13], [14] in the general case of $0 \leq p < 1$ for velocities sufficiently strongly mixing in time, and in this situation the scaling exponent is always $q = 1$, i.e., the *diffusive* scaling, and the limiting process is a Brownian motion with the diffusion coefficients given by the Taylor–Kubo formula [22]

$$D_{ij} = \int_0^{\infty} \{\mathbb{E}[V_i(t,\mathbf{0})V_j(0,\mathbf{0})] + \mathbb{E}[V_j(t,\mathbf{0})V_i(0,\mathbf{0})]\}\, dt. \tag{1.2}$$

To understand how the long-range correlation in velocity fields may change the diffusive scaling, we study the weak coupling limit for Ornstein–Uhlenbeck velocities with long-range correlations in both space and time (thus, nonmixing) defined as follows. We define the family of velocity fields with power-law spectra as follows. Let $(\Omega, \mathcal{V}, P)$ be a probability space of which each element is a velocity field $\mathbf{V}(t,\mathbf{x})$, $(t,\mathbf{x}) \in \mathbb{R} \times \mathbb{R}^d$ satisfying the following properties.

(H1) $\mathbf{V}(t,\mathbf{x})$ is time-stationary, space-homogeneous, and centered, i.e., $\mathbb{E}\mathbf{V}(0,\mathbf{0}) = \mathbf{0}$, and Gaussian. Here $\mathbb{E}$ stands for the expectation with respect to the probability measure P.

(H2) The two-point correlation tensor $\mathbf{R} = [R_{ij}]$ is given by

$$\begin{aligned} R_{ij}(t,\mathbf{x}) &= \mathbb{E}\left[V_i(t,\mathbf{x})V_j(0,\mathbf{0})\right] \\ &= \int_{\mathbb{R}^d} \cos\left(\mathbf{k}\cdot\mathbf{x}\right) \exp\left(-|\mathbf{k}|^{2\beta}t\right) \frac{\mathcal{E}(|\mathbf{k}|)}{|\mathbf{k}|^{d-1}} \\ &\qquad \times \left(\mathbf{I} - \frac{\mathbf{k}\otimes\mathbf{k}}{|\mathbf{k}|^{d-1}}\right) d\mathbf{k}, \quad \beta \geq 0, \quad d \geq 2, \end{aligned} \tag{1.3}$$

with the spatial spectral density

$$\mathcal{E}(k) = \frac{a(k)}{k^{2\alpha-1}}, \quad \alpha < 1, \tag{1.4}$$

where $a : [0, +\infty) \to \mathbb{R}_+$ is a compactly supported, continuous, nonnegative function. The factor $\mathbf{I} - \mathbf{k} \otimes \mathbf{k}/|\mathbf{k}|^2$ in (1.3) ensures the incompressibility.

Note that for $\alpha < 1$ the instantaneous two-point correlation functions $R_{ij}(0, \mathbf{x})$ decays to zero as $|\mathbf{x}|$ tends to infinity. The velocity is strongly temporally mixing if and only if $\beta = 0$ (see [20]).

We show that the scaling limit is either a Brownian motion or a persistent (i.e., superdiffusive) FBM as stated in the following theorem.

THEOREM 1. *Let the velocity field satisfy properties* (H1)–(H2) *with* $p < 1$.

Case 1. *For* $\alpha + \beta < 1$ *and the scaling exponent*

$$q = 1,$$

the solution $\boldsymbol{x}^\varepsilon(t)$ *converges in distribution, as* ε *tends to zero, to the Brownian motion with the covariance matrix given by the Kubo formula* (1.2) *plus* $\kappa\mathbf{I}$.

Case 2. *For* $1 < \alpha + \beta,\ \alpha + 2\beta < 1 + 1/p$, *and the scaling exponent*

$$q := \frac{\beta}{\alpha + 2\beta - 1}, \tag{1.5}$$

the solution $\boldsymbol{x}^\varepsilon(t)$ *converges in probability, as* ε *tends to zero, to a fractional Brownian motion* $\mathbf{B}_H(t)$ *with covariance given by*

$$Cov(\mathbf{B}_H(t_1), \mathbf{B}_H(t_2)) = \frac{1}{2}\mathbf{D}\left\{|t_1|^{2H} + |t_2|^{2H} - |t_1 - t_2|^{2H}\right\} \tag{1.6}$$

with the coefficients $\mathbf{D}$

$$\mathbf{D} = \int_{\mathbb{R}^d} \frac{e^{-|\mathbf{k}|^{2\beta}} - 1 + |\mathbf{k}|^{2\beta}}{|\mathbf{k}|^{2\alpha+4\beta-1}} \left(\mathbf{I} - \frac{\mathbf{k} \otimes \mathbf{k}}{|\mathbf{k}|^2}\right) \frac{a(0)}{|\mathbf{k}|^{d-1}} d\mathbf{k} \tag{1.7}$$

and the Hurst exponent H,

$$1/2 < H = 1/2 + \frac{\alpha + \beta - 1}{2\beta} < 1. \tag{1.8}$$

The homogenization scaling with $p = 1$ has been considered in [2], [3], [8], [15] and the corresponding scaling exponent q is the same. But the eddy diffusion matrix is no longer given by the Kubo formula.

We also establish the following results, which are very useful for understanding the simultaneous limit of the motion of multiple particles.

THEOREM 2. *Under the same assumptions of Theorem* 1, *the following approximations are valid in the respective regimes in the mean square sense for sufficiently small* ε:

Case 1.

$$\boldsymbol{x}^\varepsilon(t) = \mathbf{W}_\varepsilon(t) + o(1)$$

with

$$\begin{aligned}\mathbf{W}_\varepsilon(t) : = \int_0^t \int_{\mathbb{R}^d} & \frac{|\mathbf{k}|^\beta \mathcal{E}^{\frac{1}{2}}(|\mathbf{k}|)}{(|\mathbf{k}|^{2\beta} + \varepsilon^2)|\mathbf{k}|^{\frac{(d-1)}{2}}} \left(\mathbf{I} - \frac{\mathbf{k} \otimes \mathbf{k}}{|\mathbf{k}|^2}\right)^{\frac{1}{2}} \\ & [\cos(\varepsilon^{-p}\boldsymbol{x}^\varepsilon(s) \cdot \mathbf{k})\mathbf{W}_0(ds, d\mathbf{k}) + \sin(\varepsilon^{-p}\boldsymbol{x}^\varepsilon(s) \cdot \mathbf{k})\mathbf{W}_1(ds, d\mathbf{k})],\end{aligned} \tag{1.9}$$

where $\mathbf{W}_i(dt, d\mathbf{k})$, $i = 0, 1$, *are two independent copies of a d-dimensional space-time white-noise field (see* [16] *for a thorough discussion).*

Case 2.

$$\boldsymbol{x}^\varepsilon(t) = \boldsymbol{x}^\varepsilon(0) + \int_0^t \varepsilon^{1-2q}\mathbf{V}(\varepsilon^{-2q}s, \varepsilon^{-p}\boldsymbol{x}^\varepsilon(0))ds + o(1). \tag{1.10}$$

The surprising feature about the approximation (1.10) is that the "frozen path" approximation is asymptotically exact on the time scale of observation. Thus the multiple-point motion can be easily derived.

The process $\mathbf{W}_\varepsilon(t)$ defined by (1.9) is a continuous martingale with the quadratic variation

$$\langle \mathbf{W}_\varepsilon \rangle_t = t \int_{\mathbb{R}^d} \frac{|\mathbf{k}|^{2\beta}\mathcal{E}(|\mathbf{k}|)}{(|\mathbf{k}|^{2\beta} + \varepsilon^2)^2} \left(\mathbf{I} - \frac{\mathbf{k}\otimes\mathbf{k}}{|\mathbf{k}|^2}\right) \frac{d\mathbf{k}}{|\mathbf{k}|^{d-1}}.$$

Thus we know that $\mathbf{W}_\varepsilon(t)$, $t \geq 0$, is a Brownian motion. It is easy to check that the ratio $\langle \mathbf{W}_\varepsilon \rangle_t / t$ converges to the Kubo formula as ε tends to zero.

Theorem 1 characterizes the limit of one-point motion whereas Theorem 2 enables us to calculate the limit of multiple-point motion with each particle starting from a different point. It is straightforward to check from the corresponding approximations (1.9) and (1.10) that any two particles with a *fixed* initial separation in $\boldsymbol{x}^\varepsilon(0)$ become, in the limit $\varepsilon \to 0$, *independent* Brownian or FBMs for $p > 0$. However, if the initial separation of particles is of order ε^p, then the resulting limit processes are correlated as in the case of $p = 0$ which has been studied in [6]. The proofs of Theorem 1 and 2 use (finite) diagrammatic expansion and are given in sections 4–6.1. In the main text we present the physical explanation of the theorems in terms of the frozen path approximation. The results are shown schematically in Figure 1. In section 7 we provide a scaling argument for the case of $p > 1$ for the fractional Brownian regime.

When an additional infrared cutoff of the size ε^γ is introduced in the velocity spectrum, the results depend on whether the cutoff is *subcritical*, $\gamma < (\alpha + 2\beta - 1)^{-1}$, or *supercritical*, $\gamma > (\alpha + 2\beta - 1)^{-1}$. A supercritical cutoff does not affect the diagram, but a subcritical cutoff does. In particular, the regime of FBM limit disappears, and the limit is always a Brownian motion when the infrared cutoff is subcritical (see [2], [3]). We will not further discuss the effect of infrared cutoff in this paper.

The effect of molecular diffusion on transport may be subtle (see [18], [7]). However, for isotropic flows with monotonically decaying temporal correlation, small molecular diffusivity is negligible and will only affect results perturbatively. So we set $\kappa = 0$ from now on to simplify the presentation.

2. Brownian motion limit. Let us first consider the case of the Brownian motion limit. We express the displacement in the integral form

$$\boldsymbol{x}_\varepsilon(t) = \boldsymbol{x}_\varepsilon(0) + \frac{1}{\varepsilon}\int_0^t \mathbf{V}\left(\frac{t_1}{\varepsilon^2}, \frac{\boldsymbol{x}_\varepsilon(t_1)}{\varepsilon^p}\right) dt_1.$$

Assuming for simplicity that the spatial derivative of the velocity field is uniformly bounded, we know that the frozen path approximation

$$\boldsymbol{x}_\varepsilon(t) \approx \tilde{\boldsymbol{x}}_\varepsilon(t) = \boldsymbol{x}_\varepsilon(0) + \frac{1}{\varepsilon}\int_0^t \mathbf{V}\left(\frac{s}{\varepsilon^2}, \frac{\boldsymbol{x}_\varepsilon(0)}{\varepsilon^p}\right) ds, \quad 0 < t < \tau,$$

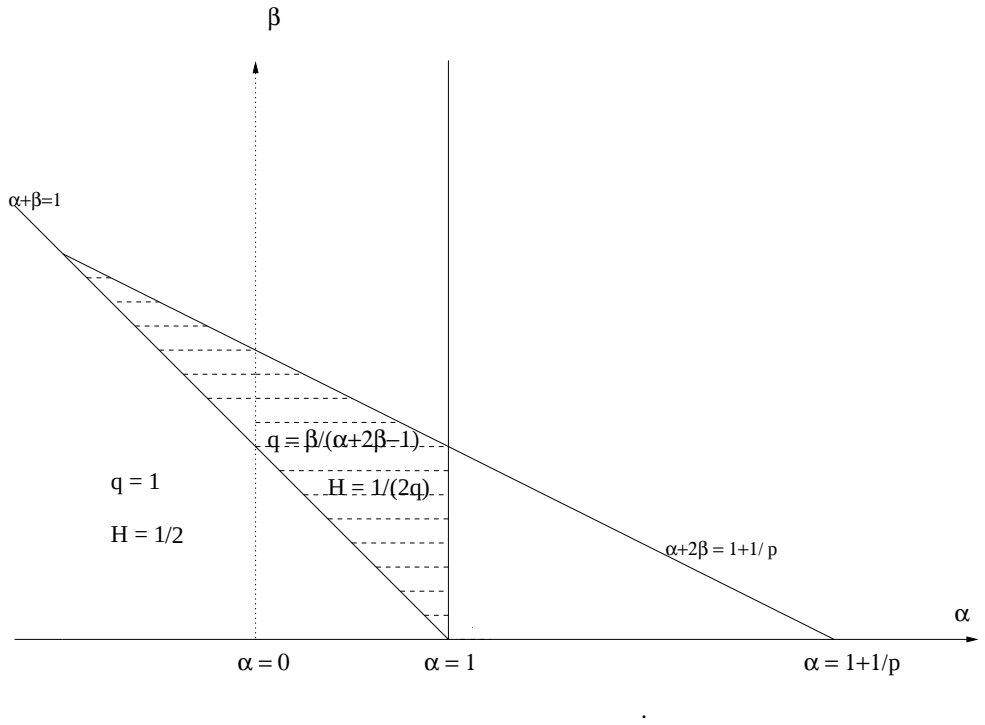


FIG. 1. *Phase diagram with supercritical infrared cutoff.*

is accurate pathwise with an error of $O(\tau^{3/2}\varepsilon^{-1-p})$ on the time scale

$$\varepsilon^2 \ll \tau \ll \varepsilon^{p+1} \tag{2.1}$$

(cf. (3.5)). One then expects that, for small ε, the displacement $\boldsymbol{x}_\varepsilon(t)$ is approximately the sum, $\tilde{\boldsymbol{x}}_\varepsilon(t)$, of t/τ random variables in the form

$$\Delta\tilde{\boldsymbol{x}}^n_\varepsilon(\tau) = \tilde{\boldsymbol{x}}_\varepsilon((n+1)\tau) - \tilde{\boldsymbol{x}}_\varepsilon(n\tau) = \varepsilon \int_{n\tau/\varepsilon^2}^{(n+1)\tau/\varepsilon^2} \mathbf{V}\left(s, \frac{\tilde{\boldsymbol{x}}_\varepsilon(n\tau)}{\varepsilon^p}\right) ds, \quad n = 0, 1, 2, \ldots.$$

Since $\tau \gg \varepsilon^2$, by the central limit theorem for processes with mixing, stationary increments (cf. [20]), the process $\Delta\tilde{\boldsymbol{x}}^n_\varepsilon(t)$,

$$\Delta\tilde{\boldsymbol{x}}^n_\varepsilon(t) = \varepsilon \int_{n\tau/\varepsilon^2}^{(n\tau+t)/\varepsilon^2} \mathbf{V}\left(s, \frac{\tilde{\boldsymbol{x}}_\varepsilon(n\tau)}{\varepsilon^p}\right) ds, \quad 0 < t \le \tau,$$

conditioned on $\tilde{\boldsymbol{x}}_\varepsilon(n\tau)$, is approximately a Brownian motion, starting at 0, with diffusion coefficient given by the Taylor–Kubo formula (1.2). Since $\tau \gg \varepsilon^2$ and the Taylor–Kubo formula converges, $\Delta\tilde{\boldsymbol{x}}^n_\varepsilon$ are nearly uncorrelated for different n and the total error made by the frozen path approximation is $O(\tau\varepsilon^{-1-p})$, which is negligible for $\tau \ll \varepsilon^{1+p}$.

The question is, What is the region in the (α, β) plane where the classical turbulent diffusion theorem, with the Taylor–Kubo formula (1.2), holds? It is easy to find the necessary condition by imposing the convergence of the Taylor–Kubo formula

(1.2). A straightforward calculation

$$\begin{aligned} D^*_{ij} &= \int_0^\infty R_{ij}(t,\mathbf{0})dt \\ &= \int_{\mathbb{R}^d} \left(\delta_{ij} - \frac{k_i k_j}{|\mathbf{k}|^2}\right) \frac{\mathcal{E}(|\mathbf{k}|)}{|\mathbf{k}|^{d-1}} \int_0^\infty \exp\left(-|\mathbf{k}|^{2\beta} t\right) dt d\mathbf{k} \\ &= \int_{\mathbb{R}^d} \left(\delta_{ij} - \frac{k_i k_j}{|\mathbf{k}|^2}\right) \frac{\mathcal{E}(|\mathbf{k}|)}{|\mathbf{k}|^{d+2\beta-1}} d\mathbf{k} \end{aligned} \tag{2.2}$$

leads to the condition

$$\alpha + \beta < 1. \tag{2.3}$$

It turns out that (2.3) is also sufficient. In other words, the classical turbulent diffusion theorem holds for this family of Gaussian velocity fields if and only if (2.3) is true (see section 6.1).

Let us see what the frozen path approximation tells us. The covariance of the Gaussian increment $\Delta\tilde{\boldsymbol{x}}^n_\varepsilon(t),\ 0 < t < \tau$ (given by (2.1)), stationary with respect to n, can be expressed as

$$\begin{aligned} & 2\varepsilon^2 \int_0^{t/\varepsilon^2} \int_0^{s_1} R_{ij}(s_1 - s_2, \mathbf{0})\ ds_2\ ds_1 \\ &= 2\varepsilon^2 \int_0^{t/\varepsilon^2} \int_{\mathbb{R}^d} \left(\delta_{ij} - \frac{k_i k_j}{|\mathbf{k}|^2}\right) \frac{\mathcal{E}(|\mathbf{k}|)}{|\mathbf{k}|^{d+2\beta-1}} (1 - e^{-s_1|\mathbf{k}|^{2\beta}}) d\mathbf{k} ds_1 \\ &= 2\int_0^t \int_{\mathbb{R}^d} \left(\delta_{ij} - \frac{k_i k_j}{|\mathbf{k}|^2}\right) \frac{\mathcal{E}(|\mathbf{k}|)}{|\mathbf{k}|^{d+2\beta-1}} (1 - e^{-t_1|\mathbf{k}|^{2\beta}/\varepsilon^2}) d\mathbf{k} dt_1 \\ &= 2D^*_{ij} t - 2\int_0^t \int_{\mathbb{R}^d} \left(\delta_{ij} - \frac{k_i k_j}{|\mathbf{k}|^2}\right) \frac{\mathcal{E}(|\mathbf{k}|)}{|\mathbf{k}|^{d+2\beta-1}} e^{-t_1|\mathbf{k}|^{2\beta}/\varepsilon^2} d\mathbf{k} dt_1 \\ &= 2D^*_{ij} t - 2\int_{\mathbb{R}^d} \left(\delta_{ij} - \frac{k_i k_j}{|\mathbf{k}|^2}\right) \frac{\mathcal{E}(|\mathbf{k}|)}{|\mathbf{k}|^{d+2\beta-1}} \times \frac{\varepsilon^2}{|\mathbf{k}|^{2\beta}} (1 - e^{-t|\mathbf{k}|^{2\beta}/\varepsilon^2})\, d\mathbf{k} \end{aligned} \tag{2.4}$$

with D^* given by the Taylor–Kubo formula (2.2). The last integral can be estimated by breaking it into two parts: $|\mathbf{k}|^{2\beta} < \varepsilon^2/t$ and $|\mathbf{k}|^{2\beta} \geq \varepsilon^2/t$. The first part has the asymptotic

$$\begin{aligned} & 2t \int_{|\mathbf{k}|^{2\beta} < \varepsilon^2/t} \left(\delta_{ij} - \frac{k_i k_j}{|\mathbf{k}|^2}\right) \frac{\mathcal{E}(|\mathbf{k}|)}{|\mathbf{k}|^{d+2\beta-1}}\, d\mathbf{k} \\ &\sim |\mathbf{k}|^{2-2\alpha-2\beta} \Big|_0^{(\frac{\varepsilon^2}{t})^{1/(2\beta)}} t \\ &= \varepsilon^{2(1-\alpha-\beta)/\beta} t^{(\alpha+2\beta-1)/\beta}, \end{aligned} \tag{2.5}$$

which, if $\alpha + 2\beta > 1$, gives rise to the *subdiffusive* FBM with the Hurst exponent

$$H = \frac{\alpha + 2\beta - 1}{2\beta} \ < \ 1/2$$

and vanishing coefficient since $\alpha + \beta < 1$. The second part can be estimated by

$$2\int_{|\mathbf{k}|^{2\beta} \geq \varepsilon^2/t} \frac{\mathcal{E}(|\mathbf{k}|)}{|\mathbf{k}|^{d+2\beta-1}} \times \frac{\varepsilon^2}{|\mathbf{k}|^{2\beta}}\, d\mathbf{k}$$

$$\sim 2\varepsilon^2\left[\left(\frac{\varepsilon^2}{t}\right)^{(1-\alpha-2\beta)/\beta} - K^{2(1-\alpha-2\beta)}\right]$$

$$= 2\varepsilon^{2(1-\alpha-\beta)/\beta}t^{(\alpha+2\beta-1)/\beta} - 2\varepsilon^2 K^{2(1-\alpha-2\beta)}. \tag{2.6}$$

Thus, if $\alpha+2\beta<1$, the second term in (2.6) dominates the first and, if $\alpha+2\beta>1$, the first dominates. But both (2.5) and (2.6) are negligible relative to the leading term $2D^*_{ij}t$.

Therefore, for $\alpha+\beta<1$, the displacement $\boldsymbol{x}_\varepsilon(t)$ behaves like the Brownian motion, with the diffusion coefficient given by the Taylor–Kubo formula (2.2), plus a correction term. When $\alpha+2\beta>1$ the correction term is like a *subdiffusive* FBM.

3. FBM limit. What happens if (2.3) is violated? The divergence of the Taylor–Kubo formula (1.2) suggests a superdiffusive behavior and, consequently, a different scaling limit.

Consider the superdiffusive scaling on the displacement

$$\boldsymbol{x}_\varepsilon(t) = \varepsilon\boldsymbol{x}\left(\frac{t}{\varepsilon^{2q}}\right), \quad q<1. \tag{3.1}$$

The equation of motion becomes

$$\frac{d\boldsymbol{x}_\varepsilon(t)}{dt} = \frac{1}{\varepsilon^{2q-1}}\mathbf{V}\left(\frac{t}{\varepsilon^{2q}}, \frac{\boldsymbol{x}_\varepsilon(t)}{\varepsilon^p}\right), \quad p<1. \tag{3.2}$$

The frozen path argument will show that for

$$\alpha+\beta>1, \quad \alpha<1, \tag{3.3}$$

and

$$q = \frac{\beta}{\alpha+2\beta-1}, \tag{3.4}$$

the solution $\boldsymbol{x}_\varepsilon(t)$ of (3.2) converges to an FBM.

First we note that the frozen path approximation

$$\boldsymbol{x}_\varepsilon(t) \approx \tilde{\boldsymbol{x}}_\varepsilon(t) = \boldsymbol{x}_\varepsilon(0) + \frac{1}{\varepsilon^{2q-1}}\int_0^t \mathbf{V}\left(\frac{t_1}{\varepsilon^{2q}}, \frac{\boldsymbol{x}_\varepsilon(0)}{\varepsilon^p}\right)dt_1$$

is accurate with the error $O(\tau^{1+1/(2q)}\varepsilon^{1-p-2q})$ on the (rescaled) time scale τ,

$$\varepsilon^{2q} \ll \tau \ll \varepsilon^{p+2q-1}, \tag{3.5}$$

provided that the scaling exponent q is the right one (i.e., $\boldsymbol{x}_\varepsilon(t), t>0$ is $O(1)$). The upper limit on τ is imposed in (3.5) because the total error made by the frozen path approximation is then $O(\tau\varepsilon^{1-p-2q})$, which is negligible.

Let us calculate the covariance of the Gaussian increment

$$\Delta\tilde{\boldsymbol{x}}^n_\varepsilon(t) = \varepsilon\int_{n\tau/\varepsilon^{2q}}^{(n\tau+t)/\varepsilon^{2q}} \mathbf{V}\left(s, \frac{\tilde{\boldsymbol{x}}_\varepsilon(n\tau)}{\varepsilon^p}\right)ds, \quad 0<t\le\tau,$$

which is stationary in n. Denoting by $\mathbf{R}_s$ the symmetric part of the covariance matrix $\mathbf{R}$, we have

$$\begin{aligned}
&\mathbb{E}[\Delta\tilde{\mathbf{x}}_\varepsilon^n(t)\otimes\Delta\tilde{\mathbf{x}}_\varepsilon^n(t)]\\
&= 2\varepsilon^2\int_0^{t/\varepsilon^{2q}}\int_0^{s_1}\mathbf{R}_s(s_1-s_2,0)\,ds_1\,ds_2\\
&= 2\varepsilon^2\int_0^{t/\varepsilon^{2q}}\int_{\mathbb{R}^d}\left(\mathbf{I}-\frac{\mathbf{k}\otimes\mathbf{k}}{|\mathbf{k}|^2}\right)\frac{\mathcal{E}(\mathbf{k})}{|\mathbf{k}|^{d+2\beta-1}}(1-e^{-s_1|\mathbf{k}|^{2\beta}})\,ds_1\,d\mathbf{k}\\
&= 2\varepsilon^{2(1-q)}\int_0^t\left(\int_{|\mathbf{k}|^{2\beta}<\varepsilon^{2q}/t_1}+\int_{|\mathbf{k}|^{2\beta}\ge\varepsilon^{2q}/t_1}\right)\\
&\qquad\times\left(\mathbf{I}-\frac{\mathbf{k}\otimes\mathbf{k}}{|\mathbf{k}|^2}\right)\frac{\mathcal{E}(\mathbf{k})}{|\mathbf{k}|^{d+2\beta-1}}(1-e^{-t_1|\mathbf{k}|^{2\beta}/\varepsilon^{2q}})\,d\mathbf{k}\,dt_1.
\end{aligned}$$

The first integral has the order of magnitude

$$\varepsilon^{2(1-q)}\int_0^t\int_{|\mathbf{k}|^{2\beta}<\varepsilon^{2q}/t_1}\frac{\mathcal{E}(\mathbf{k})}{|\mathbf{k}|^{d+2\beta-1}}\times\frac{t_1|\mathbf{k}|^{2\beta}}{\varepsilon^{2q}}\,d\mathbf{k}\,dt_1\sim\varepsilon^{2[1-q(\alpha+2\beta-1)/\beta]}t^{(\alpha+2\beta-1)/\beta}.$$

The second integral has the order of magnitude

$$\varepsilon^{2(1-q)}\int_0^t\int_{|\mathbf{k}|^{2\beta}\ge\varepsilon^{2q}/t_1}\frac{\mathcal{E}(\mathbf{k})}{|\mathbf{k}|^{d+2\beta-1}}\,d\mathbf{k}\,dt_1\sim\varepsilon^{2[1-q(\alpha+2\beta-1)/\beta]}t^{(\alpha+2\beta-1)/\beta}.$$

They are of the same sign so they do not cancel with each other. With (3.4) both terms behave like the FBM of finite, constant coefficients with the Hurst exponent $H=1/(2q)$ on the (rescaled) time scales in the range given by (3.5). In particular, for $p=0$, the FBM limit holds up to order one time as is rigorously proved in [6]. The scaling with (3.4) is *superdiffusive* since $q<1$ for $\alpha+\beta>1$. This is the result of *positive* correlation between successive increments. For the FBM-like behavior to persist up to order one times for $p>0$ the stationary increments at different times must have the right positive correlation. This is proved in section 6.

4. Estimation by diagrammatic method. We now turn to the proof of Theorem 1. We shall only calculate the mean square displacement of the particle. We make use of a spectral representation of the velocity field as follows. Let $\hat{\mathbf{V}}_0(t,d\mathbf{k}),\hat{\mathbf{V}}_1(t,d\mathbf{k})$ be two independent copies of real $\mathbb{R}^d$-valued, Gaussian, random spectral measures with the structure matrix

$$\mathbb{E}[\hat{\mathbf{V}}_i(t,d\mathbf{k})\otimes\hat{\mathbf{V}}_i(0,d\mathbf{k})]=\frac{\mathrm{e}^{-|\mathbf{k}|^{2\beta}t}\mathcal{E}(|\mathbf{k}|)}{|\mathbf{k}|^{d-1}}\left(\mathbf{I}-\frac{\mathbf{k}\otimes\mathbf{k}}{|\mathbf{k}|^2}\right)d\mathbf{k},\quad i=0,1. \tag{4.1}$$

The modes of the random measure can be intuitively thought of as mutually independent "infinitesimal" Ornstein–Uhlenbeck processes, that is, a stationary solution of a properly understood (e.g., in the sense of generalized functions) stochastic differential equation

$$\begin{aligned}
&d_t\hat{\mathbf{V}}_i(t,d\mathbf{k})\\
&\quad=-|\mathbf{k}|^{2\beta}\hat{\mathbf{V}}_i(t,d\mathbf{k})dt+|\mathbf{k}|^{(2\beta+1-d)/2}\mathcal{E}^{\frac12}(|\mathbf{k}|)\left(\mathbf{I}-\frac{\mathbf{k}\otimes\mathbf{k}}{|\mathbf{k}|^2}\right)\mathbf{W}_i(dt,d\mathbf{k}),\quad i=0,1.
\end{aligned} \tag{4.2}$$

Here $\mathbf{W}_i(dt, d\mathbf{k})$, $i = 0, 1$, are independent $\mathbb{R}^d$-valued, uncorrelated space-time white-noise random measures.

We can write then that

$$\mathbf{V}(t, \mathbf{x}) = \int \hat{\mathbf{V}}(t, \mathbf{x}, d\mathbf{k}), \tag{4.3}$$

with

$$\hat{\mathbf{V}}(t, \mathbf{x}, d\mathbf{k}) := \mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{x}}\hat{\mathbf{V}}(t, d\mathbf{k}) \tag{4.4}$$

and $\hat{\mathbf{V}}(t, \cdot)$ a $\mathbb{C}^d$-valued, componentwise Gaussian random measure given by

$$\hat{\mathbf{V}}(t, A) := \frac{1}{2}[\hat{\mathbf{V}}_0(t, A) + \hat{\mathbf{V}}_0(t, -A)] + \frac{\mathrm{i}}{2}[\hat{\mathbf{V}}_1(t, A) - \hat{\mathbf{V}}_1(t, -A)]. \tag{4.5}$$

The velocity field is temporally Markovian because for any Borel set A and $s \le t$

$$\mathbb{E}_s\hat{\mathbf{V}}(t, d\mathbf{k}) = \mathrm{e}^{-|\mathbf{k}|^{2\beta}(t-s)}\hat{\mathbf{V}}(s, d\mathbf{k}). \tag{4.6}$$

Here $\mathbb{E}_s$ denotes the conditional expectation with respect to the history of the random field determined up to time s. Another property of temporal dynamics of the field is its *reversibility*, which can be expressed in the following form. For any $s_1 \ge s_2 \ge \cdots s_n \ge 0$ and functions F, G of appropriate arguments, we have

$$\begin{aligned}&\mathbb{E}\left\{\mathbb{E}_0\left[F(\hat{\mathbf{V}}(s_1, d\mathbf{k}_1), \ldots, \hat{\mathbf{V}}(s_n, d\mathbf{k}_n))\right] G(\hat{\mathbf{V}}(0, d\mathbf{k}_{n+1}))\right\}\\ &= \mathbb{E}\left[F(\hat{\mathbf{V}}(s_1 - s_n, d\mathbf{k}_1), \ldots, \hat{\mathbf{V}}(0, d\mathbf{k}_n))\mathbb{E}_0 G(\hat{\mathbf{V}}(s_n, d\mathbf{k}_{n+1}))\right].\end{aligned} \tag{4.7}$$

As explained in the introduction, the molecular diffusion has only a perturbative effect and will be set to zero to simplify the calculation. The motion of the tracer is then described by

$$\frac{d\boldsymbol{x}(t)}{dt} = \mathbf{V}(t, \varepsilon^{1-p}\boldsymbol{x}(t)). \tag{4.8}$$

Let us set

$$\boldsymbol{x}^\varepsilon(t) = \varepsilon\int_0^{t\varepsilon^{-2q}} \mathbf{V}(s, \varepsilon^{1-p}\boldsymbol{x}(s))ds, \tag{4.9}$$

where $p < 1$ and $\mathbf{x}(t)$ is given by (4.8) and q is to be specified later.

For any $t \ge s$ define $\Delta_n(t, s) := [(s_1, \ldots, s_{n+1}) : t \ge s_1 \ge \cdots \ge s_{n+1} \ge s]$. To compute the mean square displacement of the particle we write

$$\begin{aligned}\mathbb{E}\left[\mathbf{x}_\varepsilon(t) \otimes \mathbf{x}_\varepsilon(t)\right] &= \varepsilon^2\int_0^{t\varepsilon^{-2q}} ds \int_0^s \left\{\mathbb{E}\left[\mathbf{V}(s_1, \varepsilon^{1-p}\mathbf{x}(s_1)) \otimes \mathbf{V}(0, \mathbf{0})\right]\right.\\ &\qquad \left. + \mathbb{E}\left[\mathbf{V}(0, \mathbf{0}) \otimes \mathbf{V}(s_1, \varepsilon^{1-p}\mathbf{x}(s_1))\right]\right\} ds_1\\ &= \sum_{n=0}^{N-1} \mathcal{I}_{n,\varepsilon}(t) + \mathcal{R}_{N,\varepsilon}(t),\end{aligned} \tag{4.10}$$

with $\mathcal{I}_{n,\varepsilon}(t)$ the symmetric part of the matrix

$$\mathcal{I}^0_{n,\varepsilon}(t) := 2\varepsilon^{n(1-p)+2}\int_0^{t\varepsilon^{-2q}} ds \int_{\Delta_n(s,0)}\cdots\int \mathbb{E}\left\{\mathbb{E}_0\left[\mathbf{W}_n(s_1,\ldots,s_{n+1},\mathbf{0})\right] \otimes \mathbf{V}(0,\mathbf{0})\right\} ds_1\cdots ds_{n+1} \tag{4.11}$$

and $\mathcal{R}_{N,\varepsilon}(t)$ the symmetric part of the matrix

$$\begin{aligned}&\mathcal{R}^0_{N,\varepsilon}(t)\\ &= 2\varepsilon^{N(1-p)+2}\int_0^{t\varepsilon^{-2q}} ds \int_{\Delta_N(s,0)}\cdots\int \mathbb{E}\left\{\mathbb{E}_{s_{N+1}}\left[\mathbf{W}_N(s_1,\ldots,s_{N+1},\varepsilon^{1-p}\mathbf{x}(s_{N+1}))\right] \otimes \mathbf{V}(0,\mathbf{0})\right\} ds_1\cdots ds_{N+1},\end{aligned}$$

where $\mathbf{W}_n(\cdot)$ is defined inductively by

$$\mathbf{W}_0(s_1,\mathbf{x}) := \mathbf{V}(s_1,\mathbf{x}), \tag{4.12}$$

$$\mathbf{W}_n(s_1,\ldots,s_{n+1},\mathbf{x}) := \mathbf{V}(s_{n+1},\mathbf{x}) \cdot\nabla\mathbf{W}_{n-1}(s_1,\ldots,s_n,\mathbf{x}) \qquad \text{for } n=1,2,\ldots. \tag{4.13}$$

To estimate both $\mathcal{I}_n$ and the remainder term $\mathcal{R}_{N,\varepsilon}(t)$ we shall deal with expectations of polynomial-like expressions in a Gaussian variable. To calculate the expectation of multiple product of Gaussian random variables, we use Feynman diagrams borrowed from quantum field theory (see, e.g., [9] and [10]). A *Feynman diagram* $\mathcal{F}$ (of order n = number of vertexes and rank r = number of edges) is a graph consisting of a set $B(\mathcal{F})$ of n vertexes and a set $E(\mathcal{F})$ of r edges without common endpoints. So there are r pairs of vertexes, each joined by an edge, and $n-2r$ unpaired vertexes, called *free vertexes*. Let $B(\mathcal{F})$ be a subset of positive integers. An edge whose endpoints are $m,n\in B$ is represented by $\widehat{mn}$ (unless otherwise specified, we always assume $m<n$); an edge includes its endpoints. A diagram $\mathcal{F}$ is said to be *based on* $B(\mathcal{F})$. Denote the set of free vertexes by $A(\mathcal{F})$, so $A(\mathcal{F}) = \mathcal{F}\setminus E(\mathcal{F})$. The diagram is *complete* if $A(\mathcal{F})$ is empty and *incomplete* otherwise. Denote by $\mathcal{G}(B)$ the set of all diagrams based on B, by $\mathcal{G}_c(B)$ the set of all complete diagrams based on B, and by $\mathcal{G}_{\text{in}}(B)$ the set of all incomplete diagrams based on B. A diagram $\mathcal{F}'\in\mathcal{G}_c(B)$ is called a *completion* of $\mathcal{F}\in\mathcal{G}_i(B)$ if $E(\mathcal{F})\subseteq E(\mathcal{F}')$.

Let $\mathbb{Z}_n := \{1,2,3,\ldots,n\}$. For $n\geq 1$ we define inductively a class $\mathfrak{S}_n$ of certain Feynman diagrams based on $\mathbb{Z}_n$ as follows. For $n=1$, $\mathfrak{S}_1$ consists of the trivial diagram $\mathcal{F}$ with vertex 1. Given $\mathfrak{S}_{n-1}$, $\mathfrak{S}_n$ consists of all the *descendants* of $\mathfrak{S}_{n-1}$. A descendant $\mathcal{F}'$ of $\mathcal{F}\in\mathfrak{S}_{n-1}$ is a graph based on $\mathbb{Z}_n$ such that $A(\mathcal{F}')\neq\emptyset$ and

$$\mathcal{F}'_{|n-1} = \mathcal{F}, \tag{4.14}$$

where $\mathcal{F}'_{|n-1}$ is the restriction of $\mathcal{F}'$ to $\mathbb{Z}_{n-1}$ with edges of the type $\widehat{mn}$, $m=1,2,\ldots,n-1$, deleted. We call $\mathcal{F}$ the *predecessor* of $\mathcal{F}'$. The predecessor of any $\mathcal{F}'\in\mathfrak{S}_n$ is clearly unique. For $\mathcal{F}\in\mathfrak{S}_n$ set $A_k(\mathcal{F}) = A(\mathcal{F}_{|k})$, $k=1,2,\ldots,n$.

Adopting the multi-index notation for any $N\in\mathbb{Z}^+$, $\mathbf{n}=(n_1,\ldots,n_{N+1})$, $n_j\in\{1,2,3,\ldots,d\}$, and $|\mathbf{n}| := n_1+n_2+\ldots+n_{N+1}$, we have the following formula.

LEMMA 1. *Let $N\geq 1$ and $s_1\geq s_2\geq\cdots\geq s_{N+1}\geq 0$, $i\in\{1,\ldots,d\}$, $\mathbf{x}\in\mathbb{R}^d$. We have then that*

(4.15)

$$\mathbb{E}_{s_{N+1}} W_{N,i}(s_1,\dots,s_{N+1},\mathbf{x}) = \sum_{\mathbf{n},\mathcal{F}} \int\cdots\int \exp\left\{ \mathrm{i} \sum_{m\in A_N(\mathcal{F})\cup\{N+1\}} \mathbf{k}_m\cdot\mathbf{x} \right\}$$
$$\times\ \mathrm{i}^N \prod_{j=1}^N \left(\sum_{m\in A_j(\mathcal{F})} \mathbf{k}_m \right) \exp\left\{ - \sum_{m\in A_j(\mathcal{F})} |\mathbf{k}_m|^{2\beta}(s_j - s_{j+1}) \right\}$$
$$\times\ C_{\mathbf{n},N} \prod_{\widehat{mm'}\in E(\mathcal{F})} \mathbb{E}\left[\widehat{V}_{n_m}(0,d\mathbf{k}_m)\widehat{V}_{n_{m'}}(0,d\mathbf{k}_{m'}) \right] \prod_{m\in A_N(\mathcal{F})\cup\{N+1\}} \widehat{V}_{n_m}(s_{N+1},d\mathbf{k}_m),$$

where $|C_{\mathbf{n},N}| \le 1$. *The summation extends over all integer valued multi-indices* $\mathbf{n} = (n_1,\dots,n_{N+1})$, $n_1 = i$, *and all Feynman diagrams* $\mathcal{F} \in \mathfrak{S}_N$.

The proof of Lemma 1 is a straightforward moment calculation with jointly Gaussian random variables using spectral representation (4.3)–(4.4). The free vertexes arise from centering and the edges from covariance of each pair. The condition $A(\mathcal{F}') \neq \emptyset$ is due to the gradient operation. The term $C_{\mathbf{n},N}$ contains an $O(1)$ factor like

$$\prod_{\widehat{mm'}\in E(\mathcal{F})} \left[1 - e^{-2|\mathbf{k}_{m'}|^{2\beta}(s_{m'} - s_{N+1})} \right]$$

resulting from replacing the conditional covariance by the covariance of the pairing (cf. [6]).

Using Lemma 1 we can write that

(4.16)
$$\int_0^{t\varepsilon^{-2q}} ds \int_{\Delta_N(s,0)}\cdots\int \mathbb{E}_{s_{N+1}} W_{N,i}(s_1,\dots,s_{N+1},\mathbf{x}) ds_1\cdots ds_{N+1}$$
$$= \sum \int_0^{t\varepsilon^{-2q}} ds \int_0^s ds' \int_{\Delta_{N-1}(s,s')}\cdots\int \varphi_N(\mathbf{k}_1,\dots,\mathbf{k}_N) P_N(\mathbf{x},\mathbf{k}_1,\dots,\mathbf{k}_N;\mathcal{F})$$
$$\times \prod_{\widehat{mm'}\in E(\mathcal{F})} \mathbb{E}\left[\widehat{V}_{n_m}(0,d\mathbf{k}_m)\widehat{V}_{n_{m'}}(0,d\mathbf{k}_{m'}) \right] \prod_{m\in A_N(\mathcal{F})\cup\{N+1\}} \widehat{V}_{n_m}(s',d\mathbf{k}_m)$$

for $i = 1,\dots,d$. Here,

(4.17)
$$\varphi_N(\mathbf{x},\mathbf{k}_1,\dots,\mathbf{k}_N) := \mathrm{i}^N C_{\mathbf{n},N} \prod_{j=1}^N \frac{\sum_{m\in A_j(\mathcal{F})} |\mathbf{k}_m|^{2\beta}}{1 - \exp\left\{ - \sum_{m\in A_j(\mathcal{F})} |\mathbf{k}_m|^{2\beta} t\varepsilon^{-2q}) \right\}} \times \frac{\sum_{m\in A_j(\mathcal{F})} \mathbf{k}_m}{\sum_{m\in A_j(\mathcal{F})} |\mathbf{k}_m|}$$
$$\times \exp\left\{ \mathrm{i} \sum_{m\in A_N(\mathcal{F})\cup\{N+1\}} \mathbf{k}_m\cdot\mathbf{x} \right\}$$
$$\times \int_{\Delta_{N-1}(s,s')}\cdots\int \prod_{j=1}^N \exp\left\{ - \sum_{m\in A_j(\mathcal{F})} |\mathbf{k}_m|^{2\beta}(s_j - s_{j+1}) \right\} ds_1\cdots ds_N$$

and

(4.18)
$$P_N(\mathbf{k}_1,\dots,\mathbf{k}_N;\mathcal{F}) = \prod_{j=1}^{N}\left\{\left(\sum_{m\in A_j(\mathcal{F})}|\mathbf{k}_m|\right)\times\frac{1-\exp\left\{-\sum\limits_{m\in A_j(\mathcal{F})}|\mathbf{k}_m|^{2\beta}t\varepsilon^{-2q}\right\}}{\sum\limits_{m\in A_j(\mathcal{F})}|\mathbf{k}_m|^{2\beta}}\right\}.$$

It is elementary to check that

$$|\varphi_N(\mathbf{x},\mathbf{k}_1,\dots,\mathbf{k}_N)|\le 1. \tag{4.19}$$

To further estimate the expression (4.16) we need to have more refined analysis of the graphs $\mathcal{F}\in\mathfrak{S}_N$. Define

$$r_N(\mathcal{F}) := \max\{m : m\in A_N(\mathcal{F})\}, \tag{4.20}$$
$$a_N(\mathcal{F}) := \min\{m : m\in A_N(\mathcal{F})\}. \tag{4.21}$$

We define $a_k(\mathcal{F}),\ k<N$, as

$$a_{k-1}(\mathcal{F}) = \min\{m : m\in A_{a_k}(\mathcal{F})\}$$

successively unless $a_k=1$. In other words, a_{k-1} is the smallest integer which is the left endpoint of an edge with its right endpoint greater than a_k; cf. (4.14). Below we will use the short-hand notation $a_k := a_k(\mathcal{F})$. Note that $A_N(\mathcal{F})$ and $A_{a_k-1}(\mathcal{F}),\ a_k>1$, are mutually disjoint. Let

$$\mathcal{A}(\mathcal{F}) := A_N(\mathcal{F})\bigcup_{k:\,a_k>1}A_{a_k-1}(\mathcal{F}). \tag{4.22}$$

Observe that any vertex $m\in\mathcal{A}(\mathcal{F})$ cannot be a right endpoint of any edge in $E(\mathcal{F})$. For any $m\in\mathcal{A}(\mathcal{F})$ let m^* be the nearest vertex in $A(\mathcal{F})$ to the right of m, i.e.,

$$m^* := \min[k:\ k\in\mathcal{A}(\mathcal{F}),\, k>m] \tag{4.23}$$

if the relevant set is nonempty; otherwise, set $m^*:=N$. Let

$$q_m := \#\{\widehat{pp'}\in E(\mathcal{F}) : m<p'<m^*\} \tag{4.24}$$

and let $e(\mathcal{F})$, $c(\mathcal{F})$ be the cardinalities of $E(\mathcal{F})$ and $A_N(\mathcal{F})$, respectively. It is easy to see that

$$\sum_{m\in\mathcal{A}} q_m = e(\mathcal{F}), \tag{4.25}$$

and thus

$$\sum_{m\in\mathcal{A}(\mathcal{F})} q_m + e(\mathcal{F}) + c(\mathcal{F}) = N. \tag{4.26}$$

4.1. Estimates for the remainder terms $\mathcal{R}_{N,\varepsilon}(t)$. By the Cauchy–Schwartz inequality we get that

$$
\begin{aligned}
(4.27)\quad & |\mathcal{R}_{N,\varepsilon}(t)|^2 \\
&\le 4\varepsilon^{2N(1-p)}\mathbb{E}|\mathbf{V}(0,\mathbf{0})|^2 \\
&\quad \times \max_{0\le s\le t\varepsilon^{-2}} \mathbb{E}\left|\int_0^s ds' \int_{\Delta_{N-1}(s,s')}\cdots\int \mathbb{E}_{s'}\mathbf{W}_N(s_1,\dots,s_N,s',\varepsilon\mathbf{x}(s'))ds_1\cdots ds_N\right|^2 .
\end{aligned}
$$

The stationarity of the Lagrangian velocity field implies that the right-hand side of (4.27) is equal to

$$
4\varepsilon^{2N(1-p)}\mathbb{E}|\mathbf{V}(0,\mathbf{0})|^2 \max_{0\le s\le t\varepsilon^{-2}} \mathbb{E}\left|\int_0^s ds' \int_{\Delta_N(s',0)}\cdots\int \mathbb{E}_0\mathbf{W}_N(s_1,\dots,s_N,0,\mathbf{0})ds_1\cdots ds_N\right|^2 . \tag{4.28}
$$

Subsequently using (4.16) for the multiple time integration of the conditional expectations in (4.28), we deduce that the above expression is less than or equal to

$$
\begin{aligned}
(4.29)\quad & 4C\varepsilon^{2N(1-p)}t^2\varepsilon^{4(1-2q)}\mathbb{E}\left|\sum_{\mathcal{F},\mathbf{n}}\int\cdots\int \varphi_N(\mathbf{x},\mathbf{k}_1,\dots,\mathbf{k}_N)P_N(\mathbf{k}_1,\dots,\mathbf{k}_N;\mathcal{F})\right. \\
&\left.\times \prod_{\widehat{mm'}\in E(\mathcal{F})} \mathbb{E}\left[\widehat{V}_{n_m}(0,d\mathbf{k}_m)\widehat{V}_{n_{m'}}(0,d\mathbf{k}_{m'})\right] \prod_{m\in A_N(\mathcal{F})\cup\{N+1\}} \widehat{V}_{n_m}(0,d\mathbf{k}_m)\right|^2 .
\end{aligned}
$$

The summation above extends over all Feynman diagrams $\mathcal{F}\in\mathfrak{S}_N$ and multi-indices $\mathbf{n}$.

By introducing an identical copy of the diagram which is supported on $\{N+2, N+2,\dots,2N+2\}$, the expression in (4.29) can be written in the form

$$
\begin{aligned}
& 4C\varepsilon^{2N(1-p)}t^2\varepsilon^{4(1-2q)}\sum\int\dots\int \varphi_N(\mathbf{0},\mathbf{k}_1,\dots,\mathbf{k}_N)\varphi_N(\mathbf{0},\mathbf{k}'_1,\dots,\mathbf{k}'_N) \\
&\qquad\qquad P_N(\mathbf{k}_1,\dots,\mathbf{k}_N;\mathcal{F})P_N(\mathbf{k}'_1,\dots,\mathbf{k}'_N;\mathcal{F}) \\
&\times\left|\prod_{\widehat{mm'}\in E(\mathcal{F})} \mathbb{E}\left[\widehat{V}_{n_m}(0,d\mathbf{k}_m)\widehat{V}_{n_{m'}}(0,d\mathbf{k}_{m'})\right]\right. \\
&\quad\left.\times \prod_{\widehat{mm'}\in E(\mathcal{F})} \mathbb{E}\left[\widehat{V}_{n_m}(0,d\mathbf{k}'_m)\widehat{V}_{n_{m'}}(0,d\mathbf{k}'_{m'})\right]\mathbb{E}\left[\prod_{m\in A^{(2)}(\mathcal{F})}\widehat{V}_{n_m}(0,d\mathbf{k}_m)\right]\right|,
\end{aligned}
$$

where $\mathbf{k}_{N+1+j}:=\mathbf{k}'_j$ and

$$
A^{(2)}(\mathcal{F}) = A_N(\mathcal{F})\cup\{N+1\}\cup\{j+N+1 : j\in A_N(\mathcal{F})\cup\{N+1\}\}. \tag{4.30}
$$

Using the elementary inequality

$$
\frac{1-\mathrm{e}^{-xt/\varepsilon^2}}{x}\le\frac{C}{x+\varepsilon^{2q}/t},\quad t,x>0,
$$

for a constant C independent of ε, x, we conclude that (see (4.18))

$$|P_N(\mathbf{k}_1,\dots,\mathbf{k}_N;\mathcal{F})| \le \prod_{j=1}^{N} \frac{\sum\limits_{m\in A_j(\mathcal{F})} |\mathbf{k}_m|}{\sum\limits_{m\in A_j(\mathcal{F})} |\mathbf{k}_m|^{2\beta} + \varepsilon^{2q}/t}.$$

The expression (4.29) can be now estimated by

$$\begin{aligned}(4.31)\qquad & Ct^2\varepsilon^{4(1-2q)} \sum \int_0^K \cdots \int_0^K Q_N(k_1,\dots,k_N;\mathcal{F})\, Q_N(k'_1,\dots,k'_N;\mathcal{F}) \\ & \times \prod_{\widehat{mm'}\in E(\mathcal{F})} \left[\frac{\delta(k_m - k_{m'})dk_m dk_{m'}}{k_m^{2\alpha-1}} \times \frac{\delta(k'_m - k'_{m'})dk_m dk_{m'}}{{k'_m}^{2\alpha-1}}\right] \\ & \times \prod_{\widehat{mm'}\in E(\mathcal{F}')} \frac{\delta(k_m - k_{m'})dk_m dk_{m'}}{k_m^{2\alpha-1}},\end{aligned}$$

with $k_m = |\mathbf{k}_m|$ and

$$Q_N(k_1,\dots,k_N;\mathcal{F}) := \prod_{j=1}^{N} \frac{\sum\limits_{m\in A_j(\mathcal{F})} k_m}{\sum\limits_{m\in A_j(\mathcal{F})} k_m^{2\beta} + \varepsilon^{2q}/t}.$$

The summation extends over all Feynman diagrams $\mathcal{F} \in \mathfrak{S}_N$ and all complete diagrams $\mathcal{F}'$ made of the vertexes of $A^{(2)}(\mathcal{F})$.

When $2\beta \le 1$, $Q_N(k_1,\dots,k_N;\mathcal{F})$ is bounded. This in turn implies that the expression (4.31) diverges at most at the rate $\varepsilon^{4(1-2q)}$. The estimate (4.27) implies then that $\mathcal{R}_{N,\varepsilon}(t)$ vanishes with $\varepsilon \downarrow 0$ and $N > 2/(1-p)$.

Let us assume therefore that $2\beta > 1$. There exists then a constant C, depending only on t, N, β, and K, such that

$$(4.32)\qquad \frac{\sum\limits_{m\in A_j(\mathcal{F})} k_m}{\sum\limits_{m\in A_j(\mathcal{F})} k_m^{2\beta} + \varepsilon^{2q}/t} \le C\frac{k_{m_j} + \varepsilon^{q/\beta}}{k_{m_j}^{2\beta} + \varepsilon^{2q}} \quad \forall m_j \in A_j(\mathcal{F})$$

and thus

$$(4.33)\qquad Q_N(k_1,\dots,k_N;\mathcal{F}) \le C\prod_{j=1}^{N} \frac{k_{m_j} + \varepsilon^{q/\beta}}{k_{m_j}^{2\beta} + \varepsilon^{2q}}$$

for all $m_j \in A_j(\mathcal{F})$. Hereby we make the following definite choice of m_j: let $m_j := j$ if j is *not* the right endpoint of an edge of the diagram $\mathcal{F}$. Otherwise, let m_j be the closest vertex from $\mathcal{A}(\mathcal{F})$ to the left of j.

Denote by $E'(\mathcal{F})$ the set of the edges of the diagram $\mathcal{F}$ with neither endpoint belonging to $\mathcal{A}(\mathcal{F})$ (see (4.22)) by cardinality of e'. In view of (4.25), (4.26), and the identity

$$(4.34)\qquad e'(\mathcal{F}) + \#[A(\mathcal{F}) \setminus A_N(\mathcal{F})] = e(\mathcal{F}),$$

the expression on the right-hand side of (4.33) can be written as

$$C\prod_{\widehat{mm'}\in E'(\mathcal{F})}\frac{k_m+\varepsilon^{q/\beta}}{k_m^{2\beta}+\varepsilon^{2q}}\times\prod_{m\in\mathcal{A}(\mathcal{F})\setminus A_N(\mathcal{F})}\left(\frac{k_m+\varepsilon^{q/\beta}}{k_m^{2\beta}+\varepsilon^{2q}}\right)^{q_m+1}\times\prod_{m\in A_N(\mathcal{F})}\left(\frac{k_m+\varepsilon^{q/\beta}}{k_m^{2\beta}+\varepsilon^{2q}}\right)^{q_m+1}. \tag{4.35}$$

From (4.31), (4.33), and (4.35) we conclude that

(4.36)

$$|\mathcal{R}_{N,\varepsilon}(t)|^2\le C\varepsilon^{2N(1-p)+4(1-2q)}\sum\prod_{m\in\mathcal{A}(\mathcal{F})\setminus A_N(\mathcal{F})}\left[\int_0^K\left(\frac{k+\varepsilon^{q/\beta}}{k^{2\beta}+\varepsilon^{2q}}\right)^{q_m+1}\frac{dk}{k^{2\alpha-1}}\right]^2\times\left[\int_0^K\frac{(k+\varepsilon^{q/\beta})dk}{(k^{2\beta}+\varepsilon^{2q})k^{2\alpha-1}}\right]^{2e'}\prod_{\widehat{mm'}\in E(\mathcal{F}')}\int_0^K\left(\frac{k+\varepsilon^{q/\beta}}{k^{2\beta}+\varepsilon^{2q}}\right)^{2+q_m+q_{m'}}\frac{dk}{k^{2\alpha-1}}.$$

Here the summation extends over all possible diagrams $\mathcal{F}$, $\mathcal{F}'$ as in (4.31). The meanings of q_m's related to the diagram $\mathcal{F}$ are the same as introduced in the previous section. We adopt also the convention that $q_{N+1}=q_{2N+2}=-1$ and $q_{N+1+m}:=q_m$.

4.2. Estimates for $\mathcal{I}_{n,\varepsilon}(t)$ for $n\ge 1$. The calculation is similar to that for the remainder term carried out in the previous section, so we shall sketch only the main points.

From (4.16) we infer that the i,jth entry of the matrix $\mathcal{I}_{n,\varepsilon}(t)$, given by (4.11), equals

$$2\varepsilon^{n(1-p)+2}\sum\int_0^{t\varepsilon^{-2q}}ds\int\cdots\int\varphi_{n+1}(\mathbf{k}_1,\dots,\mathbf{k}_{n+1})\times P_n(\mathbf{k}_1,\dots,\mathbf{k}_n;\mathcal{F})\left(\sum_{m\in A_{n+1}(\mathcal{F})}|\mathbf{k}_m|^{2\beta}+\varepsilon^{2q}/t\right)^{-1}\times\left|\prod_{\widehat{mm'}\in E(\mathcal{F})}\mathbb{E}\left[\widehat{V}_{l_m}(0,d\mathbf{k}_m)\widehat{V}_{l_{m'}}(0,d\mathbf{k}_{m'})\right]\mathbb{E}\left[\prod_{m\in A_{n+1}(\mathcal{F})\cup\{n+2\}}\widehat{V}_{l_m}(0,d\mathbf{k}_m)\right]\right|. \tag{4.37}$$

Here the summation extends over all multi-indices $\mathbf{l}=(l_1,\dots,l_{n+2})$ such that $l_1=i$, $l_{n+2}=j$, and all Feynman diagrams $\mathcal{F}\in\mathfrak{S}_{n+1}$. Proceeding with the same type of estimates as in the case of the remainder term we conclude that

$$|\mathcal{I}_{n,\varepsilon}(t)|\le Ct\varepsilon^{n(1-p)}\sum\prod_{m\in\mathcal{A}(\mathcal{F})\setminus A_{n+1}(\mathcal{F})}\int_0^K\left(\frac{\varepsilon^{\frac{q}{\beta}}+k}{\varepsilon^{2q}+k^{2\beta}}\right)^{q_m+1}\frac{dk}{k^{2\alpha-1}}\times\left[\int_0^K\frac{(\varepsilon^{\frac{q}{\beta}}+k)dk}{(\varepsilon^{2q}+k^{2\beta})k^{2\alpha-1}}\right]^{e'}\prod_{\widehat{mm'}\in\mathcal{F}'}\int_0^K\frac{(\varepsilon^{\frac{q}{\beta}}+k)^{2+q_m+q_{m'}-r_{m,m'}}}{(\varepsilon^{2q}+k^{2\beta})^{2+q_m+q_{m'}}}\times\frac{dk}{k^{2\alpha-1}}. \tag{4.38}$$

Here the summation extends over all Feynman diagrams $\mathcal{F}\in\mathfrak{S}_{n+1}$ and all complete diagrams $\mathcal{F}'$ made of the vertexes of $A_{n+1}(\mathcal{F})\cup\{n+2\}$; $r_{m,m'}:=\delta_{m,m_{n+1}}+\delta_{m',m_{n+1}}$. Also, we adopt convention $q_{n+2}=-1$.

5. Case 1. $\alpha+\beta<1$, $2p\beta<1$, $0\le p<1$—Brownian motion ($q=1$). We shall give the proof only in the case $2p\beta<1$. Also, for clarity we shall calculate only the asymptotic of the mean square displacement of $\mathbf{x}_\varepsilon(t)$, referring an interested reader to our paper [5], where the proof of the martingale version of our theorem has been laid out for $p=0$. A suitable adaptation of the proof to the case $p\in[0,1)$ and $2p\beta<1$ is possible along the lines of the argument we present below.

After an elementary calculation we deduce that under the assumption $\alpha+\beta<1$

$$\lim_{\varepsilon\downarrow 0}\mathcal{I}_{0,\varepsilon}(t)=\mathbf{D}t$$

with

$$\mathbf{D}=\int_{\mathbb{R}^d}\left(\mathbf{I}-\frac{\mathbf{k}\otimes\mathbf{k}}{|\mathbf{k}|^2}\right)\frac{a(|\mathbf{k}|)}{|\mathbf{k}|^{2\alpha+2\beta-1}}\,\frac{d\mathbf{k}}{|\mathbf{k}|^{d-1}},$$

provided that $q=1$.

Estimates for $\mathcal{R}_{N,\varepsilon}(t)$. We observe that

$$\int_0^K\frac{k+\varepsilon^{1/\beta}}{k^{2\beta}+\varepsilon^2}\,\frac{dk}{k^{2\alpha-1}}\le C, \tag{5.1}$$

$$\int_0^K\left(\frac{\varepsilon^{1/\beta}+k}{\varepsilon^2+k^{2\beta}}\right)^{q_m+1}\frac{dk}{k^{2\alpha-1}}\le C(1+\varepsilon^{\gamma(m)}), \tag{5.2}$$

$$\int_0^K\left(\frac{\varepsilon^{1/\beta}+k}{\varepsilon^2+k^{2\beta}}\right)^{2+q_m+q_{m'}}\frac{dk}{k^{2\alpha-1}}\le C(1+\varepsilon^{\gamma(\widehat{mm'})}), \tag{5.3}$$

with

$$\gamma(m):=\frac{1}{\beta}[2-2\alpha+(q_m+1)(1-2\beta)], \tag{5.4}$$

$$\gamma(\widehat{mm'}):=\frac{1}{\beta}[2-2\alpha+(q_m+q_{m'}+2)(1-2\beta)]. \tag{5.5}$$

We conclude therefore that

$$|\mathcal{R}_{N,\varepsilon}(t)|^2\le C\varepsilon^{\mu}, \tag{5.6}$$

with

$$\mu:=2N(1-p)-4+\kappa, \tag{5.7}$$

$$\kappa:=\frac{1}{\beta}\Big[2f'(2-\alpha-2\beta)+2f''(3-2\alpha-2\beta) +(1-2\beta)\sum{}'(q_m+q_{m'})+2(1-2\beta)\sum{}''q_m\Big], \tag{5.8}$$

where the summation $\sum'$ extends over the edges $\widehat{mm'}$ of the diagram $\mathcal{F}'$ for which $\gamma(\widehat{mm'})<0$ and $\sum''$ extends over the vertexes m of $\mathcal{A}(\mathcal{F})\setminus A_N(\mathcal{F})$, for which $\gamma(m)<0$ (see (5.4), (5.5)) and f',f'' denote the cardinalities of the respective sets of edges and vertexes. Obviously,

$$f'\le c,\quad f''\le e-e', \tag{5.9}$$

with c the cardinality of $A_N(\mathcal{F})$ and e the number of edges of $\mathcal{F}$ (cf. (4.34)). Note that $c + 2e = N$.

Using $\sum'(q_m + q_{m'}) + 2\sum'' q_m \le 2e$ and $2\beta > 1$, we can write that

$$\kappa \ge \frac{2}{\beta}\left[f'(2-\alpha-2\beta) + f''(3-2\alpha-2\beta) + e(1-2\beta)\right]. \tag{5.10}$$

Since $N = c + 2e$ we conclude from (5.9) and (5.10) that

$$\begin{aligned}\mu &\ge -4 + 2N(1-p) + \frac{2}{\beta}\left[f'(2-\alpha-2\beta) + f''(3-2\alpha-2\beta) + e(1-2\beta)\right] \\ &\ge -4 + 2(c-f')(1-p) \\ &\quad + \frac{2}{\beta}\left\{f'[2-\alpha-(1+p)\beta] + f''(3-2\alpha-2\beta) + e(1-2p\beta)\right\} > 0,\end{aligned} \tag{5.11}$$

provided that $2p\beta < 1$ (note that then necessarily $2-\alpha-(1+p)\beta > 0$) and N is sufficiently large.

Estimates for $\mathcal{I}_{n,\varepsilon}(t)$ for $n \ge 1$. Using (5.1)–(5.2) and

$$\int_0^K \frac{(k+\varepsilon^{\frac{1}{\beta}})^{2+q_m+q_{m'}-r_{m,m'}}}{(k^{2\beta}+\varepsilon^2)^{2+q_m+q_{m'}}}\,\frac{dk}{k^{2\alpha-1}} \le C(1+\varepsilon^{\tilde{\gamma}(\widehat{mm'})}),$$

with

$$\tilde{\gamma}(\widehat{mm'}) := \frac{1}{\beta}[2-2\alpha+(q_m+q_{m'}+2)(1-2\beta)-r_{m,m'}], \tag{5.12}$$

we conclude that

$$|\mathcal{I}_{n,\varepsilon}(t)| \le C\varepsilon^{\mu}, \tag{5.13}$$

where $\mu = n(1-p)+\kappa$ and

$$\begin{aligned}\kappa := \frac{1}{\beta}\Big[&2f'(2-\alpha-2\beta) + f''(3-2\alpha-2\beta) \\ &+(1-2\beta)\sum{}'(q_m+q_{m'}) + (1-2\beta)\sum{}'' q_m - 1\Big].\end{aligned}$$

The summation $\sum'$ extends over the edges $\widehat{mm'}$ of the diagram $\mathcal{F}'$ for which $\tilde{\gamma}(\widehat{mm'}) < 0$ and $\sum''$ extends over the vertexes m of $\mathcal{A}(\mathcal{F}) \setminus A_{n+1}(\mathcal{F})$ for which $\gamma(m) < 0$. f', f'' denote the cardinalities of the respective sets of edges and vertexes. Finally, obtain that

$$\begin{aligned}\mu &\ge (c+1-2f'+e)(1-p) \\ &\quad + \frac{1}{\beta}\left[2p\beta + 2f'(2-\alpha-(1+p)\beta) + f''(3-2\alpha-2\beta) + e(1-2p\beta)\right] > 0.\end{aligned}$$

In conclusion, we proved that the utmost left-hand side of (4.10) tends to $\mathbf{D}t$ as $\varepsilon \downarrow 0$, provided that $\alpha + \beta < 1$.

6. Case 2. $1 < \alpha + \beta < 1 + 1/p$, $0 \le p < 1$—FBM. For $\alpha + \beta > 1$, it is straightforward to check that

$$\lim_{\varepsilon \downarrow 0} \mathcal{I}_{0,\varepsilon} = \mathbf{D} t^{2H},$$

provided that $q = \beta/(\alpha + 2\beta - 1)$. Here

$$\mathbf{D} = \int_{\mathbb{R}^d} \frac{e^{-|\mathbf{k}|^{2\beta}} - 1 + |\mathbf{k}|^{2\beta}}{|\mathbf{k}|^{2\alpha+4\beta-1}} \left(\mathbf{I} - \frac{\mathbf{k} \otimes \mathbf{k}}{|\mathbf{k}|^2} \right) \frac{a(0)}{|\mathbf{k}|^{d-1}} d\mathbf{k}$$

and the Hurst exponent H is given by

$$1/2 < H = 1/2 + \frac{\alpha + \beta - 1}{2\beta} < 1.$$

6.1. Case 2a. We assume that

$$3/2 < \alpha + \beta, \quad \alpha + 2\beta < 1 + 1/p, \quad 0 \le p < 1.$$

We shall only carry out the estimates of $\mathcal{R}_{N,\varepsilon}(t)$. One can easily obtain the respective estimates of $\mathcal{I}_{n,\varepsilon}(t)$. These estimates are very similar to the corresponding part of section 5. We use the notation introduced there.

As before we need only to consider the case $2\beta > 1$ (cf. (4.36)). Note that

$$\int_0^K \frac{(\varepsilon^{q/\beta} + k) dk}{(\varepsilon^{2q} + k^{2\beta}) k^{2\alpha-1}} \le C(1 + \varepsilon^{\gamma}),$$

$$\int_0^K \left(\frac{\varepsilon^{q/\beta} + k}{\varepsilon^{2q} + k^{2\beta}} \right)^{q_m+1} \frac{dk}{k^{2\alpha-1}} \le C(1 + \varepsilon^{\gamma(m)}),$$

and

$$\int_0^K \left(\frac{\varepsilon^{q/\beta} + k}{\varepsilon^{2q} + k^{2\beta}} \right)^{2+q_m+q_{m'}} \times \frac{dk}{k^{2\alpha-1}} \le C(1 + \varepsilon^{\gamma(\widehat{mm'})}),$$

with

$$\gamma := \frac{3 - 2\alpha - 2\beta}{\alpha + 2\beta - 1}, \tag{6.1}$$

$$\gamma(m) := \frac{3 - 2\alpha - 2\beta + q_m(1 - 2\beta)}{\alpha + 2\beta - 1}, \tag{6.2}$$

$$\gamma(\widehat{mm'}) := \frac{4 - 2\alpha - 4\beta + (q_m + q_{m'})(1 - 2\beta)}{\alpha + 2\beta - 1} \tag{6.3}$$

(cf. (5.4)–(5.5)).

Estimating the same way as in (5.6)–(5.11) we obtain

$$|\mathcal{R}_{N,\varepsilon}(t)|^2 \le C t^4 \varepsilon^{\mu},$$

with

$$\mu := 2N(1-p) + 4(1-2q) + \kappa, \tag{6.4}$$

$$\kappa := 2(e' + f'')\frac{3-2\alpha-2\beta}{\alpha+2\beta-1}$$

$$+\frac{1}{\alpha+2\beta-1}\left[2f'(2-\alpha-2\beta) + (1-2\beta)\sum{}'(q_m + q_{m'}) + 2(1-2\beta)\sum{}'' q_m\right]$$

(cf. (5.9)). We have

$$\begin{aligned}\mu &\geq 4(1-2q) + 2N(1-p) + 2(e'+f'')\frac{3-2\alpha-2\beta}{\alpha+2\beta-1}\\ &\quad + \frac{2}{\alpha+2\beta-1}\,[f'(2-\alpha-2\beta) + e(1-2\beta)]\\ &\geq 4(1-2q) + 2(c+2e)(1-p) + 2e\frac{3-2\alpha-2\beta}{\alpha+2\beta-1}\\ &\quad + \frac{2}{\alpha+2\beta-1}\,[c(2-\alpha-2\beta) + e(1-2\beta)]\\ &\geq 4(1-2q) + 2Np\,\frac{1+1/p-\alpha-2\beta}{\alpha+2\beta-1} > 0,\end{aligned}$$

provided that N is sufficiently large. This in turn implies that $|\mathcal{R}_{N,\varepsilon}(t)|^2$ vanishes as $\varepsilon \downarrow 0$ for such a choice of N.

6.2. Case 2b. Here we assume that

$$1 < \alpha+\beta < 3/2, \quad \alpha+2\beta < 1 + 1/(2p) + (\alpha+\beta-1)/p, \quad 0 \leq p < 1.$$

In this case one can write κ in (6.4) as

$$\begin{aligned}\kappa &= 2f''\,\frac{3-2\alpha-2\beta}{\alpha+2\beta-1} + \frac{1}{\alpha+2\beta-1}\\ &\quad\times\left[2f'(2-\alpha-2\beta) + (1-2\beta)\sum{}'(q_m+q_{m'}) + 2(1-2\beta)\sum{}'' q_m\right]\end{aligned}$$

and hence

$$\begin{aligned}\mu &\geq 4(1-2q) + 2(c+2e)(1-p)\\ &\quad + 2f''\,\frac{3-2\alpha-2\beta}{\alpha+2\beta-1} + \frac{2}{\alpha+2\beta-1}\,[f'(2-\alpha-2\beta) + e(1-2\beta)]\\ &\geq 4(1-2q) + 2ep\,\frac{(\alpha+\beta-1)/p + 1/(2p) - \alpha - 2\beta}{\alpha+2\beta-1} + 2f''\,\frac{3-2\alpha-2\beta}{\alpha+2\beta-1}\\ &\quad + 2(c-f')(1-p) + 2f'p\,\frac{1+1/p-\alpha-2\beta}{\alpha+2\beta-1} > 0,\end{aligned}$$

provided that N is sufficiently large. This in turn implies that $|\mathcal{R}_{N,\varepsilon}(t)|^2$ vanishes as $\varepsilon \downarrow 0$ for such a choice of N.

7. FBM limit with $p \geq 1$: Heuristics. In this section we give an argument indicating that the FBM limit holds for $p > 1$. The argument is similar to the one given in [8].

Let $\mathbf{U}^\varepsilon(t, \mathbf{x})$ be the Gaussian velocity with energy spectrum given by

$$\mathcal{E}_\varepsilon(k) = \frac{a(\varepsilon^p k)}{k^{2\alpha-1}}. \tag{7.1}$$

Then it follows from the spectral representation of the velocity correlation function that $\mathbf{U}^\varepsilon$ is related to $\mathbf{V}$ via

$$\mathbf{V}\left(\frac{t}{\varepsilon^{2q}}, \frac{\mathbf{x}}{\varepsilon^p}\right) = \varepsilon^{p(1-\alpha)}\mathbf{U}^\varepsilon\left(\frac{t}{\varepsilon^{2(q-p\beta)}}, \mathbf{x}\right).$$

With a unique pair of parameters q, η_ε,

$$q = \beta/(\alpha + 2\beta - 1), \quad \eta_\varepsilon = \varepsilon^{1+p-p(\alpha+2\beta)}, \tag{7.2}$$

the equation of motion can be written as

$$\frac{d\mathbf{x}^\varepsilon(t)}{dt} = \frac{1}{\eta_\varepsilon^{2q-1}}\mathbf{U}^\varepsilon\left(\frac{t}{\eta_\varepsilon^{2q}}, \mathbf{x}^\varepsilon(t)\right). \tag{7.3}$$

Since η_ε must tend to zero we require that

$$\alpha + 2\beta < 1 + \frac{1}{p}. \tag{7.4}$$

Condition (7.4) is also related to the fact that the velocity $\mathbf{U}^\varepsilon$ has increasingly smaller scales as ε tends to zero.

The following physical argument shows that, under the conditions (7.4) and

$$\alpha + \beta > 1,$$

the ultraviolet divergence in $\mathbf{U}^\varepsilon$ has no physical significance. The small-scale velocity associated with high wave number $|\mathbf{k}|$ has the amplitude

$$\left(\int_{c_1|\mathbf{k}| \leq |\mathbf{k}'| \leq c_2|\mathbf{k}|} \mathcal{E}(\mathbf{k}')d|\mathbf{k}'|\right)^{1/2} \sim |\mathbf{k}|^{1-\alpha}, \quad |\mathbf{k}| \gg 1,$$

and the correlation time is of the order $|\mathbf{k}|^{-2\beta}$. Then particles transported by small-scale velocity travel a distance less than or equal to the sum of $t|\mathbf{k}|^{2\beta}$ number uncorrelated random variables of magnitude $|\mathbf{k}|^{1-\alpha}|\mathbf{k}|^{-2\beta}$. Thus, on the time scale $t \sim \eta_\varepsilon^{-2q}$, the displacement caused by high wave number $\mathbf{k}$ is of the order less than or equal to $\sqrt{\eta_\varepsilon^{-2q}|\mathbf{k}|^{2\beta}}|\mathbf{k}|^{1-\alpha-2\beta}$, as suggested by the turbulent diffusion limit theorem for mixing flows [4], which equals $\eta_\varepsilon^{-q}|\mathbf{k}|^{1-\alpha-\beta}$ and is always smaller than η_ε^{-1} (the spatial scale of observation) if $\alpha + \beta > 1$ and $q < 1$ (superdiffusive scaling). With (7.2) the two conditions ($\alpha + \beta > 1$ and $q < 1$) are equivalent. It is clear that for $|\mathbf{k}| = O(1)$ the previous argument is still valid.

Now, if we neglect the high wave numbers in (7.3) the equation becomes

$$d\boldsymbol{x}^\varepsilon(t)/dt = \eta_\varepsilon^{1-2q}\mathbf{V}(t/\eta_\varepsilon^{2q}, \boldsymbol{x}^\varepsilon(t)), \tag{7.5}$$

which has the asymptotic solution

$$\boldsymbol{x}^{\varepsilon}(0) + \eta_{\varepsilon} \int_0^{t/\eta_{\varepsilon}^{2q}} \mathbf{V}(\boldsymbol{x}^{\varepsilon}(0), s) ds \tag{7.6}$$

converging to an FBM (Theorems 1 and 2).

REFERENCES

[1] G. Dagan, *Flow and Transport in Porous Formations*, Springer-Verlag, Berlin, New York, 1989.

[2] A. Fannjiang, *Phase diagram for turbulent transport: Sampling drift, eddy diffusivity and variational principles*, Phy. D, 136 (2000), pp. 145–174.

[3] A. Fannjiang, *Erratum to: "Phase diagram for turbulent transport: sampling drift, eddy diffusivity and variational principle" [Phy. D* 136 (2000)*, no.* 1-2, 145–147*]*, Phys. D, 157 (2001), pp. 166–168.

[4] A. Fannjiang and T. Komorowski, *Turbulent diffusion in Markovian flows*, Ann. Appl. Probab., 9 (1999), pp. 591–610.

[5] A. Fannjiang and T. Komorowski, *Diffusion approximation for particle convection in Markovian flows*, Bull. Polish Acad. Sci. Math., 48 (2000), pp. 253–275.

[6] A. Fannjiang and T. Komorowski, *The fractional Brownian motion limit for turbulent transport*, Ann. Appl. Probab., 10 (2000), pp. 1100–1120.

[7] A. Fannjiang and T. Komorowski, *Fractional Brownian motions and enhanced diffusion in a unidirectional wave-like turbulence*, J. Statist. Phys., 100 (2000), pp. 1071–1095.

[8] A. Fannjiang and T. Komorowski, *Diffusive and nondiffusive limits of transport in nonmixing flows*, SIAM J. Appl. Math., 62 (2002), pp. 909–923.

[9] J. Glimm and A. Jaffe, *Quantum Physics*, Springer-Verlag, New York, 1981.

[10] S. Janson, *Gaussian Hilbert Spaces*, Cambridge University Press, Cambridge, UK, 1997.

[11] H. Kesten and G. C. Papanicolaou, *A limit theorem for turbulent diffusion*, Comm. Math. Phys., 65 (1979), pp. 97–128.

[12] J. Klafter, M. F. Shlesinger, and G. Zumofen, *Beyond Brownian motion*, Phys. Today, 49 (1993), p. 33.

[13] T. Komorowski, *Diffusion approximation for the advection of particles in a strongly turbulent random environment*, Ann. Probab., 24 (1996), pp. 346–376.

[14] T. Komorowski, *Application of the parametrix method to diffusions in a turbulent Gaussian environment*, Stochastic Process Appl., 74 (1998), pp. 165–193.

[15] T. Komorowski and S. Olla, *On the Superdiffusive Behavior or Passive Tracer with a Gaussian Drift*, preprint, 2002.

[16] H. H. Kuo, *White Noise Distribution Theory*, CRC Press, Boca Raton, FL, 1996.

[17] B. B. Mandelbrot and J. W. Van Ness, *Fractional Brownian motions, fractional noises and applications*, SIAM Rev., 10 (1968), pp. 422–437.

[18] A. J. Majda and P. R. Kramer, *Simplified models for turbulent diffusion: Theory, numerical modeling, and physical phenomena*, Phys. Rep., 314 (1999), pp. 237–574.

[19] R. Metzler and J. Klafter, *The random walk's guide to anomalous diffusion: A fractional dynamics approach*, Phys. Rep., 339 (2000), pp. 1–77.

[20] M. Rosenblatt, *Markov Processes. Structure and Asymptotic Behavior*, Springer-Verlag, Berlin, Heidelberg, New York, 1971.

[21] M. F. Shlesinger, G. Zaslavsky, and J. Klafter, *Strange kinetics*, Nature, 363 (1993), pp. 31–37.

[22] G. I. Taylor, *Diffusions by continuous movements*, Proc. London Math. Soc. Ser. 2, 20 (1923), pp. 196–211.

SIAM J. APPL. MATH.
Vol. 63, No. 6, pp. 2063–2086

STABILITY AND TRAVELING FRONTS IN LOTKA–VOLTERRA COMPETITION MODELS WITH STAGE STRUCTURE*

J. F. M. AL-OMARI† AND S. A. GOURLEY†

Abstract. This paper is concerned with a delay differential equation model for the interaction between two species, the adult members of which are in competition. The competitive effects are of the Lotka–Volterra kind, and in the absence of competition it is assumed that each species evolves according to the predictions of a simple age-structured model which reduces to a single equation for the total adult population. For each of the two species the model incorporates a time delay which represents the time from birth to maturity of that species. Thus, the time delays appear in the adult recruitment terms.

The dynamics of the model are determined, and global stability results are established for each equilibrium. The equilibria of the model involve the maturation delays. The criteria for global convergence to each equilibrium are sharp and involve these delays.

A reaction-diffusion extension of the model is also studied for the case when only the adult members of each species can diffuse. We prove the existence of a traveling front solution connecting the two boundary equilibria for the case when there is no coexistence equilibrium. This represents invasion by the stronger species of territory previously inhabited only by the weaker. The proof of the existence of such a front uses Wu and Zou's theory for traveling front solutions of delayed reaction-diffusion systems.

Key words. competition, stage structure, time delay, global stability, reaction-diffusion, traveling front

AMS subject classifications. 92D25, 34D23, 35K57

DOI. 10.1137/S0036139902416500

1. Introduction.

This paper is concerned with the following delayed Lotka–Volterra-type model for the adult members of two species U and V in competition,

$$\begin{aligned} \frac{dU(t)}{dt} &= \alpha_u \int_0^\infty f_u(s) e^{-\gamma_u s} U(t-s)\, ds - \beta_u U^2(t) - c_1 U(t) V(t), \\ \frac{dV(t)}{dt} &= \alpha_v \int_0^\infty f_v(s) e^{-\gamma_v s} V(t-s)\, ds - \beta_v V^2(t) - c_2 U(t) V(t) \end{aligned} \tag{1.1}$$

and also with a reaction-diffusion extension thereof. In proposing the system (1.1), it is assumed that competition effects are of the classical Lotka–Volterra kind. It is also assumed that the adult numbers of each species, in the absence of the other species, evolve according to a delay equation of the form

$$\frac{du(t)}{dt} = \alpha \int_0^\infty f(s) e^{-\gamma s} u(t-s)\, ds - \beta u^2(t). \tag{1.2}$$

Equation (1.2) can be regarded as a generalization of the second equation of the system

$$\begin{aligned} u_i'(t) &= \alpha u_m(t) - \gamma u_i(t) - \alpha\, e^{-\gamma\tau} u_m(t-\tau), \\ u_m'(t) &= \alpha\, e^{-\gamma\tau} u_m(t-\tau) - \beta u_m^2(t), \end{aligned} \tag{1.3}$$

*Received by the editors October 23, 2002; accepted for publication (in revised form) March 21, 2003; published electronically September 17, 2003.
http://www.siam.org/journals/siap/63-6/41650.html
†Department of Mathematics and Statistics, University of Surrey, Guildford, Surrey GU2 7XH, United Kingdom (s.gourley@surrey.ac.uk).

which was proposed by Aiello and Freedman [1] as a model of a single species, where u_i denotes the number of sexually immature members of the species and u_m the number of mature adult members. Note that, in system (1.3), the second equation is uncoupled from the first and thus it is sufficient to consider this second equation on its own. The parameter τ in (1.3) measures the time from birth to maturity, and the $e^{-\gamma\tau}$ terms allow for the fact that not all immatures survive to maturity (γ in (1.3) is the death rate of immatures, while β measures deaths of matures). An assumption behind (1.3) is that all individuals take the same amount of time τ to become mature. If we instead allow the possibility that an individual could become mature at any age, and if we denote by $f(s)\,ds$ the probability that the maturation time is between s and $s + ds$ with ds infinitesimal, and $\int_0^\infty f(s)\,ds = 1$, then simple modeling leads to (1.2) as a generalization of the second equation of (1.3). Of course, $f(s)$ will be small when s is small, and when s is large, and it is quite reasonable to take $f(s) = 0$ for all s above some finite value, as we shall do in much of the paper.

Thus, in our model (1.1), we assume that each species on its own grows not according to the logistic equation or delayed logistic equation but according to the predictions of the simple stage structured model (1.2) which is arguably more realistic as a model of a single species. The c_i terms in (1.1) are the competition effects; c_1 and c_2 measure the competitive effect of V on U, and U on V, respectively. Of course, U and V in (1.1) refer only to the *adult* numbers of the two species. Thus, it is assumed that competition occurs only between the adults. Since many species strongly protect their young, we feel this is not too unrealistic an assumption.

Competition systems with time delays have been studied by many authors. For example, Gopalsamy [3] studied the two-species delayed competition system

$$\begin{aligned} du/dt &= u\left(\gamma_1 - a_1 u - b_1 \int_{-T}^{0} K_1(s) v(t+s)\,ds\right), \\ dv/dt &= v\left(\gamma_2 - a_2 \int_{-T}^{0} K_2(s) u(t+s)\,ds - b_2 v\right), \end{aligned}$$

in which the delays are in the interspecies competition terms, and established a result on the global stability of the coexistence equilibrium, showing that when the intraspecies competition is stronger than the interspecies competition, nonconstant oscillatory solutions are not possible.

In the present paper we will show that the global dynamics of our system (1.1) can be completely determined, except in the case when the boundary equilibria $\hat{E}_u$ and $\hat{E}_v$ are both linearly stable. In this case one expects that the outcome of the competition will depend on the initial conditions. For other cases, the possibilities can be enumerated in a similar way to the classical Lotka–Volterra competition model without delay, as described in the book by Murray [6]. If one boundary equilibrium is linearly stable and the other unstable, we show that solutions converge globally to the stable boundary equilibrium. When both boundary equilibria are linearly unstable, we show that solutions converge globally to the coexistence equilibrium. Thus, the conditions for global stability are sharp and can be interpreted ecologically. Furthermore, the conditions involve the maturation delay kernels $f_u(s)$ and $f_v(s)$, and thus the role of these maturation delays can be elucidated. This is important since the comparison approach we adopt in this paper does not always, in other applications, elucidate clearly the role of the time delays and tends rather to furnish conditions which are merely sufficient for convergence and only involve the other parameters of the model.

It is straightforward to show that the solutions of system (1.1), subject to (2.8) below, satisfy $U(t), V(t) > 0$ on $(0, \infty)$. This fact is important for both the modeling and the analysis.

2. Equilibria and their stability. System (1.1) has four equilibria, determined by setting $dU/dt = dV/dt = 0$ in (1.1), and these are

$$E_0 = (0,0), \qquad \hat{E}_u = \left(\frac{\alpha_u}{\beta_u}\int_0^\infty f_u(s)e^{-\gamma_u s}ds, 0\right), \qquad \hat{E}_v = \left(0, \frac{\alpha_v}{\beta_v}\int_0^\infty f_v(s)e^{-\gamma_v s}ds\right),$$

and

$$\hat{E} = (\hat{U}, \hat{V}),$$

where

$$\hat{U} = \frac{\beta_v\alpha_u\int_0^\infty f_u(s)e^{-\gamma_u s}ds - c_1\alpha_v\int_0^\infty f_v(s)e^{-\gamma_v s}ds}{\beta_u\beta_v - c_1c_2}$$

and

$$\hat{V} = \frac{\beta_u\alpha_v\int_0^\infty f_v(s)e^{-\gamma_v s}ds - c_2\alpha_u\int_0^\infty f_u(s)e^{-\gamma_u s}ds}{\beta_u\beta_v - c_1c_2}$$

provided $\hat{U}, \hat{V} > 0$. Of course, the feasibility of the fourth equilibrium $\hat{E}$ depends on the parameters. As we shall show, it is feasible if either (i) the boundary equilibria $\hat{E}_u$ and $\hat{E}_v$ are both linearly unstable, or (ii) the boundary equilibria are both linearly stable.

We will begin by investigating the linearized stability of each equilibrium. Starting with $E_0 = (0,0)$, the linearization of (1.1) about this equilibrium is

$$\begin{aligned} \frac{dU(t)}{dt} &= \alpha_u\int_0^\infty f_u(s)e^{-\gamma_u s}U(t-s)\,ds, \\ \frac{dV(t)}{dt} &= \alpha_v\int_0^\infty f_v(s)e^{-\gamma_v s}V(t-s)\,ds. \end{aligned} \tag{2.1}$$

The characteristic equation resulting from (2.1) is

$$\left(\lambda - \alpha_u\int_0^\infty f_u(s)e^{-s(\gamma_u+\lambda)}ds\right)\left(\lambda - \alpha_v\int_0^\infty f_v(s)e^{-s(\gamma_v+\lambda)}ds\right) = 0,$$

the roots of which are the zeros of the first and the second bracketed factors, and in each of these the existence of a real positive root λ can be seen by plotting against λ the graphs of $y = \lambda$, $y = \alpha_u\int_0^\infty f_u(s)e^{-s(\gamma_u+\lambda)}ds$, and $y = \alpha_v\int_0^\infty f_v(s)e^{-s(\gamma_v+\lambda)}ds$. Therefore, $(0,0)$ is linearly unstable.

The linearization of (1.1) about $\hat{E}_u$ is

$$\begin{aligned} \frac{dU}{dt} &= \alpha_u\int_0^\infty f_u(s)e^{-\gamma_u s}U(t-s)\,ds - 2\alpha_u U(t)\int_0^\infty f_u(s)e^{-\gamma_u s}\,ds \\ &\quad -\frac{c_1\alpha_u}{\beta_u}V(t)\int_0^\infty f_u(s)e^{-\gamma_u s}ds, \\ \frac{dV}{dt} &= \alpha_v\int_0^\infty f_v(s)e^{-\gamma_v s}V(t-s)\,ds - \frac{c_2\alpha_u}{\beta_u}V(t)\int_0^\infty f_u(s)e^{-\gamma_u s}\,ds. \end{aligned} \tag{2.2}$$

The characteristic equation resulting from (2.2) is

$$\left(\lambda + c_2\alpha_u\beta_u^{-1}\int_0^\infty f_u(s)e^{-\gamma_u s}\,ds - \alpha_v\int_0^\infty f_v(s)e^{-s(\gamma_v+\lambda)}ds\right)$$
$$\times\left(\lambda + 2\alpha_u\int_0^\infty f_u(s)e^{-\gamma_u s}\,ds - \alpha_u\int_0^\infty f_u(s)e^{-s(\gamma_u+\lambda)}\,ds\right) = 0.$$

The eigenvalues of the linearization about $\hat{E}_u$ are therefore the roots λ of the equation

$$\lambda + 2\alpha_u\int_0^\infty f_u(s)e^{-\gamma_u s}ds = \alpha_u\int_0^\infty f_u(s)e^{-s(\gamma_u+\lambda)}\,ds \tag{2.3}$$

together with the roots λ of the equation

$$\lambda + c_2\alpha_u\beta_u^{-1}\int_0^\infty f_u(s)e^{-\gamma_u s}\,ds = \alpha_v\int_0^\infty f_v(s)e^{-s(\gamma_v+\lambda)}ds. \tag{2.4}$$

It is not difficult to see that all the roots of (2.3) satisfy $\operatorname{Re}\lambda < 0$. We shall now find the condition which determines that all roots of (2.4) satisfy $\operatorname{Re}\lambda < 0$. Assume, for contradiction, that there exists a root λ^* of (2.4) such that $\operatorname{Re}\lambda^* \geq 0$. Then

$$\begin{aligned}\left|\lambda^* + c_2\alpha_u\beta_u^{-1}\int_0^\infty f_u(s)e^{-\gamma_u s}\,ds\right| &= \left|\alpha_v\int_0^\infty f_v(s)e^{-s(\gamma_v+\lambda)}ds\right| \\ &\leq \alpha_v\int_0^\infty f_v(s)e^{-s\gamma_v}|e^{-s\lambda^*}|ds \\ &= \alpha_v\int_0^\infty f_v(s)e^{-s\gamma_v}e^{-s\operatorname{Re}\lambda^*}ds \\ &\leq \alpha_v\int_0^\infty f_v(s)e^{-s\gamma_v}ds,\end{aligned}$$

since $\operatorname{Re}\lambda^* \geq 0$. This implies that λ^* is in the circle in the complex λ plane centered at $\lambda = -c_2\alpha_u\beta_u^{-1}\int_0^\infty f_u(s)e^{-\gamma_u s}\,ds$ and of radius $\alpha_v\int_0^\infty f_v(s)e^{-s\gamma_v}ds$. Accordingly, we shall have a contradiction if

$$c_2\alpha_u\int_0^\infty f_u(s)e^{-\gamma_u s}\,ds > \beta_u\alpha_v\int_0^\infty f_v(s)e^{-\gamma_v s}ds. \tag{2.5}$$

Therefore, if condition (2.5) holds, then the equilibrium $\hat{E}_u$ is linearly stable.

In a similar way, we can show that $\hat{E}_v$ is linearly stable if

$$c_1\alpha_v\int_0^\infty f_v(s)e^{-\gamma_v s}\,ds > \beta_v\alpha_u\int_0^\infty f_u(s)e^{-\gamma_u s}ds. \tag{2.6}$$

Thus, if (2.5) and (2.6) both hold, then $\hat{E}_u$ and $\hat{E}_v$ are both linearly stable, and the numerators of the components $\hat{U}$, $\hat{V}$ of the equilibrium $\hat{E}$ are both negative. But, at the same time, (2.5) and (2.6) imply that

$$\beta_v < \frac{c_1\alpha_v\int_0^\infty f_v(s)e^{-\gamma_v s}\,ds}{\alpha_u\int_0^\infty f_u(s)e^{-\gamma_u s}ds}, \qquad \beta_u < \frac{c_2\alpha_u\int_0^\infty f_u(s)e^{-\gamma_u s}\,ds}{\alpha_v\int_0^\infty f_v(s)e^{-\gamma_v s}ds}$$

so that $\beta_v\beta_u < c_1c_2$, i.e., the denominators of $\hat{U}, \hat{V}$ are negative too. Thus, under these circumstances, $\hat{U}, \hat{V} > 0$ so that the equilibrium $\hat{E}$ is feasible.

In a similar way, we can see that if (2.5) and (2.6) are both reversed, then the boundary equilibria $\hat{E}_u$ and $\hat{E}_v$ are both linearly unstable and again $\hat{E}$ is feasible under these circumstances. However, if one of the boundary equilibria $\hat{E}_u$, $\hat{E}_v$ is stable and the other unstable, then the coexistence equilibrium $\hat{E}$ is not feasible.

In the next two sections, we shall prove theorems on the global asymptotic stability of the equilibria $\hat{E}_u$, $\hat{E}_v$, and $\hat{E}$ for the case when the kernels $f_u(s)$, $f_v(s)$ have compact support, that is, $f_u(s) = f_v(s) = 0$ for all $s \geq \tau$, for some $\tau > 0$, and normalized such that $\int_0^\tau f_u(s)\,ds = \int_0^\tau f_v(s)\,ds = 1$. In this case the system (1.1) becomes

$$\begin{aligned} \frac{dU(t)}{dt} &= \alpha_u \int_0^\tau f_u(s)e^{-\gamma_u s}U(t-s)\,ds - \beta_u U^2(t) - c_1 U(t)V(t), \\ \frac{dV(t)}{dt} &= \alpha_v \int_0^\tau f_v(s)e^{-\gamma_v s}V(t-s)\,ds - \beta_v V^2(t) - c_2 U(t)V(t). \end{aligned} \tag{2.7}$$

For initial data, we assume that

$$U(t), V(t) \geq 0 \text{ for } -\tau \leq t \leq 0, \quad \text{with} \quad U(0), V(0) > 0. \tag{2.8}$$

Before proceeding, we shall need the following theorem.

THEOREM 1. *Let $u(t)$ be the solution of*

$$\frac{du(t)}{dt} = \alpha_u \int_0^\tau f_u(s)e^{-\gamma_u s}u(t-s)\,ds - \beta_u u^2(t) - A\,u(t), \tag{2.9}$$

where $u(t) > 0$ for $-\tau \leq t \leq 0$. If

$$0 \leq A < \alpha_u \int_0^\tau f_u(s)e^{-\gamma_u s}\,ds,$$

then $\lim_{t\to\infty} u(t) = \hat{u}$, where

$$\hat{u} = \frac{1}{\beta_u}\left[\alpha_u \int_0^\tau f_u(s)e^{-\gamma_u s}\,ds - A\right]. \tag{2.10}$$

Proof. Let us first deal with the cases when $u(t)$ is eventually monotonically decreasing or eventually monotonically increasing. In the former case, positivity of solutions immediately yields that $u(t)$ must approach some limit $\hat{u} \geq 0$. This limit must be an equilibrium of (2.9) and is therefore either zero or the value stated. Zero is ruled out since a standard linearized analysis yields that the zero solution of (2.9) is linearly unstable under the stated condition on A.

If $u(t)$ is eventually monotonically increasing, then, when t is sufficiently large, $u(t-s) \leq u(t)$ for all $s \in [0, \tau]$ so that

$$\frac{du(t)}{dt} \leq \alpha_u u(t) \int_0^\tau f_u(s)e^{-\gamma_u s}\,ds - \beta_u u^2(t) - A\,u(t)$$

and hence $u(t)$ must be bounded above. Hence $\hat{u} = \lim_{t\to\infty} u(t)$ exists and is an equilibrium of (2.9), and obviously $\hat{u} > 0$ in this case. The conclusion follows immediately.

The remaining case to consider is that in which $u(t)$ is neither eventually monotonically decreasing nor increasing. Of the various possibilities which then arise, we shall treat in detail the case in which $u(t)$ has an infinite sequence of local maxima

$\{t_j\}$, $j = 1, 2, 3, \ldots$, all greater than $\hat{u}$, with $t_j \to \infty$. Other cases can be dealt with similarly. At the times t_j, $u(t_j) > \hat{u}$, $\dot{u}(t_j) = 0$, and $\ddot{u}(t_j) < 0$. We claim that $\sup_{t \geq t_1} u(t) = u(t_k)$ for some integer k. If this is false, it means that after every local maximum $u(t_j)$ there is another that is higher, and therefore that a subsequence of $\{t_j\}$ (still denoted $\{t_j\}$) can be found with the property that $u(t) < u(t_j)$ for all $t_1 \leq t < t_j$ and each j. (The subsequence is generated by including each local maximum that is higher than every one before it.) Then, for each j,

$$\begin{aligned} 0 = \dot{u}(t_j) &= \alpha_u \int_0^\tau f_u(s) e^{-\gamma_u s} u(t_j - s)\, ds - \beta_u u^2(t_j) - A\, u(t_j) \\ &\leq u(t_j) \left(\alpha_u \int_0^\tau f_u(s) e^{-\gamma_u s}\, ds - \beta_u u(t_j) - A \right) \\ &< u(t_j) \left(\alpha_u \int_0^\tau f_u(s) e^{-\gamma_u s}\, ds - \beta_u \hat{u} - A \right) \\ &= 0, \end{aligned}$$

which is a contradiction. Thus, $\sup_{t \geq t_1} u(t) = u(t_k)$ for some integer k, and we let $s_1 = t_k$. Now, by applying this same argument to the interval $t \geq t_{k+1}$ we can infer the existence of a t_l ($l > k$) with $\sup_{t \geq t_{k+1}} u(t) = u(t_l)$, and we let $s_2 = t_l$. This process can be continued to generate an infinite sequence $\{s_j\}$ of times such that $s_{j+1} > s_j$, $s_j \to \infty$, $u(t) \leq u(s_j)$ for all $t > s_j$, and $\dot{u}(s_j) = 0$.

Let $y(t) = u(t) - \hat{u}$; then we wish to prove that $y(t) \to 0$ as $t \to \infty$. We have $y(s_j) \geq y(s_{j+1}) > 0$ (since $u(s_j) \geq u(s_{j+1})$ and $u(s_j) > \hat{u}$), and it is now enough to show that $y(s_j) \to 0$ as $j \to \infty$. In terms of y, equation (2.9) becomes, at $t = s_j$,

$$0 = \dot{y}(s_j) = \alpha_u \int_0^\tau f_u(s) e^{-\gamma_u s} y(s_j - s)\, ds - 2\beta_u y(s_j)\hat{u} - \beta_u y^2(s_j) - A y(s_j)$$

so that

$$\begin{aligned} &\alpha_u \int_0^\tau f_u(s) e^{-\gamma_u s} y(s_j - s)\, ds \\ &= 2\beta_u y(s_j)\hat{u} + \beta_u y^2(s_j) + A y(s_j) \\ &= 2\beta_u y(s_j)\hat{u} + \beta_u y^2(s_j) + \left[\alpha_u \int_0^\tau f_u(s) e^{-\gamma_u s}\, ds - \beta_u \hat{u} \right] y(s_j) \\ &= \beta_u y(s_j)\hat{u} + \beta_u y^2(s_j) + \alpha_u y(s_j) \int_0^\tau f_u(s) e^{-\gamma_u s}\, ds \\ &\geq \left(\alpha_u \int_0^\tau f_u(s) e^{-\gamma_u s}\, ds + \beta_u \hat{u} \right) y(s_j), \end{aligned}$$

where we have used (2.10). From the sequence $\{s_j\}$ we shall now extract a further subsequence, still denoted $\{s_j\}$, such that $s_j - \tau \geq s_{j-1}$. Then $y(s_j - s) \leq y(s_{j-1})$ for all $s \in [0, \tau]$ and therefore

$$y(s_j) \leq \frac{\alpha_u \int_0^\tau f_u(s) e^{-\gamma_u s} y(s_j - s)\, ds}{\alpha_u \int_0^\tau f_u(s) e^{-\gamma_u s}\, ds + \beta_u \hat{u}} \leq S\, y(s_{j-1}),$$

where

$$S = \frac{\alpha_u \int_0^\tau f_u(s) e^{-\gamma_u s}\, ds}{\alpha_u \int_0^\tau f_u(s) e^{-\gamma_u s}\, ds + \beta_u \hat{u}}.$$

Now, $S < 1$ and S is independent of j. Therefore, $y(s_j) \to 0$ as $j \to \infty$. We conclude that $\lim_{t\to\infty} u(t) = \hat{u}$, and the proof of Theorem 1 is complete.

2.1. Global stability of $\hat{E}_u$. We shall prove a theorem on the global stability of the equilibrium point

$$\hat{E}_u = \left(\frac{\alpha_u}{\beta_u}\int_0^\tau f_u(s)e^{-\gamma_u s}ds, 0\right)$$

of system (2.7), in the situation when the other boundary equilibrium

$$\hat{E}_v = \left(0, \frac{\alpha_v}{\beta_v}\int_0^\tau f_v(s)e^{-\gamma_v s}ds\right)$$

of (2.7) is linearly unstable. This means that the competition between the two species U and V is strong and the species cannot coexist. One of them, in this case the V population, dies out.

In the proof of Theorem 2 below, and in subsequent theorems, we shall use a comparison principle. Comparison principles do not always hold for delay equations; it depends very much on how the delay appears in the equations. For scalar equations, the essential requirement for a comparison principle to hold is that the reaction term be a nondecreasing function of the delayed variable (see, for example, Martin and Smith [5]). The following proposition will be useful and follows easily from the results in [5].

PROPOSITION 1. *Let $v(t)$ be a solution of*

$$\frac{dv(t)}{dt} = \alpha\int_0^\tau f(s)e^{-\gamma s}v(t-s)\,ds - \beta v^2(t) - \lambda v(t), \qquad t > 0,$$

and $u(t)$ some function satisfying

$$\frac{du(t)}{dt} \geq \alpha\int_0^\tau f(s)e^{-\gamma s}u(t-s)\,ds - \beta u^2(t) - \lambda u(t), \qquad t > 0. \tag{2.11}$$

Assume also that $u(s) \geq v(s)$ for all $s \in [-\tau, 0]$. Then $u(t) \geq v(t)$ for all $t > 0$.

Remarks. An analogous result holds with the inequalities reversed, and we shall need this also. In our applications of these comparison results we shall often find that a differential inequality of the form (2.11) holds only for t above some value, say t_1, and not for all $t > 0$. In that case the initial time is simply thought of as t_1 rather than 0, and $u(t) \geq v(t)$ is arranged to hold for $t \leq t_1$ by appropriate definition of $v(t)$ for values of $t \leq t_1$. In the interests of clarity, we shall not always elaborate on this latter point in detail.

THEOREM 2. *Let the initial data satisfy* (2.8)*, and assume that*

$$c_2\alpha_u\int_0^\tau f_u(s)e^{-\gamma_u s}\,ds > \beta_u\alpha_v\int_0^\tau f_v(s)e^{-\gamma_v s}\,ds \tag{2.12}$$

and

$$c_1\alpha_v\int_0^\tau f_v(s)e^{-\gamma_v s}\,ds < \beta_v\alpha_u\int_0^\tau f_u(s)e^{-\gamma_u s}\,ds. \tag{2.13}$$

Then $U(t) \to \frac{\alpha_u}{\beta_u}\int_0^\tau f_u(s)e^{-\gamma_u s}\,ds$ and $V(t) \to 0$ as $t \to \infty$.

Proof. Let $\overline{U} = \limsup_{t\to\infty} U(t)$, $\underline{U} = \liminf_{t\to\infty} U(t)$, $\overline{V} = \limsup_{t\to\infty} V(t)$, and $\underline{V} = \liminf_{t\to\infty} V(t)$. Now, since

$$\begin{aligned}\frac{dU(t)}{dt} &= \alpha_u \int_0^\tau f_u(s)e^{-\gamma_u s}U(t-s)\,ds - \beta_u U^2(t) - c_1 U(t)V(t)\\ &\le \alpha_u \int_0^\tau f_u(s)e^{-\gamma_u s}U(t-s)\,ds - \beta_u U^2(t),\end{aligned}$$

we can conclude from this and Theorem 1 that $\overline{U} \le U_B$, where

$$U_B = \frac{\alpha_u}{\beta_u}\int_0^\tau f_u(s)e^{-\gamma_u s}\,ds$$

is the U component of the equilibrium $\hat{E}_u$. By positivity of $V(t)$ we also know that $\underline{V} \ge 0$. To complete the proof it suffices to find two sequences $\{M_m^u\}$, $\{N_m^v\}$ with the properties that $\underline{U} \ge M_m^u$ for each m with $M_m^u \to U_B$ as $m \to \infty$ (so that $\underline{U} \ge U_B$), and $\overline{V} \le N_m^v$ for each m with $N_m^v \to 0$ as $m \to \infty$. As a first step in this process, let $v_1(t)$ satisfy

$$\frac{dv_1(t)}{dt} = \alpha_v \int_0^\tau f_v(s)e^{-\gamma_v s}v_1(t-s)\,ds - \beta_v v_1^2(t), \qquad t > 0,$$

with, for $s \le 0$, $v_1(s) \equiv \max\{V(s),\ s \in [-\tau, 0]\} > 0$. Then

$$\lim_{t\to\infty} v_1(t) = \frac{\alpha_v}{\beta_v}\int_0^\tau f_v(s)e^{-\gamma_v s}\,ds.$$

Since $U(t)$ and $V(t)$ are nonnegative,

$$\begin{aligned}\frac{dV(t)}{dt} &= \alpha_v \int_0^\tau f_v(s)e^{-\gamma_v s}V(t-s)\,ds - \beta_v V^2(t) - c_2 U(t)V(t)\\ &\le \alpha_v \int_0^\tau f_v(s)e^{-\gamma_v s}V(t-s)\,ds - \beta_v V^2(t).\end{aligned}$$

By comparison, $V(t) \le v_1(t)$ and therefore

$$\overline{V} = \limsup_{t\to\infty} V(t) \le \lim_{t\to\infty} v_1(t) = \frac{\alpha_v}{\beta_v}\int_0^\tau f_v(s)e^{-\gamma_v s}\,ds := N_1^v.$$

Let $\varepsilon > 0$ be sufficiently small such that

$$0 < \varepsilon < \frac{\beta_v\alpha_u\int_0^\tau f_u(s)e^{-\gamma_u s}\,ds - c_1\alpha_v\int_0^\tau f_v(s)e^{-\gamma_v s}\,ds}{\beta_v c_1}. \tag{2.14}$$

There exists $t_1 > \tau$ such that $V(t) \le N_1^v + \varepsilon$ for all $t \ge t_1$. For $t > t_1$ let $u_1(t)$ evolve according to

$$\frac{du_1(t)}{dt} = \alpha_u \int_0^\tau f_u(s)e^{-\gamma_u s}u_1(t-s)\,ds - \beta_u u_1^2(t) - c_1 u_1(t)(N_1^v + \varepsilon),$$

and for $t \in [t_1 - \tau, t_1]$ let

$$u_1(t) \equiv \min\{U(t),\ t \in [t_1 - \tau, t_1]\},$$

which is strictly positive, since $U(t) > 0$ on $(0, \infty)$. It is not necessary to define $u_1(t)$ for $t < t_1 - \tau$ since Proposition 1 is now being applied with initial time t_1 rather than 0.

Since ε satisfies (2.14), Theorem 1 yields that

$$\lim_{t\to\infty} u_1(t) = \frac{1}{\beta_u}\left[\alpha_u \int_0^\tau f_u(s)e^{-\gamma_u s}\,ds - c_1(N_1^v + \varepsilon)\right].$$

Now, since $N_1^v + \varepsilon \geq V(t)$ for $t \geq t_1$, we have, for such t,

$$\begin{aligned}\frac{dU(t)}{dt} &= \alpha_u \int_0^\tau f_u(s)e^{-\gamma_u s}U(t-s)\,ds - \beta_u U^2(t) - c_1U(t)V(t)\\ &\geq \alpha_u \int_0^\tau f_u(s)e^{-\gamma_u s}U(t-s)\,ds - \beta_u U^2(t) - c_1U(t)(N_1^v + \varepsilon).\end{aligned}$$

By comparison, $U(t) \geq u_1(t)$ and therefore

$$\underline{U} = \liminf_{t\to\infty} U(t) \geq \lim_{t\to\infty} u_1(t) = \frac{1}{\beta_u}\left[\alpha_u \int_0^\tau f_u(s)e^{-\gamma_u s}\,ds - c_1(N_1^v + \varepsilon)\right].$$

Since this is true for any $\varepsilon > 0$ satisfying (2.14), it follows that $\underline{U} \geq M_1^u$, where

$$M_1^u = \frac{1}{\beta_u}\left[\alpha_u \int_0^\tau f_u(s)e^{-\gamma_u s}ds - c_1N_1^v\right].$$

Let $\varepsilon > 0$. There exists $t_2 > 0$ such that $U(t) \geq M_1^u - \varepsilon$ for all $t \geq t_2$. For $t > t_2$ let $v_2(t)$ be the solution of

$$\frac{dv_2(t)}{dt} = \alpha_v \int_0^\tau f_v(s)e^{-\gamma_v s}v_2(t-s)\,ds - \beta_v v_2^2(t) - \frac{\beta_u\alpha_v \int_0^\tau f_v(s)e^{-\gamma_v s}\,ds}{\alpha_u \int_0^\tau f_u(s)e^{-\gamma_u s}\,ds}(M_1^u - \varepsilon)v_2$$

with appropriate "initial data" on the interval $[t_2 - \tau, t_2]$. Now

$$\begin{aligned}\frac{dV(t)}{dt} &= \alpha_v \int_0^\tau f_v(s)e^{-\gamma_v s}V(t-s)\,ds - \beta_v V^2(t) - c_2U(t)V(t)\\ &\leq \alpha_v \int_0^\tau f_v(s)e^{-\gamma_v s}V(t-s)\,ds - \beta_v V^2(t) - \frac{\beta_u\alpha_v \int_0^\tau f_v(s)e^{-\gamma_v s}\,ds}{\alpha_u \int_0^\tau f_u(s)e^{-\gamma_u s}\,ds}(M_1^u - \varepsilon)V(t),\end{aligned}$$

where we have used (2.12). By comparison, $V(t) \leq v_2(t)$. But, by Theorem 1, and using the fact that $M_1^u < (\alpha_u/\beta_u)\int_0^\tau f_u(s)e^{-\gamma_u s}\,ds$,

$$\lim_{t\to\infty} v_2(t) = \frac{1}{\beta_v}\left[\alpha_v \int_0^\tau f_v(s)e^{-\gamma_v s}ds - \frac{\beta_u\alpha_v \int_0^\tau f_v(s)e^{-\gamma_v s}ds}{\alpha_u \int_0^\tau f_u(s)e^{-\gamma_u s}ds}(M_1^u - \varepsilon)\right].$$

Hence

$$\begin{aligned}\overline{V} &= \limsup_{t\to\infty} V(t) \leq \lim_{t\to\infty} v_2(t)\\ &= \frac{1}{\beta_v}\left[\alpha_v \int_0^\tau f_v(s)e^{-\gamma_v s}\,ds - \frac{\beta_u\alpha_v \int_0^\tau f_v(s)e^{-\gamma_v s}\,ds}{\alpha_u \int_0^\tau f_u(s)e^{-\gamma_u s}\,ds}(M_1^u - \varepsilon)\right].\end{aligned}$$

Since ε is arbitrary, we conclude that $\overline{V} \le N_2^v$, where

$$N_2^v = \frac{1}{\beta_v}\left[\alpha_v \int_0^\tau f_v(s)e^{-\gamma_v s}\,ds - \frac{\beta_u\alpha_v \int_0^\tau f_v(s)e^{-\gamma_v s}\,ds}{\alpha_u \int_0^\tau f_u(s)e^{-\gamma_u s}\,ds} M_1^u\right]. \tag{2.15}$$

Now, let $\varepsilon > 0$ be sufficiently small such that the expression given below for $\lim_{t\to\infty} u_2(t)$ is positive. That this is possible follows from the second inequality (2.13) in the hypotheses of Theorem 2, together with the fact that N_2^v satisfies $N_2^v < (\alpha_v/\beta_v)\int_0^\tau f_v(s)e^{-\gamma_v s}\,ds$.

There exists $t_3 > 0$ such that $V(t) \le N_2^v + \varepsilon$ for all $t \ge t_3$. For $t > t_3$ let $u_2(t)$ be a suitable solution of

$$\frac{du_2(t)}{dt} = \alpha_u \int_0^\tau f_u(s)e^{-\gamma_u s}u_2(t-s)\,ds - \beta_u u_2^2(t) - c_1(N_2^v + \varepsilon)u_2(t).$$

Then, since

$$\begin{aligned}\frac{dU(t)}{dt} &= \alpha_u \int_0^\tau f_u(s)e^{-\gamma_u s}U(t-s)\,ds - \beta_u U^2(t) - c_1U(t)V(t)\\ &\ge \alpha_u \int_0^\tau f_u(s)e^{-\gamma_u s}U(t-s)\,ds - \beta_u U^2(t) - c_1(N_2^v + \varepsilon)U(t),\end{aligned}$$

we have $U(t) \ge u_2(t)$. Also

$$\lim_{t\to\infty} u_2(t) = \frac{1}{\beta_u}\left[\alpha_u \int_0^\tau f_u(s)e^{-\gamma_u s}\,ds - c_1(N_2^v + \varepsilon)\right].$$

Hence

$$\underline{U} \ge \frac{1}{\beta_u}\left[\alpha_u \int_0^\tau f_u(s)e^{-\gamma_u s}\,ds - c_1(N_2^v + \varepsilon)\right].$$

By the arbitrariness of $\varepsilon > 0$, $\underline{U} \ge M_2^u$, where

$$M_2^u = \frac{1}{\beta_u}\left[\alpha_u \int_0^\tau f_u(s)e^{-\gamma_u s}\,ds - c_1N_2^v\right]. \tag{2.16}$$

Continuing this process, we obtain two sequences N_m^v, M_m^u, $m = 1, 2, 3, \ldots$, such that, for $m \ge 2$,

$$N_m^v = \frac{1}{\beta_v}\left[\alpha_v \int_0^\tau f_v(s)e^{-\gamma_v s}ds - \frac{\beta_u\alpha_v \int_0^\tau f_v(s)e^{-\gamma_v s}ds}{\alpha_u \int_0^\tau f_u(s)e^{-\gamma_u s}ds} M_{m-1}^u\right] \tag{2.17}$$

and

$$M_m^u = \frac{1}{\beta_u}\left[\alpha_u \int_0^\tau f_u(s)e^{-\gamma_u s}ds - c_1N_m^v\right]. \tag{2.18}$$

Combining these,

$$N_m^v = \frac{c_1\alpha_v \int_0^\tau f_v(s)e^{-\gamma_v s}\,ds}{\beta_v\alpha_u \int_0^\tau f_u(s)e^{-\gamma_u s}\,ds} N_{m-1}^v,$$

which confirms that all the N_m^v are positive. Furthermore, by assumption (2.13), $N_m^v \to 0$ as $m \to \infty$. Hence, by (2.18),

$$\lim_{m\to\infty} M_m^u = \frac{\alpha_u}{\beta_u}\int_0^\tau f_u(s)e^{-\gamma_u s}ds = U_B.$$

Therefore

$$\lim_{t\to\infty} U(t) = U_B$$

and

$$\lim_{t\to\infty} V(t) = 0,$$

which completes the proof of Theorem 2.

The following theorem is an analogue of Theorem 2 for the situation when the equilibrium $\hat{E}_u$ is unstable and $\hat{E}_v$ is asymptotically stable. The proof is similar to that of Theorem 2.

THEOREM 3. *Let the initial data satisfy* (2.8)*, and assume that*

$$c_2\alpha_u\int_0^\tau f_u(s)e^{-\gamma_u s}\,ds < \beta_u\alpha_v\int_0^\tau f_v(s)e^{-\gamma_v s}\,ds$$

and

$$c_1\alpha_v\int_0^\tau f_v(s)e^{-\gamma_v s}\,ds > \beta_v\alpha_u\int_0^\tau f_u(s)e^{-\gamma_u s}\,ds.$$

Then $V(t) \to \frac{\alpha_v}{\beta_v}\int_0^\tau f_v(s)e^{-\gamma_v s}\,ds$ *and* $U(t) \to 0$*, as* $t \to \infty$.

2.2. Global stability of the coexistence state $\hat{E}$. We will prove a theorem on the global stability of the coexistence equilibrium $\hat{E} = (\hat{U}, \hat{V})$ of system (2.7), where

$$\hat{U} = \frac{\beta_v\alpha_u\int_0^\tau f_u(s)e^{-\gamma_u s}ds - c_1\alpha_v\int_0^\tau f_v(s)e^{-\gamma_v s}ds}{\beta_u\beta_v - c_1c_2}$$

and

$$\hat{V} = \frac{\beta_u\alpha_v\int_0^\tau f_v(s)e^{-\gamma_v s}ds - c_2\alpha_u\int_0^\tau f_u(s)e^{-\gamma_u s}ds}{\beta_u\beta_v - c_1c_2}.$$

The hypotheses in Theorem 4 below are those which imply linear instability of both of the boundary equilibria $\hat{E}_u$ and $\hat{E}_v$. In this case the coexistence equilibrium $\hat{E}$ is globally asymptotically stable. The conditions (2.19) and (2.20) below have various ecological interpretations including weak interspecific competition and significant adult mortality. (Recall that deaths rates of *immatures* are measured by γ_u and γ_v, which arise in the conditions in a somewhat different way.)

THEOREM 4. *If the initial data satisfies* (2.8)*, and if the following two conditions hold,*

$$c_2\alpha_u\int_0^\tau f_u(s)e^{-\gamma_u s}\,ds < \beta_u\alpha_v\int_0^\tau f_v(s)e^{-\gamma_v s}ds \tag{2.19}$$

and

$$c_1\alpha_v\int_0^\tau f_v(s)e^{-\gamma_v s}\,ds < \beta_v\alpha_u\int_0^\tau f_u(s)e^{-\gamma_u s}ds, \tag{2.20}$$

then $U(t) \to \hat{U}$ *and* $V(t) \to \hat{V}$ *as* $t \to \infty$.

Proof. Recall that in the proof of Theorem 2 we were able to establish $\overline{U} \le U_B$ and $\underline{V} \ge 0$ in one step, and thereafter we established a sequence of lower bounds M_m^u for $\underline{U}$ and a sequence of upper bounds N_m^v for $\overline{V}$. These sequences approached limits U_B and 0, respectively, establishing our result.

Our approach to proving Theorem 4 is similar, but the situation is more complicated since we are concerned with the coexistence equilibrium. We shall need four sequences, N_m^u, N_m^v, M_m^u, and M_m^v, $m = 1, 2, 3, \ldots$. It is helpful to remember that N_m denotes an upper bound and M_m a lower bound on the limsup and liminf, respectively, as $t \to \infty$, of the variable in the superscript. We shall derive recursion formulae for these bounds and use them to deduce the result.

From positivity of solutions we immediately obtain N_1^u as follows:

$$\frac{dU(t)}{dt} \le \alpha_u \int_0^\tau f_u(s) e^{-\gamma_u s} U(t-s)\, ds - \beta_u U^2(t).$$

Hence

$$\overline{U} = \limsup_{t\to\infty} U(t) \le \frac{\alpha_u}{\beta_u} \int_0^\tau f_u(s) e^{-\gamma_u s}\, ds := N_1^u.$$

In a similar way, we have

$$\overline{V} \le \frac{\alpha_v}{\beta_v} \int_0^\tau f_v(s) e^{-\gamma_v s}\, ds := N_1^v.$$

Let $\varepsilon > 0$ be sufficiently small such that

$$\varepsilon < \frac{\beta_u \alpha_v \int_0^\tau f_v(s) e^{-\gamma_v s}\, ds - c_2 \alpha_u \int_0^\tau f_u(s) e^{-\gamma_u s}\, ds}{\beta_u c_2}, \tag{2.21}$$

which is possible by (2.19). Let $t_1 > 0$ be such that $U(t) \le N_1^u + \varepsilon$ for all $t \ge t_1$, and for $t > t_1$ let $m_1^v(t)$ be a solution of

$$\frac{dm_1^v(t)}{dt} = \alpha_v \int_0^\tau f_v(s) e^{-\gamma_v s} m_1^v(t-s)\, ds - \beta_v (m_1^v(t))^2 - c_2 (N_1^u + \varepsilon) m_1^v(t)$$

with appropriate initial data on $[t_1 - \tau, t_1]$. Since ε satisfies (2.21), Theorem 1 applies and yields

$$\lim_{t\to\infty} m_1^v(t) = \frac{1}{\beta_v} \left[\alpha_v \int_0^\tau f_v(s) e^{-\gamma_v s} ds - c_2 (N_1^u + \varepsilon) \right].$$

Since $N_1^u + \varepsilon \ge U(t)$ for $t \ge t_1$,

$$\begin{aligned} \frac{dV(t)}{dt} &= \alpha_v \int_0^\tau f_v(s) e^{-\gamma_v s} V(t-s)\, ds - \beta_v V^2(t) - c_2 U(t) V(t) \\ &\ge \alpha_v \int_0^\tau f_v(s) e^{-\gamma_v s} V(t-s)\, ds - \beta_v V^2(t) - c_2 (N_1^u + \varepsilon) V(t) \end{aligned}$$

so that $V(t) \ge m_1^v(t)$, and hence

$$\underline{V} = \liminf_{t\to\infty} V(t) \ge \lim_{t\to\infty} m_1^v(t) = \frac{1}{\beta_v} \left[\alpha_v \int_0^\tau f_v(s) e^{-\gamma_v s}\, ds - c_2 (N_1^u + \varepsilon) \right].$$

This is true for any $\varepsilon > 0$ satisfying (2.21), and hence

$$\underline{V} \geq \frac{1}{\beta_v}\left[\alpha_v \int_0^\tau f_v(s)e^{-\gamma_v s}\,ds - c_2 N_1^u\right] := M_1^v.$$

In exactly the same way, we can show that

$$\underline{U} \geq \frac{1}{\beta_u}\left[\alpha_u \int_0^\tau f_u(s)e^{-\gamma_u s}\,ds - c_1 N_1^v\right] := M_1^u,$$

and, in doing so, the assumption (2.20) is used.

We shall now show how to find new upper bounds N_2^u, N_2^v in terms of the old lower bounds M_1^v, M_1^u, respectively. New lower bounds are then found from the *new* upper bounds by following the procedure already described. It will then be clear how to proceed from the $(m-1)$th to the mth step in this process.

Let $\varepsilon > 0$. There exists $t_2 > 0$ such that $V(t) \geq M_1^v - \varepsilon$ for all $t \geq t_2$. Then, for $t \geq t_2$,

$$\begin{aligned}\frac{dU(t)}{dt} &= \alpha_u \int_0^\tau f_u(s)e^{-\gamma_u s}U(t-s)\,ds - \beta_u U^2(t) - c_1 U(t)V(t)\\ &\leq \alpha_u \int_0^\tau f_u(s)e^{-\gamma_u s}U(t-s)\,ds - \beta_u U^2(t) - c_1(M_1^v - \varepsilon)U(t).\end{aligned}$$

Thus, if for $t > t_2$ we denote by $n_2^u(t)$ the solution of

$$\frac{dn_2^u(t)}{dt} = \alpha_u \int_0^\tau f_u(s)e^{-\gamma_u s}n_2^u(t-s)\,ds - \beta_u(n_2^u(t))^2 - c_1(M_1^v - \varepsilon)n_2^u(t)$$

with appropriate initial data, then $U(t) \leq n_2^u(t)$ and thus

$$\overline{U} \leq \lim_{t\to\infty} n_2^u(t) = \frac{1}{\beta_u}\left[\alpha_u \int_0^\tau f_u(s)e^{-\gamma_u s}\,ds - c_1(M_1^v - \varepsilon)\right].$$

(We have used assumption (2.20) to deduce that $n_2^u(t)$ has this limiting behavior.) Since $\varepsilon > 0$ is arbitrary,

$$\overline{U} \leq \frac{1}{\beta_u}\left[\alpha_u \int_0^\tau f_u(s)e^{-\gamma_u s}\,ds - c_1 M_1^v\right] := N_2^u.$$

In the same way, and using (2.19), we deduce the following estimate for $\overline{V}$:

$$\overline{V} \leq \frac{1}{\beta_v}\left[\alpha_v \int_0^\tau f_v(s)e^{-\gamma_v s}\,ds - c_2 M_1^u\right] := N_2^v.$$

One now sees that the transition from the $(m-1)$th to the mth step in this iterative process is given by

$$N_m^u = \frac{1}{\beta_u}\left[\alpha_u \int_0^\tau f_u(s)e^{-\gamma_u s}\,ds - c_1 M_{m-1}^v\right],$$

$$N_m^v = \frac{1}{\beta_v}\left[\alpha_v \int_0^\tau f_v(s)e^{-\gamma_v s}\,ds - c_2 M_{m-1}^u\right],$$

$$M_m^u = \frac{1}{\beta_u}\left[\alpha_u \int_0^\tau f_u(s)e^{-\gamma_u s}\,ds - c_1 N_m^v\right],$$
$$M_m^v = \frac{1}{\beta_v}\left[\alpha_v \int_0^\tau f_v(s)e^{-\gamma_v s}\,ds - c_2 N_m^u\right],$$

and, of course,

$$M_m^u \le \underline{U} \le \overline{U} \le N_m^u \quad \text{and} \quad M_m^v \le \underline{V} \le \overline{V} \le N_m^v$$

for each $m = 1, 2, 3, \ldots$. We need to show that M_m^u and N_m^u both approach $\hat{U}$ as $m \to \infty$ and that M_m^v and N_m^v both approach $\hat{V}$.

We see at once that

$$N_m^u = \frac{\alpha_u\beta_v \int_0^\tau f_u(s)e^{-\gamma_u s}\,ds - c_1\alpha_v \int_0^\tau f_v(s)e^{-\gamma_v s}\,ds}{\beta_u\beta_v} + \frac{c_1c_2}{\beta_u\beta_v}N_{m-1}^u. \tag{2.22}$$

Note that (2.19) and (2.20) imply that

$$\frac{c_1c_2}{\beta_u\beta_v} < 1.$$

We claim that N_m^u is a monotonically decreasing sequence that is bounded below by $\hat{U}$. The boundedness below by $\hat{U}$ follows immediately from (2.22) by induction. Then, by (2.22), and using (2.20),

$$\begin{aligned}\frac{N_m^u}{N_{m-1}^u} &= \frac{\alpha_u\beta_v \int_0^\tau f_u(s)e^{-\gamma_u s}\,ds - c_1\alpha_v \int_0^\tau f_v(s)e^{-\gamma_v s}\,ds}{\beta_u\beta_v N_{m-1}^u} + \frac{c_1c_2}{\beta_u\beta_v}\\ &\le \frac{\alpha_u\beta_v \int_0^\tau f_u(s)e^{-\gamma_u s}\,ds - c_1\alpha_v \int_0^\tau f_v(s)e^{-\gamma_v s}\,ds}{\beta_u\beta_v \hat{U}} + \frac{c_1c_2}{\beta_u\beta_v}\\ &= 1\end{aligned}$$

so that N_m^u is monotonically decreasing. Hence N_m^u converges to a limit which, by (2.22), equals $\hat{U}$.

Of course, convergence of N_m^u implies convergence of M_m^v, and it is easily checked that M_m^v has the limit $\hat{V}$. The analysis for the remaining two sequences is similar. The proof of the theorem is complete.

3. Traveling wave front solutions. In this section we shall explore the existence of a traveling wave front solution of a reaction-diffusion extension of system (1.1) between the two boundary equilibria $\hat{E}_v$ and $\hat{E}_u$, in the situation when V is the weaker competitor and there is no coexistence equilibrium. Ecologically, this situation corresponds to a one-dimensional habitat initially inhabited only by the weaker V species, and then some of the U species are introduced at one end. The U species then invades the domain, driving the V species to extinction. The end result is that the domain is inhabited only by U.

The approach we shall use to prove the existence of such a traveling front is the upper-lower solution technique and the monotone iteration method recently developed by Wu and Zou [9] (in particular, Theorem 3.6 of that paper) for delayed reaction-diffusion systems. To be precise, we shall consider the competitive system

$$\begin{aligned}\frac{\partial U_1(t)}{\partial t} &= d_1\frac{\partial^2 U_1}{\partial x^2} + \alpha_1 e^{-\gamma_1\tau_1}U_1(x,t-\tau_1) - \beta_1 U_1^2(x,t) - c_1U_1(x,t)U_2(x,t),\\ \frac{\partial U_2(t)}{\partial t} &= d_2\frac{\partial^2 U_2}{\partial x^2} + \alpha_2 e^{-\gamma_2\tau_2}U_2(x,t-\tau_2) - \beta_2 U_2^2(x,t) - c_2U_1(x,t)U_2(x,t),\end{aligned} \tag{3.1}$$

where all the parameters are nonnegative constants.

It is well known that the addition of diffusion to a system with time delays can be problematic from the point of view of ecological realism. One expects that in general a time-delayed term would need to become a weighted spatial average involving the diffusivity. However, system (3.1) is realistic in the case when the *immature* members of the species are not moving (the immatures do not explicitly feature in (3.1)). For many species such an assumption is entirely realistic. For example, insects which go through a larval stage do not move, or move hardly at all, during the larval phase, but on becoming adults they can of course travel tremendous distances. In such cases, each individual on reaching adulthood is still at the same location as when it was born, and only on becoming an adult does it start moving.

The situation when the immatures *do* move can be studied by replacing the time delay terms in (3.1) with more complicated delay terms involving integral convolutions in space (see, for example, Al-Omari and Gourley [2]). Of course, the immatures and matures might diffuse at different rates. The methods of Wu and Zou [9] should be applicable to this case also (see, for example, So, Wu, and Zou [7] who treated a scalar delay equation containing such an integral convolution in space). In this paper we shall concentrate on the case when the immatures are immobile.

System (3.1) has four spatially uniform equilibria: $(\frac{\alpha_1}{\beta_1}e^{-\gamma_1\tau_1}, 0)$, $(0, \frac{\alpha_2}{\beta_2}e^{-\gamma_2\tau_2})$, $(0,0)$, and the coexistence equilibrium (if feasible)

$$\left(\frac{\beta_2\alpha_1e^{-\gamma_1\tau_1} - c_1\alpha_2e^{-\gamma_2\tau_2}}{\beta_1\beta_2 - c_1c_2}, \quad \frac{\beta_1\alpha_2e^{-\gamma_2\tau_2} - c_2\alpha_1e^{-\gamma_1\tau_1}}{\beta_1\beta_2 - c_1c_2}\right).$$

The situation of interest here is that in which the equilibrium $(0, \frac{\alpha_2}{\beta_2}e^{-\gamma_2\tau_2})$ is unstable and when $(\frac{\alpha_1}{\beta_1}e^{-\gamma_1\tau_1}, 0)$ is stable, both as solutions of the spatially uniform model. The conditions on the parameters can be extracted as a particular case of Theorem 2 and are

$$c_2\alpha_1e^{-\gamma_1\tau_1} > \beta_1\alpha_2e^{-\gamma_2\tau_2} \quad \text{and} \quad c_1\alpha_2e^{-\gamma_2\tau_2} < \beta_2\alpha_1e^{-\gamma_1\tau_1}. \tag{3.2}$$

Therefore, we assume that (3.2) holds throughout this section. Note that in this case the coexistence equilibrium is not feasible since its components are of opposite sign.

3.1. Wu and Zou's theory. We shall summarize those aspects of the work of Wu and Zou [9] which are relevant here. Applications of their techniques include work by Al-Omari and Gourley [2] on a structured model of a single species, by So and Zou [8] on the diffusive Nicholson's blowflies equation, by Huang and Zou [4] on a cooperative Lotka–Volterra system with delay, and by Wu and Zou themselves [9] on a delayed Fisher equation and a delayed Belousov–Zhabotinskii reaction model.

The theory is for reaction diffusion systems of the form

$$\frac{\partial u(x,t)}{\partial t} = D\frac{\partial^2 u(x,t)}{\partial x^2} + f(u_t(x)), \qquad x \in \mathbf{R},\ t \geq 0, \tag{3.3}$$

with $u \in \mathbf{R}^n$, D a diagonal diffusion matrix with positive entries, $f(\cdot)$ a functional, and its argument $u_t(x)$ the function

$$u_t(x)(s) = u(x, t+s), \qquad s \in [-\tau, 0].$$

This set-up allows more than one time delay. Conversion to traveling wave form, by setting $u(x,t) = \phi(z) \in \mathbf{R}^n$, $z = x + ct$ with $c \geq 0$, yields a system of equations of the form

$$D\phi''(z) - c\phi'(z) + f_c(\phi_z) = \mathbf{0}, \qquad z \in \mathbf{R}, \tag{3.4}$$

where $\phi_z(\zeta) = \phi(\zeta + z)$ and the new functional f_c is defined by $f_c(\psi) = f(\psi^c)$, where $\psi^c(s) := \psi(cs)$, $s \in [-\tau, 0]$. The theory establishes, under certain conditions, existence of a solution of (3.4) satisfying

$$\phi(-\infty) = \mathbf{0}, \qquad \phi(\infty) = \mathbf{K}, \tag{3.5}$$

where $\mathbf{0}$ and $\mathbf{K}$ are equilibria of (3.3).

The theory presumes that there are no other equilibria u with $\mathbf{0} < u < \mathbf{K}$ (here, the ordering is the standard partial ordering in $\mathbf{R}^n$, i.e., $u \le v$ if $u_i \le v_i$, $i = 1, \dots, n$, and $u < v$ if $u \le v$ but $u \ne v$). Actually, this condition is not quite satisfied for us, for reasons which will be explained later. We shall argue, by a detailed examination of its proof, that Wu and Zou's Theorem 3.6 applies nevertheless.

Additionally [9, p. 659], there must exist a matrix $\delta = \operatorname{diag}(\delta_1, \dots, \delta_n)$ with $\delta_i \ge 0$ such that

$$f_c(\phi) - f_c(\psi) + \delta(\phi(0) - \psi(0)) > \mathbf{0} \tag{3.6}$$

for all continuous functions ϕ, ψ such that $\mathbf{0} \le \psi(s) \le \phi(s) \le \mathbf{K}$, $s \in [-c\tau, 0]$. This is called a quasi-monotonicity condition.

The following set is called the *profile set*:

$$\Gamma = \{\phi \in C(\mathbf{R}, \mathbf{R}^n) : \ \phi(z) \text{ is nondecreasing, } \phi(-\infty) = \mathbf{0}, \text{ and } \phi(\infty) = \mathbf{K}\}.$$

Additionally, one must find an *upper solution*, i.e., a function $\overline{\phi}$ that is twice differentiable almost everywhere and satisfies

$$D\overline{\phi}''(z) - c\overline{\phi}'(z) + f_c(\overline{\phi}_z) \le \mathbf{0}.$$

A *lower solution* $\underline{\phi}$ is also required, being a function defined as above but with the inequality reversed. These two functions must be constructed such that

$$\mathbf{0} \le \underline{\phi}(z) \le \overline{\phi}(z) \le \mathbf{K}$$

and such that at least one component of the lower solution is not identically zero. Actually only one component of our lower solution is nonzero, but we shall be arguing that refinements of Wu and Zou's approach (see Remark 4.6 in [9]) lead to their Theorem 3.6 remaining valid.

If all of the above conditions are met, and if the upper solution $\overline{\phi} \in \Gamma$ (the lower solution need not lie in Γ), Theorem 3.6 in [9] guarantees the existence of a solution of (3.4) satisfying (3.5).

3.2. Existence of a traveling wave front. In this section, we will prove the existence of a traveling front solution of (3.1) by using Wu and Zou's theory.

To seek a traveling front solution of system (3.1), set $U_1(x,t) = \phi_1(z)$ and $U_2(x,t) = \phi_2(z)$, where $z = x + ct$ and $c > 0$ is the wave speed. The system (3.1) becomes

$$\begin{aligned} &d_1\phi_1''(z) - c\phi_1'(z) + \alpha_1 e^{-\gamma_1\tau_1}\phi_1(z - c\tau_1) - \beta_1\phi_1^2(z) - c_1\phi_1(z)\phi_2(z) = 0, \\ &d_2\phi_2''(z) - c\phi_2'(z) + \alpha_2 e^{-\gamma_2\tau_2}\phi_2(z - c\tau_2) - \beta_2\phi_2^2(z) - c_2\phi_1(z)\phi_2(z) = 0, \end{aligned} \tag{3.7}$$

which is to be solved subject to

$$\phi_1(-\infty) = 0, \quad \phi_2(-\infty) = \frac{\alpha_2}{\beta_2}e^{-\gamma_2\tau_2}, \quad \phi_1(\infty) = \frac{\alpha_1}{\beta_1}e^{-\gamma_1\tau_1}, \quad \phi_2(\infty) = 0,$$

and also, for ecological realism, $\phi_1(z), \phi_2(z) \geq 0$ for all $z \in (-\infty, \infty)$. The latter is only possible for c exceeding a certain minimum value, as can be seen from the following linearized analysis. As $z \to -\infty$, the first equation of (3.7) becomes, approximately,

$$d_1\phi_1''(z) - c\phi_1'(z) + \alpha_1 e^{-\gamma_1\tau_1}\phi_1(z - c\tau_1) - \frac{c_1\alpha_2}{\beta_2}e^{-\gamma_2\tau_2}\phi_1(z) = 0,$$

and, seeking solutions of this proportional to $\exp(\lambda z)$, one finds that $\Delta_1(\lambda) = 0$, where

$$\Delta_1(\lambda) = \alpha_1 e^{-\gamma_1\tau_1}e^{-\lambda c\tau_1} - \frac{c_1\alpha_2}{\beta_2}e^{-\gamma_2\tau_2} - \left(c\lambda - d_1\lambda^2\right). \tag{3.8}$$

Similarly, if we let $z \to \infty$ and approximate the second equation of (3.7) suitably, trial solutions of the form $\exp(\lambda z)$ exist when $\Delta_2(\lambda) = 0$, where

$$\Delta_2(\lambda) = \alpha_2 e^{-\gamma_2\tau_2}e^{-\lambda c\tau_2} - \frac{c_2\alpha_1}{\beta_1}e^{-\gamma_1\tau_1} - \left(c\lambda - d_2\lambda^2\right). \tag{3.9}$$

Since positive and monotone solutions are sought, any decay to zero as $z \to \pm\infty$ must be nonoscillatory. Now, $\Delta_1(\lambda) = 0$ relates to the situation as $z \to -\infty$, so it is necessary that this equation should have at least one real positive root, while $\Delta_2(\lambda) = 0$ should have a real negative root since the latter equation is for $z \to \infty$.

Keeping in mind (3.2), simple graphical arguments show that $\Delta_1(\lambda) = 0$ has no real positive roots if c is very small, but that if c is increased, two real roots appear. We denote by c^* the infimum of the set of values of c for which there are two real positive roots. For any $c > c^*$ the two roots are denoted by $0 < \lambda_1 < \lambda_2$. Furthermore

$$\Delta_1(\lambda) = \begin{cases} > 0 & \text{for} \quad \lambda < \lambda_1, \\ < 0 & \text{for} \quad \lambda \in (\lambda_1, \lambda_2), \\ > 0 & \text{for} \quad \lambda > \lambda_2. \end{cases}$$

The existence of a real negative root (which we shall denote by λ_3) of $\Delta_2(\lambda) = 0$ follows immediately from (3.2) and a graphical argument, without further restriction on c. This equation also has a real positive root λ_4. Furthermore, $\Delta_2(\lambda) > 0$ when $\lambda < \lambda_3$, $\Delta_2(\lambda) < 0$ when $\lambda_3 < \lambda < \lambda_4$, and $\Delta_2(\lambda) > 0$ when $\lambda > \lambda_4$.

We are thus led to conjecture that ecologically relevant traveling fronts exist only for $c > c^*$. Furthermore, note that c^* depends on both of the delays τ_1 and τ_2.

Wu and Zou's theory presumes that the equilibria of the traveling wave equations are $\mathbf{0}$ and $\mathbf{K}$, where $\mathbf{K}$ is a vector with positive components. This is not so in our problem as currently posed but can be made so by the change of variables $\tilde{U}_1 = U_1$ and $\tilde{U}_2 = \frac{\alpha_2}{\beta_2}e^{-\gamma_2\tau_2} - U_2$ under which the system (3.1) is transformed into

$$\begin{aligned} \frac{\partial\tilde{U}_1(t)}{\partial t} &= d_1\frac{\partial^2\tilde{U}_1}{\partial x^2} + \alpha_1 e^{-\gamma_1\tau_1}\tilde{U}_1(x, t-\tau_1) - \beta_1\tilde{U}_1^2(x,t) - \frac{c_1\alpha_2}{\beta_2}e^{-\gamma_2\tau_2}\tilde{U}_1(x,t) \\ &\quad + c_1\tilde{U}_1(x,t)\tilde{U}_2(x,t), \\ \frac{\partial\tilde{U}_2(t)}{\partial t} &= d_2\frac{\partial^2\tilde{U}_2}{\partial x^2} + \alpha_2 e^{-\gamma_2\tau_2}\tilde{U}_2(x, t-\tau_2) + \beta_2\tilde{U}_2^2(x,t) - 2\alpha_2 e^{-\gamma_2\tau_2}\tilde{U}_2(x,t) \\ &\quad + \frac{c_2\alpha_2}{\beta_2}e^{-\gamma_2\tau_2}\tilde{U}_1(x,t) - c_2\tilde{U}_1(x,t)\tilde{U}_2(x,t). \end{aligned} \tag{3.10}$$

The equilibria of interest are now $(0,0)$ and

$$\mathbf{K} := \left(\frac{\alpha_1}{\beta_1}e^{-\gamma_1\tau_1}, \frac{\alpha_2}{\beta_2}e^{-\gamma_2\tau_2}\right), \tag{3.11}$$

which are unstable and stable, respectively, as solutions of (3.10) to spatially uniform perturbation. We have noted that the coexistence equilibrium is unfeasible, being in either the second or the fourth quadrant of the (U_1, U_2) plane, but in the latter case it is mapped into the open first quadrant of the $(\tilde{U}_1, \tilde{U}_2)$ plane. Fortunately, (3.2) yields that its components are larger than those of $\mathbf{K}$, so that Wu and Zou's results remain applicable. The origin is mapped to $\tilde{U}_1 = 0$, $\tilde{U}_2 = (\alpha_2/\beta_2)e^{-\gamma_2\tau_2}$. This equilibrium lies in $(\mathbf{0}, \mathbf{K})$ according to the meaning of $<$, and thus the assumption that there is no equilibrium u with $\mathbf{0} < u < \mathbf{K}$ does not actually hold. We shall return to this point later. It is not a serious problem and can be dealt with as suggested in Remark 4.6 of [9].

In traveling wave form, system (3.10) becomes (with tildes dropped)

$$\begin{aligned}
&d_1\phi_1''(z) - c\phi_1'(z) + \alpha_1 e^{-\gamma_1\tau_1}\phi_1(z - c\tau_1) - \beta_1\phi_1^2(z) - \frac{c_1\alpha_2}{\beta_2}e^{-\gamma_2\tau_2}\phi_1(z) \\
&\qquad + c_1\phi_1(z)\phi_2(z) = 0, \\
&d_2\phi_2''(z) - c\phi_2'(z) + \alpha_2 e^{-\gamma_2\tau_2}\phi_2(z - c\tau_2) + \beta_2\phi_2^2(z) - 2\alpha_2 e^{-\gamma_2\tau_2}\phi_2(z) + \frac{c_2\alpha_2}{\beta_2}e^{-\gamma_2\tau_2}\phi_1 \\
&\qquad - c_2\phi_1(z)\phi_2(z) = 0,
\end{aligned} \tag{3.12}$$

and the boundary conditions are now

$$\begin{aligned}
&\phi_1(-\infty) = 0, \quad \phi_1(\infty) = \frac{\alpha_1}{\beta_1}e^{-\gamma_1\tau_1}, \\
&\phi_2(-\infty) = 0, \quad \phi_2(\infty) = \frac{\alpha_2}{\beta_2}e^{-\gamma_2\tau_2}.
\end{aligned} \tag{3.13}$$

We will prove the following theorem. For ecological relevance the theorem is formulated in terms of the original competition system (3.1). However, for the proof we shall work with the transformed system and its associated traveling wave equations (3.12).

THEOREM 5. *Assume* (3.2) *holds and that* $c > c^*$. *Assume also that*

$$\Delta_2(\lambda_1) \le 2\left(\alpha_2 e^{-\gamma_2\tau_2} - \frac{c_2\alpha_1}{\beta_1}e^{-\gamma_1\tau_1}\right). \tag{3.14}$$

Then there exists a traveling wave front for (3.1) *with speed* c, *connecting the equilibria* $(0, \frac{\alpha_2}{\beta_2}e^{-\gamma_2\tau_2})$ *and* $(\frac{\alpha_1}{\beta_1}e^{-\gamma_1\tau_1}, 0)$.

Remarks. It is inconvenient to have to impose the condition (3.14). We feel that the condition is probably not necessary for Theorem 5 to hold. However, it is needed in our proof. Let us satisfy ourselves that the condition can be satisfied under ecologically realistic circumstances. Note that the right-hand side of (3.14) is negative (by (3.2)), so $\Delta_2(\lambda_1)$ must be negative too. This can certainly be arranged, (for example, by taking d_2 sufficiently small) and then the question is, Under what circumstances will $\Delta_2(\lambda_1)$ be sufficiently negative to satisfy (3.14)? Let us view both sides of (3.14) as functions of c_2 and imagine that c_2 approaches, from above, the critical c_2^* at which (3.2) ceases to hold. Then the right-hand side of (3.14) approaches zero while, noting that λ_1 does not depend on c_2, the left-hand side $\Delta_2(\lambda_1)$ approaches some strictly negative number. Thus, (3.14) is certainly satisfied provided c_2 is not too much greater than the minimum value of c_2 consistent with (3.2).

Proof of Theorem 5. Let $\phi = (\phi_1, \phi_2)$. To prove the theorem we need to show that the quasi-monotonicity condition holds and that upper and lower solutions $\overline{\phi}$ and $\underline{\phi}$ can be found as described in section 3.1, with at least one component of the lower solution not identically zero. (We shall explain later why, in our case, we do not require both components to be nonzero.) For the system (3.12) the functional $f_c(\phi) = (f_{c1}(\phi), f_{c2}(\phi))$ referred to in section 3.1 is given by

$$\begin{aligned} f_{c1}(\phi) &= \alpha_1 e^{-\gamma_1\tau_1}\phi_1(-c\tau_1) - \beta_1\phi_1^2(0) - \frac{c_1\alpha_2}{\beta_2}e^{-\gamma_2\tau_2}\phi_1(0) + c_1\phi_1(0)\phi_2(0), \\ f_{c2}(\phi) &= \alpha_2 e^{-\gamma_2\tau_2}\phi_2(-c\tau_2) + \beta_2\phi_2^2(0) - 2\alpha_2 e^{-\gamma_2\tau_2}\phi_2(0) + \frac{c_2\alpha_2}{\beta_2}e^{-\gamma_2\tau_2}\phi_1(0) \\ &\quad - c_2\phi_1(0)\phi_2(0). \end{aligned}$$

Let us verify that this f_c satisfies the quasi-monotonicity condition. Let $\tau = \max\{\tau_1, \tau_2\}$. For any $\phi = (\phi_1, \phi_2)$ and $\psi = (\psi_1, \psi_2) \in C([-c\tau; 0]; \mathbf{R}^2)$, with $\mathbf{0} \le \psi(s) \le \phi(s) \le \mathbf{K}$ for all $s \in [-c\tau, 0]$, we have

$$\begin{aligned} &f_{c1}(\phi) - f_{c1}(\psi) \\ &= \alpha_1 e^{-\gamma_1\tau_1}\phi_1(-c\tau_1) - \beta_1\phi_1^2(0) - \frac{c_1\alpha_2}{\beta_2}e^{-\gamma_2\tau_2}\phi_1(0) + c_1\phi_1(0)\phi_2(0) \\ &\quad - \alpha_1 e^{-\gamma_1\tau_1}\psi_1(-c\tau_1) + \beta_1\psi_1^2(0) + \frac{c_1\alpha_2}{\beta_2}e^{-\gamma_2\tau_2}\psi_1(0) - c_1\psi_1(0)\psi_2(0) \\ &= \alpha_1 e^{-\gamma_1\tau_1}\phi_1(-c\tau_1) - \alpha_1 e^{-\gamma_1\tau_1}\psi_1(-c\tau_1) - \beta_1\left[\phi_1^2(0) - \psi_1^2(0)\right] \\ &\quad - \frac{c_1\alpha_2}{\beta_2}e^{-\gamma_2\tau_2}\left(\phi_1(0) - \psi_1(0)\right) - c_1\left(\psi_1(0)\psi_2(0) - \phi_1(0)\phi_2(0)\right) \\ &\ge -\beta_1\left(\phi_1(0) - \psi_1(0)\right)\left(\phi_1(0) + \psi_1(0)\right) - \frac{c_1\alpha_2}{\beta_2}e^{-\gamma_2\tau_2}\left(\phi_1(0) - \psi_1(0)\right) \\ &\ge \left[-2\alpha_1 e^{-\gamma_1\tau_1} - \frac{c_1\alpha_2}{\beta_2}e^{-\gamma_2\tau_2}\right]\left(\phi_1(0) - \psi_1(0)\right) \end{aligned}$$

and hence

$$\begin{aligned} &f_{c1}(\phi) - f_{c1}(\psi) + \delta_1\left[\phi_1(0) - \psi_1(0)\right] \\ &\ge \left[-2\alpha_1 e^{-\gamma_1\tau_1} - \frac{c_1\alpha_2}{\beta_2}e^{-\gamma_2\tau_2} + \delta_1\right]\left(\phi_1(0) - \psi_1(0)\right) \ge 0 \end{aligned}$$

provided δ_1 is chosen such that

$$\delta_1 \ge 2\alpha_1 e^{-\gamma_1\tau_1} + \frac{c_1\alpha_2}{\beta_2}e^{-\gamma_2\tau_2}.$$

Similarly, we have

$$\begin{aligned} f_{c2}(\phi) - f_{c2}(\psi) &= \alpha_2 e^{-\gamma_2\tau_2}\phi_2(-c\tau_2) - \alpha_2 e^{-\gamma_2\tau_2}\psi_2(-c\tau_2) - 2\alpha_2 e^{-\gamma_2\tau_2}\left(\phi_2(0) - \psi_2(0)\right) \\ &\quad + \beta_2\left(\phi_2^2(0) - \psi_2^2(0)\right) + \frac{c_2\alpha_2}{\beta_2}e^{-\gamma_2\tau_2}\left(\phi_1(0) - \psi_1(0)\right) - c_2\left(\phi_1(0)\phi_2(0) - \psi_1(0)\psi_2(0)\right) \\ &\ge -2\alpha_2 e^{-\gamma_2\tau_2}\left(\phi_2(0) - \psi_2(0)\right) + \frac{c_2\alpha_2}{\beta_2}e^{-\gamma_2\tau_2}\left(\phi_1(0) - \psi_1(0)\right) \\ &\quad + \beta_2\left(\phi_2(0) - \psi_2(0)\right)\left(\phi_2(0) + \psi_2(0)\right) - c_2\left(\phi_1(0)\phi_2(0) - \psi_1(0)\psi_2(0)\right) \\ &= -2\alpha_2 e^{-\gamma_2\tau_2}\left(\phi_2(0) - \psi_2(0)\right) + \beta_2\left(\phi_2(0) - \psi_2(0)\right)\left(\phi_2(0) + \psi_2(0)\right) \\ &\quad + \frac{c_2\alpha_2}{\beta_2}e^{-\gamma_2\tau_2}\left(\phi_1(0) - \psi_1(0)\right) - c_2\phi_2(0)\left(\phi_1(0) - \psi_1(0)\right) - c_2\psi_1(0)\left(\phi_2(0) - \psi_2(0)\right) \end{aligned}$$

$$\geq -2\alpha_2 e^{-\gamma_2\tau_2}\left(\phi_2(0)-\psi_2(0)\right)+\beta_2\left(\phi_2(0)-\psi_2(0)\right)\left(\phi_2(0)+\psi_2(0)\right)$$
$$+\frac{c_2\alpha_2}{\beta_2}e^{-\gamma_2\tau_2}\left(\phi_1(0)-\psi_1(0)\right)-c_2\frac{\alpha_2}{\beta_2}e^{-\gamma_2\tau_2}\left(\phi_1(0)-\psi_1(0)\right)$$
$$-c_2\frac{\alpha_1}{\beta_1}e^{-\gamma_1\tau_1}\left(\phi_2(0)-\psi_2(0)\right)$$
$$\geq\left[-2\alpha_2 e^{-\gamma_2\tau_2}-\frac{c_2\alpha_1}{\beta_1}e^{-\gamma_1\tau_1}\right]\left(\phi_2(0)-\psi_2(0)\right).$$

Therefore

$$f_{c2}(\phi)-f_{c2}(\psi)+\delta_2\left[\phi_2(0)-\psi_2(0)\right]$$
$$\geq\left[\delta_2-2\alpha_2 e^{-\gamma_2\tau_2}-\frac{c_2\alpha_1}{\beta_1}e^{-\gamma_1\tau_1}\right]\left(\phi_2(0)-\psi_2(0)\right)\geq 0$$

provided

$$\delta_2\geq 2\alpha_2 e^{-\gamma_2\tau_2}+\frac{c_2\alpha_1}{\beta_1}e^{-\gamma_1\tau_1}.$$

Consequently,

$$f_c(\phi)-f_c(\psi)+\delta\left[\phi(0)-\psi(0)\right]\geq 0,$$

where the matrix δ is given by $\delta=\operatorname{diag}(\delta_1,\delta_2)$. Thus, the quasi-monotonicity condition holds.

Our search for solutions of system (3.12) shall be confined to the profile set

$$\Gamma=\left\{\phi\in C(\mathbf{R},\mathbf{R}^2):\ \begin{array}{l}\text{(i) } \phi \text{ is componentwise nondecreasing in } \mathbf{R},\\ \text{(ii) } \lim\limits_{z\to-\infty}\phi(z)=\mathbf{0} \text{ and } \lim\limits_{z\to\infty}\phi(z)=\mathbf{K}\end{array}\right\},$$

where $\mathbf{K}$ is given by (3.11).

Next, we shall find a pair of upper and lower solutions as required by Wu and Zou's theory. Let $\tau_1,\tau_2>0$, and define

$$\overline{\phi}_1(z)=\min\left\{\frac{\alpha_1}{\beta_1}e^{-\gamma_1\tau_1}e^{\lambda_1 z},\frac{\alpha_1}{\beta_1}e^{-\gamma_1\tau_1}\right\},$$

$$\overline{\phi}_2(z)=\min\left\{\frac{\alpha_2}{\beta_2}e^{-\gamma_2\tau_2}e^{\lambda_1 z},\frac{\alpha_2}{\beta_2}e^{-\gamma_2\tau_2}\right\}.$$

We claim that $\overline{\phi}(z):=(\overline{\phi}_1(z),\overline{\phi}_2(z))^T$ is an upper solution of (3.12) and $\overline{\phi}\in\Gamma$.

Certainly, $\overline{\phi}\in\Gamma$ is clear. We need to verify the two differential inequalities obtained from (3.12) by replacing $=$ by $\leq$, and each needs to be checked separately for $z>0$ and $z<0$. For $\overline{\phi}_1$ we have two cases:

(i) If $z>0$, $\overline{\phi}_1=\frac{\alpha_1}{\beta_1}e^{-\gamma_1\tau_1}$, $\overline{\phi}_1(z-c\tau_1)\leq\frac{\alpha_1}{\beta_1}e^{-\gamma_1\tau_1}$, and $\overline{\phi}_2(z)=\frac{\alpha_2}{\beta_2}e^{-\gamma_2\tau_2}$. Then

$$d_1\overline{\phi}_1''(z)-c\overline{\phi}_1'(z)+\alpha_1 e^{-\gamma_1\tau_1}\overline{\phi}_1(z-c\tau_1)-\beta_1\overline{\phi}_1^2(z)-\frac{c_1\alpha_2}{\beta_2}e^{-\gamma_2\tau_2}\overline{\phi}_1(z)+c_1\overline{\phi}_1(z)\overline{\phi}_2(z)$$
$$\leq\alpha_1 e^{-\gamma_1\tau_1}\frac{\alpha_1}{\beta_1}e^{-\gamma_1\tau_1}-\beta_1\frac{\alpha_1^2}{\beta_1^2}e^{-2\gamma_1\tau_1}-\frac{c_1\alpha_2}{\beta_2}e^{-\gamma_2\tau_2}\frac{\alpha_1}{\beta_1}e^{-\gamma_1\tau_1}+c_1\frac{\alpha_1}{\beta_1}e^{-\gamma_1\tau_1}\frac{\alpha_2}{\beta_2}e^{-\gamma_2\tau_2}$$
$$=0.$$

(ii) If $z<0$, $\overline{\phi}_1=\frac{\alpha_1}{\beta_1}e^{-\gamma_1\tau_1}e^{\lambda_1 z}$, $\overline{\phi}_1(z-c\tau_1)=\frac{\alpha_1}{\beta_1}e^{-\gamma_1\tau_1}e^{\lambda_1(z-c\tau_1)}$, and $\overline{\phi}_2(z)=\frac{\alpha_2}{\beta_2}e^{-\gamma_2\tau_2}e^{\lambda_1 z}$. Thus

$$\begin{aligned}
&d_1\overline{\phi}_1''(z)-c\overline{\phi}_1'(z)+\alpha_1e^{-\gamma_1\tau_1}\overline{\phi}_1(z-c\tau_1)-\beta_1\overline{\phi}_1^2(z)-\frac{c_1\alpha_2}{\beta_2}e^{-\gamma_2\tau_2}\overline{\phi}_1(z)\\
&\quad+c_1\overline{\phi}_1(z)\overline{\phi}_2(z)\\
&=\frac{\alpha_1}{\beta_1}e^{-\gamma_1\tau_1}e^{\lambda_1 z}\left\{d_1\lambda_1^2-c\lambda_1+\alpha_1e^{-\gamma_1\tau_1}e^{-\lambda_1 c\tau_1}-\frac{c_1\alpha_2}{\beta_2}e^{-\gamma_2\tau_2}\right\}\\
&\quad+\frac{\alpha_1}{\beta_1}e^{-\gamma_1\tau_1}e^{2\lambda_1 z}\left\{\frac{c_1\alpha_2}{\beta_2}e^{-\gamma_2\tau_2}-\alpha_1e^{-\gamma_1\tau_1}\right\}\\
&<\frac{\alpha_1}{\beta_1}e^{-\gamma_1\tau_1}e^{\lambda_1 z}\,\Delta_1(\lambda_1)=0,
\end{aligned}$$

where we have used $\Delta_1(\lambda_1)=0$ and $\alpha_1e^{-\gamma_1\tau_1}>\frac{c_1\alpha_2}{\beta_2}e^{-\gamma_2\tau_2}$.

Therefore, we have proved

$$\begin{aligned}
&d_1\overline{\phi}_1''(z)-c\overline{\phi}_1'(z)+\alpha_1e^{-\gamma_1\tau_1}\overline{\phi}_1(z-c\tau_1)-\beta_1\overline{\phi}_1^2(z)-\frac{c_1\alpha_2}{\beta_2}e^{-\gamma_2\tau_2}\overline{\phi}_1(z)+c_1\overline{\phi}_1(z)\overline{\phi}_2(z)\\
&\le 0 \text{ (a.e.) on } \mathbf{R}.
\end{aligned}$$

For $\overline{\phi}_2$, again we need to check the cases $z>0$ and $z<0$ separately. The former case is trivial while, for $z<0$, we have

$$\begin{aligned}
&d_2\overline{\phi}_2''(z)-c\overline{\phi}_2'(z)+\alpha_2e^{-\gamma_2\tau_2}\overline{\phi}_2(z-c\tau_2)+\beta_2\overline{\phi}_2^2(z)-2\alpha_2e^{-\gamma_2\tau_2}\overline{\phi}_2(z)\\
&\quad+\frac{c_2\alpha_2}{\beta_2}e^{-\gamma_2\tau_2}\overline{\phi}_1(z)-c_2\overline{\phi}_1(z)\overline{\phi}_2(z)\\
&=\frac{\alpha_2}{\beta_2}e^{-\gamma_2\tau_2}e^{\lambda_1 z}\left\{d_2\lambda_1^2-c\lambda_1+\alpha_2e^{-\gamma_2\tau_2}e^{-\lambda_1 c\tau_2}-2\alpha_2e^{-\gamma_2\tau_2}+\frac{c_2\alpha_1}{\beta_1}e^{-\gamma_1\tau_1}\right\}\\
&\quad+\frac{\alpha_2}{\beta_2}e^{-\gamma_2\tau_2}e^{2\lambda_1 z}\underbrace{\left(-\frac{c_2\alpha_1}{\beta_1}e^{-\gamma_1\tau_1}+\alpha_2e^{-\gamma_2\tau_2}\right)}_{<0}\\
&<\frac{\alpha_2}{\beta_2}e^{-\gamma_2\tau_2}e^{\lambda_1 z}\left(\Delta_2(\lambda_1)-2\alpha_2e^{-\gamma_2\tau_2}+\frac{2c_2\alpha_1}{\beta_1}e^{-\gamma_1\tau_1}\right)\\
&\le 0,
\end{aligned}$$

where we have used (3.2) and (3.14). We have shown that $\overline{\phi}=(\overline{\phi}_1,\overline{\phi}_2)$ is an upper solution of (3.12).

Now let us construct a lower solution. Recall that λ_1 and λ_2 are the two real positive roots of $\Delta_1(\lambda)=0$. Now, let $\varepsilon>0$ be sufficiently small such that $\lambda_1<\lambda_1+\varepsilon<\lambda_2$ (so that $\Delta_1(\lambda_1+\varepsilon)<0$) and also such that $\lambda_1+\varepsilon\le 2\lambda_1$. Let $M>1$ be a number to be chosen later. Our candidate for a lower solution is

$$\underline{\phi}_1(z)=\begin{cases}(1-Me^{\varepsilon z})e^{\lambda_1 z}, & z<z_1,\\ 0, & z\ge z_1,\end{cases}\qquad \underline{\phi}_2(z)=0,$$

where $z_1=-(1/\varepsilon)\ln M<0$. Then $\underline{\phi}(z)\ge\mathbf{0}$ for all z. Recall that the lower solution is not required to lie in the profile set Γ.

We must verify that $\underline{\phi}_1$, $\underline{\phi}_2$ satisfy the differential inequalities obtained by replacing $=$ by $\ge$ in (3.12). The second such inequality is trivially satisfied. The first

differential inequality needs to be checked separately for the intervals $z > z_1 + c\tau_1$, $z_1 < z \le z_1 + c\tau_1$, and $z \le z_1$. It trivially holds in the first two of these intervals since in these cases we have $\underline{\phi}_1(z) = \underline{\phi}_1(z - c\tau_1) = 0$ and $\underline{\phi}_1(z) = 0$, $\underline{\phi}_1(z - c\tau_1) \ge 0$, respectively. When $z \le z_1$,

$$\begin{aligned}
&d_1\underline{\phi}_1''(z) - c\underline{\phi}_1'(z) + \alpha_1 e^{-\gamma_1\tau_1}\underline{\phi}_1(z - c\tau_1) - \beta_1\underline{\phi}_1^2(z) \\
&\quad - \frac{c_1\alpha_2}{\beta_2}e^{-\gamma_2\tau_2}\underline{\phi}_1(z) + c_1\underline{\phi}_1(z)\underline{\phi}_2(z) \\
&= e^{\lambda_1 z}\underbrace{\left(d_1\lambda_1^2 - c\lambda_1 + \alpha_1 e^{-\gamma_1\tau_1}e^{-\lambda_1 c\tau_1} - \frac{c_1\alpha_2}{\beta_2}e^{-\gamma_2\tau_2}\right)}_{=\Delta_1(\lambda_1)=0} \\
&\quad - M\,e^{(\lambda_1+\varepsilon)z}\left(d_1(\lambda_1+\varepsilon)^2 - c(\lambda_1+\varepsilon) + \alpha_1 e^{-\gamma_1\tau_1}e^{-(\lambda_1+\varepsilon)c\tau_1} - \frac{c_1\alpha_2}{\beta_2}e^{-\gamma_2\tau_2}\right) \\
&\quad - \beta_1 e^{2\lambda_1 z}(1 - Me^{\varepsilon z})^2 \\
&= -M\,e^{(\lambda_1+\varepsilon)z}\Delta_1(\lambda_1+\varepsilon) - \beta_1 e^{2\lambda_1 z}(1 - Me^{\varepsilon z})^2.
\end{aligned}$$

Now, since $z \le z_1 < 0$, we have $0 \le 1 - Me^{\varepsilon z} < 1$. Also, $2\lambda_1 \ge \lambda_1 + \varepsilon$ so that, since $z < 0$, $e^{2\lambda_1 z} \le e^{(\lambda_1+\varepsilon)z}$. Therefore

$$\begin{aligned}
&d_1\underline{\phi}_1''(z) - c\underline{\phi}_1'(z) + \alpha_1 e^{-\gamma_1\tau_1}\underline{\phi}_1(z - c\tau_1) - \beta_1\underline{\phi}_1^2(z) \\
&\quad - \frac{c_1\alpha_2}{\beta_2}e^{-\gamma_2\tau_2}\underline{\phi}_1(z) + c_1\underline{\phi}_1(z)\underline{\phi}_2(z) \\
&\ge -M\,e^{(\lambda_1+\varepsilon)z}\Delta_1(\lambda_1+\varepsilon) - \beta_1 e^{(\lambda_1+\varepsilon)z} \\
&= Me^{(\lambda_1+\varepsilon)z}\left(\underbrace{-\Delta_1(\lambda_1+\varepsilon)}_{>0} - \frac{\beta_1}{M}\right).
\end{aligned}$$

We now choose M sufficiently large to ensure (a) strict positivity of the right-hand side of the above, and (b) that

$$\sup_{z\in\mathbf{R}}\underline{\phi}_1(z) < \frac{\alpha_1}{\beta_1}e^{-\gamma_1\tau_1},$$

which is easily shown to be possible. We then need to arrange that $\underline{\phi}_1(z) \le \overline{\phi}_1(z)$ for all z, which is not automatically true. But note that $\underline{\phi}_1(z)$ and $\overline{\phi}_1(z)$ have the same exponential decay rate as $z \to -\infty$. Accordingly, we can arrange to have $\underline{\phi}_1(z) \le \overline{\phi}_1(z)$ for all z by replacing our upper solution $\overline{\phi}(z)$ by a leftward shifted translate $\overline{\phi}(z + A)$ thereof, for a suitably large value of $A > 0$. Since our problem is invariant to translations in z, any such translate of $\overline{\phi}$ is still an upper solution and, furthermore, is still in Γ.

We have now established the existence of an upper and a lower solution. We have mentioned that certain hypotheses of Wu and Zou do not quite apply to our problem, namely, the requirement that each component of the lower solution $\underline{\phi}$ be not identically zero and that there be no equilibria u with $\mathbf{0} < u < \mathbf{K}$. But Wu and Zou stress (see Remark 4.6 in [9]) that these very assumptions are required only for the final stage in their proof, i.e., that their solution satisfies $\phi(\infty) = \mathbf{K}$ and that any replacement which also ensures this is perfectly valid. Indeed, their construction of a solution to (3.4) is based on an iterative scheme which starts with the upper solution and

converges to a (nondecreasing) function $\phi(z)$ satisfying (3.4) and $\underline{\phi}(z) \le \phi(z) \le \overline{\phi}(z)$. These facts already imply $\phi(-\infty) = \mathbf{0}$.

In view of the above remarks, Wu and Zou's construction assures us of a solution to (3.12) satisfying $\phi(-\infty) = \mathbf{0}$, and we must now confirm that $\phi(\infty) = \mathbf{K}$ with $\mathbf{K}$ given by (3.11). Indeed (see [9, p. 666]), we can conclude, since $\underline{\phi}_1(z) \not\equiv 0$, that $\lim_{z\to\infty} \phi_1(z) := \phi_1^*$ exists and $\phi_1^* \in (0, \frac{\alpha_1}{\beta_1} e^{-\gamma_1\tau_1}]$. For $\phi_2(z)$, since $\underline{\phi}_2(z) \equiv 0$, we are still assured of the existence of $\lim_{z\to\infty} \phi_2(z) := \phi_2^*$, but we only know that $\phi_2^* \in [0, \frac{\alpha_2}{\beta_2} e^{-\gamma_2\tau_2}]$. However, (ϕ_1^*, ϕ_2^*) must be an equilibrium of (3.10). Thus

$$(\phi_1^*, \phi_2^*) = \left(\frac{\alpha_1}{\beta_1} e^{-\gamma_1\tau_1},\ \frac{\alpha_2}{\beta_2} e^{-\gamma_2\tau_2} \right).$$

The proof of the theorem is complete.

4. Conclusion. For the purely time dependent model studied in sections 1 and 2, it is apparent that the dynamics depends on the values of the four quantities

$$c_2\alpha_u \int_0^\tau f_u(s) e^{-\gamma_u s}\, ds, \qquad \beta_u\alpha_v \int_0^\tau f_v(s) e^{-\gamma_v s}\, ds, \qquad c_1\alpha_v \int_0^\tau f_v(s) e^{-\gamma_v s}\, ds,$$

and

$$\beta_v\alpha_u \int_0^\tau f_u(s) e^{-\gamma_u s}\, ds.$$

Accordingly, the dynamics depends heavily on the maturation delays as represented by the probability density functions $f_u(s)$ and $f_v(s)$. To see the role of these delays it will be helpful to consider the particular case when $f_u(s) = \delta(s-\tau_u)$ and $f_v(s) = \delta(s-\tau_v)$, so that all members of the U species take time τ_u to mature, and the V species take time τ_v to mature. Then, the criteria for U to win and V to be driven to extinction (Theorem 2) become

$$c_2\alpha_u e^{-\gamma_u\tau_u} > \beta_u\alpha_v e^{-\gamma_v\tau_v} \quad \text{and} \quad c_1\alpha_v e^{-\gamma_v\tau_v} < \beta_v\alpha_u e^{-\gamma_u\tau_u}.$$

These conditions are automatically satisfied if the V species has a long maturation time τ_v, a large immature mortality rate γ_v, or insufficient live births or eggs laid per adult per unit time (this is what α_v represents). Significant *adult* mortality among the V species, as measured by β_v, does not automatically imply extinction of that species. The two species will coexist if

$$c_2\alpha_u e^{-\gamma_u\tau_u} < \beta_u\alpha_v e^{-\gamma_v\tau_v} \quad \text{and} \quad c_1\alpha_v e^{-\gamma_v\tau_v} < \beta_v\alpha_u e^{-\gamma_u\tau_u},$$

and these conditions are satisfied when there is little interspecific competition and/or significant adult mortality in *both* species.

While the conclusions listed above are not counterintuitive, let us emphasize that they explicitly state how the survival or extinction of a species depends on the maturation time for the species, or more generally on a weighted average thereof.

For the reaction-diffusion model, we have proved the existence of a traveling wave front connecting the two boundary equilibria in the situation when the stronger species is sufficiently dominant that there is no possibility of coexistence. In addition to the conditions which one expects to have to impose, based on linearized analysis (i.e., conditions (3.2)), our existence proof requires (3.14) to be satisfied. In practice,

this will happen whenever (i) the diffusivity d_2 of the weaker competitor is relatively small and (ii) the parameter c_2, which measures the competitive pressure exerted by the stronger competitor on the weaker, is not too much greater than the minimum value consistent with (3.2). Smallness of d_2 (i.e., inability to move about quickly) can be interpreted as one of the weaker competitor's weaknesses. The fact that we have to restrict c_2 almost certainly has no ecological interpretation.

We have also calculated the minimum speed c^* implicitly. From simple graphical arguments one easily determines that the circumstances under which the minimum speed will be *reduced* are (i) if one or more of the parameters c_1, α_2, γ_1, or τ_1 is *increased*, or (ii) if one or more of the parameters d_1, α_1, β_2, γ_2, or τ_2 is *decreased*. Of course, one must ensure (3.2) remains satisfied.

Ecologically speaking, invasion of the domain by the dominant species therefore slows down under one or more of the following circumstances: reduction of diffusivity of the dominant species; reduction of the dominant species' reproductive activity; reduction of adult mortality, infant mortality, or maturation time for the weaker species. The invasion speed is also lowered if the weaker species increases its reproductive activity or exerts increased competitive pressure on the stronger, or if the stronger species suffers an increase in infant mortality or its maturation delay.

The minimum speed as determined from the linearized analysis does not depend on d_2 or c_2. Of course, the value of c_2 is important in determining whether conditions are right for colonization by the stronger competitor, driving the weaker to extinction. However, if these conditions are satisfied, then our analysis predicts that the actual invasion speed does not depend on c_2. Also, while the invasion speed depends strongly on the diffusivity d_1 of the stronger competitor, it does not depend at all on the diffusivity d_2 of the weaker.

REFERENCES

[1] W.G. Aiello and H.I. Freedman, *A time-delay model of single species growth with stage structure*, Math. Biosci., 101 (1990), pp. 139–153.

[2] J. Al-Omari and S.A. Gourley, *Monotone travelling fronts in an age-structured reaction-diffusion model of a single species*, J. Math. Biol., 45 (2002), pp. 294–312.

[3] K. Gopalsamy, *Time lags and global stability in two-species competition*, Bull. Math. Biol., 42 (1980), pp. 729–737.

[4] J. Huang and X. Zou, *Traveling wavefronts in diffusive and cooperative Lotka-Volterra system with delays*, J. Math. Anal. Appl., 271 (2002), pp. 455–466.

[5] R.H. Martin and H. Smith, *Abstract functional differential equations and reaction-diffusion systems*, Trans. Amer. Math. Soc., 321 (1990), pp. 1–44.

[6] J.D. Murray, *Mathematical Biology*, 2nd ed., Springer, Berlin, 1993.

[7] J. W.-H. So, J. Wu, and X. Zou, *A reaction-diffusion model for a single species with age structure.* I. *Travelling wavefronts on unbounded domains*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., 457 (2001), pp. 1841–1853.

[8] J. W.-H. So and X. Zou, *Travelling waves for the diffusive Nicholson's blowflies equation*, Appl. Math. Comput., 122 (2001), pp. 385–392.

[9] J. Wu and X. Zou, *Travelling wave fronts of reaction-diffusion systems with delay*, J. Dynam. Differential Equations, 13 (2001), pp. 651–687.

SIAM J. APPL. MATH.
Vol. 63, No. 6, pp. 2087–2127

A HIERARCHY OF MODELS FOR SUPERCONDUCTING THIN FILMS*

S. J. CHAPMAN† AND D. R. HERON†

Abstract. A hierarchy of models for type-II superconducting thin films is presented. Through appropriate asymptotic limits this hierarchy passes from the mesoscopic Ginzburg–Landau model to the London model with isolated vortices as δ-function singularities to vortex-density models and finally to macroscopic critical-state models. At each stage it is found that a key nondimensional parameter is $\Lambda = \lambda^2/dL$, where λ is the penetration depth of the magnetic field, a material parameter, and d and L are a typical thickness and lateral dimension of the film, respectively. The models simplify greatly if this parameter is large or small.

Key words. superconductivity, vortices, thin films, homogenization, critical state

AMS subject classification. 82D55

DOI. 10.1137/S0036139902410333

1. Introduction. The response of a bulk superconducting material to an applied magnetic field is conveniently described by Figure 1, which shows the phase the superconductor adopts as a function of the external magnetic field H_{ext} and the material parameter κ (known as the Ginzburg–Landau parameter), which determines the type of superconducting material; $\kappa < 1/\sqrt{2}$ describes what are known as type-I superconductors, while $\kappa > 1/\sqrt{2}$ describes what are known as type-II superconductors.

For type-I superconductors in sufficiently low magnetic fields the material is in the superconducting state, and the field is excluded from the interior of the sample except in thin boundary layers (this effect is known as the Meissner effect). However, there is a critical magnetic field, H_c, above which the material will revert to the normally conducting (normal) state, and the magnetic field will penetrate it fully.

In type-II superconductors this critical magnetic field splits into a lower critical field, H_{c_1}, and an upper critical field, H_{c_2}. For magnetic fields below H_{c_1} the material is in the superconducting state and the field is excluded from the interior, while for magnetic fields above H_{c_2} the material is in the normal state and the field penetrates it fully. For magnetic fields between H_{c_1} and H_{c_2} a third state exists, known as the "mixed state," in which there is a partial penetration of the magnetic field into the superconducting material, which occurs by means of thin filaments of nonsuperconducting material carrying magnetic flux ("flux tubes") and circled by a vortex of superconducting current (hence these filaments are often referred to as vortices).

A hierarchy of models for bulk type-II superconductors has been derived recently in [6]. The starting point for this hierarchy is the Ginzburg–Landau model, which applies on mesoscopic lengthscales $\sim 0.01\mu m$, and which is generally believed to describe the behavior of superconductors well (at least for low-temperature superconductors not too far from the critical temperature T_c). The Ginzburg–Landau model is quite complicated, and through the asymptotic limit $\kappa \to \infty$ may be simplified to the London model, a linear equation in which the vortices appear as line singularities. These

*Received by the editors June 26, 2002; accepted for publication (in revised form) March 13, 2003; published electronically September 17, 2003.
http://www.siam.org/journals/siap/63-6/41033.html
†Mathematical Institute, 24-29 St. Giles', Oxford OX1 3LB, UK (chapman@maths.ox.ac.uk, dale.heron@philips.com).

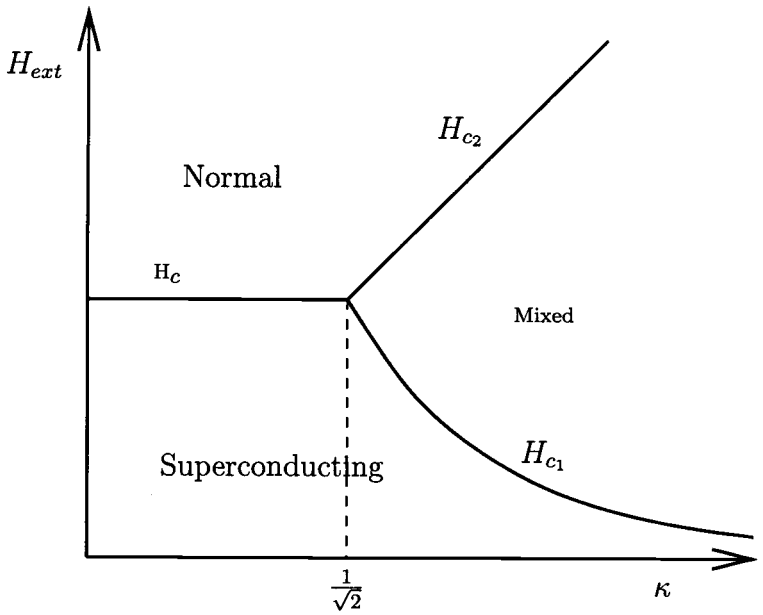


FIG. 1. *The response of a superconducting material as a function of the applied magnetic field H_{ext} and the Ginzburg–Landau parameter κ.*

may then be averaged to produce vortex-density models. Finally, if vortex pinning by inhomogeneities is included in these vortex-density models, so-called critical-state models can be derived.

Here we are interested in the simplifications which arise when the superconducting material comprises a thin film, possibly of varying thickness. Such a situation is very common experimentally and technologically because of the relative ease with which thin films can be manufactured. We will derive a hierarchy of models for superconducting thin films corresponding exactly to those for bulk superconductors. We will find that a key nondimensional parameter is $\Lambda = \lambda_{\text{eff}}/L$, where L is a typical lateral dimension of the film and the effective screening length [24] $\lambda_{\text{eff}} = \lambda^2/d$, where λ is the penetration depth of the magnetic field, a material parameter, and d is a typical thickness of the film. The canonical scaling is for Λ to be of order one as $d/L \to 0$, in which case the problems for the electric current in the film and the magnetic field outside it are coupled. The problem simplifies greatly if Λ is either large or small, since in each case the problems for the electric current and the magnetic field decouple. Many of the models we arrive at are new; our aims are to show where the existing models fit into the general framework and to fill in the gaps. In particular we will find that the thin-film limit of the Ginzburg–Landau model considered in [8] corresponds to the limit $\Lambda \to \infty$, while the thin-film critical-state models studied recently in [29] and [31] correspond to the limit $\Lambda \to 0$.

Before we begin let us first make a note of some of the lengthscales in the problem. There are two material parameters which are lengthscales, both of which appear in the Ginzburg–Landau model. These are λ, the aforementioned penetration depth, which is the typical lengthscale for the decay of magnetic field away from a vortex, and ξ, the coherence length, which is the typical lengthscale for the variation in the number density of superconducting electrons, and is the vortex core radius. The ratio of these

two lengthscales is the Ginzburg–Landau parameter $\kappa = \lambda/\xi$; as we have already said, for type-II superconductors $\kappa > 1/\sqrt{2}$, and the London model corresponds to the limit in which $\kappa \to \infty$. However, we will find that the important parameter in determining the behavior of thin films is not κ but $\kappa_{\text{eff}} = \lambda_{\text{eff}}/\xi = (\lambda/d)\kappa$. Thus when d is small compared to λ, the Ginzburg–Landau parameter is enhanced. The situation is not dissimilar to that of lubrication theory, in which the key parameter is the reduced Reynolds number rather than the Reynolds number itself.

To go with these two material lengths we have three geometrical lengths in the problem. Since we will allow the film to vary in thickness, besides the typical thickness d and lateral dimension L of the film we also have the typical lengthscale for the thickness variations, which we will denote by δ. (We assume that the amplitude of the thickness variations is of the same order as the thickness itself, which is the canonical case.) When pinning through inhomogeneities is introduced we have the typical lengthscale for variation of the pinning potential, which we denote by ε. Finally, we have the typical separation of vortices, which we denote by ν, and which will be determined by the strength of the applied magetic field.

Thus there are seven lengthscales in the problem, and it is the relative sizes of these which will determine which is the relevant thin-film model in any given situation.

In section 2 we introduce the Ginzburg–Landau theory which underpins our hierarchy of models and consider its thin-film limit. In section 3 we consider the London limit $\kappa_{\text{eff}} \to \infty$ of the thin-film Ginzburg–Landau model and show that this is the same as the thin-film limit of the bulk London model. In section 4 we let the separation of vortices $\nu/L \to 0$ and average the thin-film London model to produce a thin-film vortex-density model. We show that this is the same as the thin-film limit of the bulk vortex-density model.

In section 5 we let the lengthscale of the pinning potential $\varepsilon \to 0$ and homogenize the pinning force to produce a thin-film critical-state model. Finally, in section 6, we present our conclusions.

2. Ginzburg–Landau models. The starting point for our discussion of thin-film models of superconductivity is the Ginzburg–Landau equations. In their steady form these equations were written down by Ginzburg and Landau in [17], through the minimization of a phenomenologically developed free-energy functional. Later it was shown by Gor'kov [18] that they could be derived as a limit of the microscopic Bardeen, Cooper, and Schrieffer (BCS) model [2]. Time-dependent versions of the Ginzburg–Landau equations were written down by Schmidt [27], and in 1968 Gor'kov and Éliashberg [19] demonstrated that (1)–(2) could be derived from the BCS model for a superconductor with paramagnetic impurities. These correspond to a gauge-invariant gradient flow of the Ginzburg–Landau energy functional and as such are the simplest time-dependent equations whose solutions evolve to the minimizers of that functional. The dimensionless time-dependent Ginzburg–Landau equations for a superconducting material occupying a region $\Omega \subseteq \mathbb{R}^3$ are

$$\frac{1}{\kappa^2}\frac{\partial \Psi}{\partial t} + \frac{i\Psi\phi}{\kappa} = \left(\frac{1}{\kappa}\nabla - i\boldsymbol{A}\right)^2 \Psi + \Psi\left(1 - |\Psi|^2\right) \text{ in } \Omega, \tag{1}$$

$$-(\operatorname{curl})^2 \boldsymbol{A} - \frac{\sigma}{\kappa^2}\left(\frac{\partial \boldsymbol{A}}{\partial t} + \nabla\phi\right) = \frac{i}{2\kappa}\left(\Psi^*\nabla\Psi - \Psi\nabla\Psi^*\right) + |\Psi|^2\boldsymbol{A} \text{ in } \Omega, \tag{2}$$

where Ψ is the normalized superconducting order parameter, so that $|\Psi|^2$ represents the number density of superconducting electrons, with $|\Psi| = 1$ representing wholly

superconducting material and $\Psi = 0$ representing wholly nonsuperconducting (normal) material; $\boldsymbol{A}$ and ϕ are the magnetic vector potential and electric scalar potential, respectively, which are such that the magnetic field $\boldsymbol{H}$ and electric field $\boldsymbol{E}$ are given by

$$\boldsymbol{H} = \operatorname{curl} \boldsymbol{A}, \tag{3}$$

$$\boldsymbol{E} = -\frac{\partial \boldsymbol{A}}{\partial t} - \nabla \phi. \tag{4}$$

$\boldsymbol{A}$ is determined up to a gradient; once $\boldsymbol{A}$ is given ϕ is determined up to a function of t. The constant σ is a measure of the normal conductivity of the superconducting material (in this nondimensional form σ is the ratio of the timescale for diffusion of magnetic field in the normal state to the timescale for the relaxation of the order parameter), and κ is the Ginzburg–Landau parameter. Length has been scaled with the penetration depth λ, which is the natural lengthscale for variations in the magnetic field. In these units the thermodynamic critical field is given by $H_c = 1/\sqrt{2}$, and the upper and lower critical fields are given by $H_{c_2} = \kappa$ and $H_{c_1} \sim \log(\kappa)/2\kappa$ for large κ, respectively.

The most common boundary conditions on (1)–(2) are the natural conditions

$$\boldsymbol{n} \cdot (\nabla \Psi - i\boldsymbol{A}\Psi) = 0 \qquad \text{on } \partial\Omega \tag{5}$$

and continuity of $\boldsymbol{A}$, $\boldsymbol{H}$, and $\boldsymbol{E}$ across $\partial\Omega$ (assuming that the permeability and permittivity of the region exterior to the superconductor is equal to that of the superconducting material; the modification if this is not the case is easy to make). Equation (5) is applicable when the region adjacent to the superconductor is an insulator, which is the case we will consider henceforth. Note that (5) implies that no supercurrent passes through the boundary, and we assume also that no normal current is applied directly to the superconductor.

Outside Ω we have Maxwell's equations (neglecting displacement current)

$$\operatorname{curl} \boldsymbol{H} = \boldsymbol{J}_{ext}, \tag{6}$$

$$\operatorname{div} \boldsymbol{H} = 0, \tag{7}$$

$$\boldsymbol{H}_t + \operatorname{curl} \boldsymbol{E} = \boldsymbol{0}, \tag{8}$$

$$\operatorname{div} \boldsymbol{E} = 0, \tag{9}$$

where $\boldsymbol{J}_{ext}$ is the externally imposed current which is driving the system. The most commonly considered situation is that in which a uniform magnetic field is applied at infinity, in which case $\boldsymbol{J}_{ext}$ is zero and

$$\boldsymbol{H} \to \boldsymbol{H}_{ext} \text{ as } |\boldsymbol{x}| \to \infty \tag{10}$$

(assuming a bounded superconducting region). The continuities of $\boldsymbol{A}$, $\boldsymbol{H}$, and $\boldsymbol{E}$ are the usual boundary conditions on Maxwell's equations at an interface between two media. Note though that these conditions are not all independent, since, for example, the continuity of the normal component of $\boldsymbol{H}$ arises by taking (7) to hold in a generalized sense everywhere, but this equation is automatic from (3), which, if it holds in a generalized sense, gives continuity of the tangential components of $\boldsymbol{A}$.

Equations (1)–(2) are gauge invariant in the sense that they are invariant under transformations of the form

$$\boldsymbol{A} \to \boldsymbol{A} + \nabla \omega, \qquad \phi \to \phi - \frac{\partial \omega}{\partial t}, \qquad \Psi \to \Psi e^{i\kappa\omega}.$$

We take advantage of this invariance to write the equations in terms of real variables by writing

$$\Psi = fe^{i\chi}, \qquad \boldsymbol{Q} = \boldsymbol{A} - \frac{1}{\kappa}\nabla\chi, \qquad \Phi = \phi + \frac{1}{\kappa}\frac{\partial\chi}{\partial t} \tag{11}$$

to give

$$-\frac{1}{\kappa^2}\frac{\partial f}{\partial t} + \frac{1}{\kappa^2}\nabla^2 f = f^3 - f + f|\boldsymbol{Q}|^2 \quad \text{in } \Omega, \tag{12}$$

$$f^2\Phi + \operatorname{div}(f^2\boldsymbol{Q}) = 0 \quad \text{in } \Omega, \tag{13}$$

$$-(\operatorname{curl})^2\boldsymbol{Q} = \frac{\sigma}{\kappa^2}\left(\frac{\partial\boldsymbol{Q}}{\partial t} + \nabla\Phi\right) + f^2\boldsymbol{Q} \quad \text{in } \Omega, \tag{14}$$

$$\boldsymbol{H} = \operatorname{curl}\boldsymbol{Q}, \tag{15}$$

$$\boldsymbol{E} = -\frac{\partial\boldsymbol{Q}}{\partial t} - \nabla\Phi. \tag{16}$$

The vortex solutions characteristic of type-II superconductors can be illustrated by seeking a solution of the form

$$\Psi = f(r)e^{in\theta}, \tag{17}$$

$$\boldsymbol{A} = A(r)\boldsymbol{e}_\theta \tag{18}$$

on an infinite domain, where n is an integer known as the vortex number, r and θ are polar coordinates, and $\boldsymbol{e}_\theta$ is the unit vector in the azimuthal direction. Substituting into (1)–(2) gives

$$\frac{1}{\kappa^2}\frac{1}{r}\frac{d}{dr}\left(r\frac{df}{dr}\right) - \left(A - \frac{n}{\kappa r}\right)^2 f = f^3 - f, \tag{19}$$

$$\frac{d}{dr}\left(\frac{1}{r}\frac{d}{dr}(rA)\right) = f^2\left(A - \frac{n}{\kappa r}\right), \tag{20}$$

$$f, \quad A \text{ bounded} \quad \text{as } r \to 0, \tag{21}$$

$$f \to 1, \quad A \to 0 \text{ as } r \to \infty. \tag{22}$$

The existence of a solution which necessarily has $f(0) = 0$ has been shown by Berger and Chen [3]. The supercurrent is given by

$$\boldsymbol{J} = -f^2\left(A - \frac{n}{\kappa r}\right)\boldsymbol{e}_\theta, \tag{23}$$

which shows the vortex nature of this solution. The axial magnetic field carried by the vortex is

$$\int_{\mathbb{R}^2} \boldsymbol{H}\cdot\boldsymbol{dS} = \frac{2\pi n}{\kappa}, \tag{24}$$

which is quantized in units of $2\pi/\kappa$, with n the number of flux quanta carried by the vortex. Note that for large values of κ, $f \approx 1$ except in a region of order κ^{-1} from the origin, which is the vortex core.

Since the flux quantum is $O(1/\kappa)$ it is common to rescale magnetic field with $1/\kappa$ when considering vortex solutions. However, when considering the thin-film limit we must first rescale length so that we are working on the lateral dimension of the

sample, L say, which requires a corresponding rescaling of time with L^2/λ^2. The canonical scale for the magnetic field then involves rescaling with $\lambda^2/(L^2\kappa)$ (if we assume that the vortex separation is $O(L)$; see [10], for example), so that $\boldsymbol{Q}$ and $\boldsymbol{J}$ must be rescaled with $\lambda/(L\kappa)$. This gives

$$\frac{\xi^2}{L^2}\left(-\frac{\partial f}{\partial t}+\nabla^2 f\right)=f^3-f+\frac{\xi^2 f|\boldsymbol{Q}|^2}{L^2} \text{ in } \Omega, \tag{25}$$

$$f^2\Phi+\operatorname{div}(f^2\boldsymbol{Q})=0 \text{ in } \Omega, \tag{26}$$

$$-\lambda^2/L^2 \operatorname{curl}\boldsymbol{H}=-\boldsymbol{J}=\frac{\xi^2\sigma}{L^2}\left(\frac{\partial \boldsymbol{Q}}{\partial t}+\nabla\Phi\right)+f^2\boldsymbol{Q} \text{ in } \Omega, \tag{27}$$

$$\boldsymbol{H}=\operatorname{curl}\boldsymbol{Q}, \tag{28}$$

remembering that $\kappa=\lambda/\xi$. Note also the equations

$$\operatorname{div}\boldsymbol{H}=0, \tag{29}$$

$$\operatorname{div}\boldsymbol{J}=0, \tag{30}$$

which follow from (27) and (28) and will prove useful in the thin-film analysis. With these equations we have the boundary conditions

$$\frac{\partial f}{\partial n}=0 \text{ on } \partial\Omega, \tag{31}$$

$$\boldsymbol{Q}\cdot\boldsymbol{n}=0 \text{ on } \partial\Omega, \tag{32}$$

$$\boldsymbol{J}\cdot\boldsymbol{n}=0 \text{ on } \partial\Omega, \tag{33}$$

with $\boldsymbol{H}$ and $\boldsymbol{E}$ continuous across $\partial\Omega$.

2.1. Vortex pinning in the Ginzburg–Landau model. Before we go on to consider the thin-film version of (25)–(33) let us first describe the modifications which can be made to model the pinning of vortices in this framework.

In most technological applications superconductors are required to carry a transport current. The interaction of this current with the current circling a vortex causes the vortex to move. (This is often considered to be the result of the "Lorentz force" on the magnetic flux line carried by the vortex due to the transport current.) The motion of the vortex dissipates energy, leads to an electric field and hence a nonzero resistivity, and is therefore undesirable. In practice attempts are made to "pin" vortices at certain sites in the material in order to impede their motion. It is found that any form of inhomogeneity (for example, impurities, dislocations, or grain boundaries) will help to pin vortices. Such impurities have the effect of impeding locally the ability of the material to become superconducting. A popular way of modeling this inhomogeneity in the Ginzburg–Landau framework is to allow the equilibrium density of superconducting electrons to vary spatially [23, 7]. In the simplest case this leads to

$$\frac{\xi^2}{L^2}\left(-\frac{\partial f}{\partial t}+\nabla^2 f\right)=f^3-a(\boldsymbol{x})f+\frac{\xi^2 f|\boldsymbol{Q}|^2}{L^2} \text{ in } \Omega, \tag{34}$$

$$f^2\Phi+\operatorname{div}(f^2\boldsymbol{Q})=0 \text{ in } \Omega, \tag{35}$$

$$-\lambda^2/L^2 \operatorname{curl}\boldsymbol{H}=-\boldsymbol{J}=\frac{\xi^2\sigma}{L^2}\left(\frac{\partial \boldsymbol{Q}}{\partial t}+\nabla\Phi\right)+f^2\boldsymbol{Q} \text{ in } \Omega, \tag{36}$$

where the equilibrium density of superconducting electrons is denoted by $a(\boldsymbol{x})$. Of course, more generally we may allow the coefficient of $|\Psi|^2\Psi$ as well as λ, ξ, κ, and σ to vary spatially, but we will consider here the case when these parameters are constant.

Numerical simulations of these equations in two dimensions show that the vortices are attracted to minima of a, as we would have hoped [7, 14].

2.2. Thin-film limit of the Ginzburg–Landau model. Let the film be given, here and throughout, by

$$\Omega = \{(x,y,z) : (x,y) \in D, (\zeta - g/2)\epsilon < z < (\zeta + g/2)\epsilon\}, \tag{37}$$

where $D \subset \mathbb{R}^2$ is the projection of the film in the (x,y) plane, $\zeta(x,y)$ is the height of the centersurface of the film, $g(x,y)$ is the film thickness, and $\epsilon = d/L$ is the aspect ratio.

We write $\xi/L = \Xi$ and $\lambda^2/L^2 = \epsilon\Lambda$, since the canonical scaling will turn out to be $\Lambda = O(1)$. Then

$$\Xi^2\left(-\frac{\partial f}{\partial t} + \nabla^2 f\right) = f^3 - a(\boldsymbol{x})f + \Xi^2 f|\boldsymbol{Q}|^2 \text{ in } \Omega, \tag{38}$$

$$f^2\Phi + \operatorname{div}(f^2\boldsymbol{Q}) = 0 \text{ in } \Omega, \tag{39}$$

$$-\epsilon\Lambda \operatorname{curl} \boldsymbol{H} = -\boldsymbol{J} = \Xi^2\sigma\left(\frac{\partial \boldsymbol{Q}}{\partial t} + \nabla\Phi\right) + f^2\boldsymbol{Q} \text{ in } \Omega. \tag{40}$$

We consider first the problem in the film, where we rescale z with ϵ by setting $z = \epsilon Z$. Henceforth for clarity we use $\boldsymbol{h}$ to denote the magnetic field outside the film and $\boldsymbol{H}$ to denote it inside the film. Then

$$\Xi^2\left(-\frac{\partial f}{\partial t} + \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{1}{\epsilon^2}\frac{\partial^2 f}{\partial Z^2}\right) = f^3 - f + \Xi^2 f(Q_1^2 + Q_2^2 + Q_3^2) \text{ in } \Omega, \tag{41}$$

$$f^2\Phi + \frac{\partial(f^2Q_1)}{\partial x} + \frac{\partial(f^2Q_2)}{\partial y} + \frac{1}{\epsilon}\frac{\partial(f^2Q_3)}{\partial Z} = 0 \text{ in } \Omega, \tag{42}$$

$$-\epsilon\Lambda\left(\frac{\partial H_3}{\partial y} - \frac{1}{\epsilon}\frac{\partial H_2}{\partial Z}\right) = -J_1 = \Xi^2\sigma\left(\frac{\partial Q_1}{\partial t} + \frac{\partial\Phi}{\partial x}\right) + f^2Q_1 \text{ in } \Omega, \tag{43}$$

$$-\epsilon\Lambda\left(\frac{1}{\epsilon}\frac{\partial H_1}{\partial Z} - \frac{\partial H_3}{\partial x}\right) = -J_2 = \Xi^2\sigma\left(\frac{\partial Q_2}{\partial t} + \frac{\partial\Phi}{\partial y}\right) + f^2Q_2 \text{ in } \Omega, \tag{44}$$

$$-\epsilon\Lambda\left(\frac{\partial H_2}{\partial x} - \frac{\partial H_1}{\partial y}\right) = -J_3 = \Xi^2\sigma\left(\frac{\partial Q_3}{\partial t} + \frac{1}{\epsilon}\frac{\partial\Phi}{\partial Z}\right) + f^2Q_3 \text{ in } \Omega, \tag{45}$$

$$\frac{\partial Q_3}{\partial y} - \frac{1}{\epsilon}\frac{\partial Q_2}{\partial Z} = H_1, \tag{46}$$

$$\frac{1}{\epsilon}\frac{\partial Q_1}{\partial Z} - \frac{\partial Q_3}{\partial x} = H_2, \tag{47}$$

$$\frac{\partial Q_2}{\partial x} - \frac{\partial Q_1}{\partial y} = H_3, \tag{48}$$

$$\frac{\partial H_1}{\partial x} + \frac{\partial H_2}{\partial y} + \frac{1}{\epsilon}\frac{\partial H_3}{\partial Z} = 0, \tag{49}$$

$$\frac{\partial J_1}{\partial x} + \frac{\partial J_2}{\partial y} + \frac{1}{\epsilon}\frac{\partial J_3}{\partial Z} = 0. \tag{50}$$

With these equations we have the boundary conditions on the upper and lower surfaces of the film

$$-\epsilon^2\frac{\partial}{\partial x}(\zeta \pm g/2)\frac{\partial f}{\partial x}(x,y,\zeta\pm g/2) - \epsilon^2\frac{\partial}{\partial y}(\zeta\pm g/2)\frac{\partial f}{\partial y}(x,y,\zeta\pm g/2) + \frac{\partial f}{\partial Z}(x,y,\zeta\pm g/2) = 0 \text{ for } (x,y)\in D, \tag{51}$$

$$-\epsilon\frac{\partial}{\partial x}(\zeta\pm g/2)\,Q_1(x,y,\zeta\pm g/2) - \epsilon\frac{\partial}{\partial y}(\zeta\pm g/2)\,Q_2(x,y,\zeta\pm g/2) + Q_3(x,y,\zeta\pm g/2) = 0 \text{ for } (x,y)\in D, \tag{52}$$

$$-\epsilon\frac{\partial}{\partial x}(\zeta\pm g/2)\,J_1(x,y,\zeta\pm g/2) - \epsilon\frac{\partial}{\partial y}(\zeta\pm g/2)\,J_2(x,y,\zeta\pm g/2) + J_3(x,y,\zeta\pm g/2) = 0 \text{ for } (x,y)\in D, \tag{53}$$

$$\boldsymbol{h}(x,y,z=\epsilon(\zeta\pm g/2)) = \boldsymbol{H}(x,y,Z=\zeta\pm g/2). \tag{54}$$

We formally expand all variables in a power series in ϵ as $\epsilon\to 0$ as

$$\boldsymbol{H} = \boldsymbol{H}^{(0)} + \epsilon\boldsymbol{H}^{(1)} + \cdots,$$

etc. Then at leading order in (41)–(50)

$$\frac{\partial^2 f^{(0)}}{\partial Z^2} = 0 \text{ in } \Omega, \tag{55}$$

$$\frac{\partial((f^{(0)})^2 Q_3^{(0)})}{\partial Z} = 0 \text{ in } \Omega, \tag{56}$$

$$\Lambda\frac{\partial H_2^{(0)}}{\partial Z} = -J_1^{(0)} = \Xi^2\sigma\left(\frac{\partial Q_1^{(0)}}{\partial t} + \frac{\partial\Phi^{(0)}}{\partial x}\right) + (f^{(0)})^2 Q_1^{(0)} \text{ in } \Omega, \tag{57}$$

$$-\Lambda\frac{\partial H_1^{(0)}}{\partial Z} = -J_2^{(0)} = \Xi^2\sigma\left(\frac{\partial Q_2^{(0)}}{\partial t} + \frac{\partial\Phi^{(0)}}{\partial y}\right) + (f^{(0)})^2 Q_2^{(0)} \text{ in } \Omega, \tag{58}$$

$$0 = \frac{\partial\Phi^{(0)}}{\partial Z} \text{ in } \Omega, \tag{59}$$

$$\frac{\partial Q_2^{(0)}}{\partial Z} = 0, \tag{60}$$

$$\frac{\partial Q_1^{(0)}}{\partial Z} = 0, \tag{61}$$

$$\frac{\partial Q_2^{(0)}}{\partial x} - \frac{\partial Q_1^{(0)}}{\partial y} = H_3^{(0)}, \tag{62}$$

$$\frac{\partial H_3^{(0)}}{\partial Z} = 0, \tag{63}$$

$$\frac{\partial J_3^{(0)}}{\partial Z} = 0, \tag{64}$$

with boundary conditions

$$\frac{\partial f^{(0)}}{\partial Z}(x,y,\zeta\pm g/2) = 0 \text{ for } (x,y)\in D, \tag{65}$$

$$Q_3^{(0)}(x,y,\zeta \pm g/2) = 0 \text{ for } (x,y) \in D, \tag{66}$$

$$J_3^{(0)}(x,y,\zeta \pm g/2) = 0 \text{ for } (x,y) \in D, \tag{67}$$

$$\boldsymbol{h}^{(0)}(x,y,z=\pm 0) = \boldsymbol{H}^{(0)}(x,y,Z=\zeta \pm g/2). \tag{68}$$

Equations (55) and (65) give $f^{(0)} = f^{(0)}(x,y,t)$. From (64) and (67) we see that $J_3^{(0)} = 0$. We see from (59)–(61) that $Q_1^{(0)}$, $Q_2^{(0)}$, and $\Phi^{(0)}$ are independent of Z, so that $J_1^{(0)}$ and $J_2^{(0)}$ are independent of Z. Then (57), and (58) may be integrated to give

$$H_1^{(0)} = \frac{ZJ_2^{(0)}}{\Lambda} + a(x,y,t), \tag{69}$$

$$H_2^{(0)} = -\frac{ZJ_1^{(0)}}{\Lambda} + b(x,y,t). \tag{70}$$

Similarly (63) gives

$$H_3^{(0)} = c(x,y,t). \tag{71}$$

Now evaluating on the top and bottom of the film $Z = \zeta \pm g/2$ and using (68) give

$$[\boldsymbol{h}^{(0)}] = \frac{g}{\Lambda}(J_2^{(0)}, -J_1^{(0)}, 0) \quad \text{for} \quad (x,y) \in D, \tag{72}$$

where the square bracket indicates the jump in the quantity enclosed across D. Writing this jump condition in the more usual form gives

$$\left[\boldsymbol{e}_z \wedge \boldsymbol{h}^{(0)}\right] = \frac{g}{\Lambda}\boldsymbol{J}^{(0)} \quad \text{for} \quad (x,y) \in D, \tag{73}$$

$$\left[\boldsymbol{e}_z \cdot \boldsymbol{h}^{(0)}\right] = 0 \quad \text{for} \quad (x,y) \in D, \tag{74}$$

where $\boldsymbol{e}_z$ is the unit vector in the z-direction. Equations (73)–(74) form boundary conditions on the problem for the external magnetic field, which satisfies Maxwell's equations

$$\operatorname{curl} \boldsymbol{h}^{(0)} = \boldsymbol{J}_{ext}^{(0)}, \tag{75}$$

$$\operatorname{div} \boldsymbol{h}^{(0)} = 0, \tag{76}$$

$$\boldsymbol{h}^{(0)} \to \boldsymbol{h}_{ext}^{(0)} \quad \text{as } |\boldsymbol{r}| \to \infty. \tag{77}$$

As we might expect, $\boldsymbol{h}^{(0)}$ is simply the magnetic field generated if the total current in the film were distributed on a sheet, plus the applied (externally generated) magnetic field. We can therefore calculate $\boldsymbol{h}^{(0)}$ once we have found $\boldsymbol{J}^{(0)}$.

Returning to the film, we see from (56) and (66) that $Q_3^{(0)} = 0$. What remains is to find an equation for $f^{(0)}$, $Q_1^{(0)}$, $Q_2^{(0)}$, and $\Phi^{(0)}$. To do this we need to proceed to higher orders in the expansion of (41) and (51).

At first order in (41) and (51) we find

$$\frac{\partial f^{(1)}}{\partial Z} = 0. \tag{78}$$

At second order we find

$$\Xi^2\left(-\frac{\partial f^{(0)}}{\partial t} + \frac{\partial^2 f^{(0)}}{\partial x^2} + \frac{\partial^2 f^{(0)}}{\partial y^2} + \frac{\partial^2 f^{(2)}}{\partial Z^2}\right) = (f^{(0)})^3 - a(\boldsymbol{x})f^{(0)} + \Xi^2 f^{(0)}((Q_1^{(0)})^2 + (Q_2^{(0)})^2) \text{ in } \Omega, \tag{79}$$

with

$$-\frac{\partial}{\partial x}(\zeta \pm g/2)\frac{\partial f^{(0)}}{\partial x}(x,y,\zeta \pm g/2) - \frac{\partial}{\partial y}(\zeta \pm g/2)\frac{\partial f^{(0)}}{\partial y}(x,y,\zeta \pm g/2) + \frac{\partial f^{(2)}}{\partial Z}(x,y,\zeta \pm g/2) = 0 \text{ for } (x,y) \in D. \tag{80}$$

Integrating (79) from $\zeta - g/2$ to $\zeta + g/2$ and using (80) give

$$\Xi^2\left(-\frac{\partial f^{(0)}}{\partial t} + \frac{1}{g}\nabla\cdot\left(g\nabla f^{(0)}\right)\right) = (f^{(0)})^3 - f^{(0)} + \Xi^2 f^{(0)}((Q_1^{(0)})^2 + (Q_2^{(0)})^2) \text{ for } (x,y) \in D. \tag{81}$$

We are still missing an equation for the divergence of $\boldsymbol{Q}^{(0)}$, which should come from (42). Proceeding to next order in this equation gives

$$(f^{(0)})^2\Phi^{(0)} + \frac{\partial((f^{(0)})^2 Q_1^{(0)})}{\partial x} + \frac{\partial((f^{(0)})^2 Q_2^{(0)})}{\partial y} + \frac{\partial((f^{(0)})^2 Q_3^{(1)})}{\partial Z} = 0 \text{ in } \Omega. \tag{82}$$

At next order in the boundary condition (52) we find

$$-\frac{\partial}{\partial x}(\zeta \pm g/2)\, Q_1^{(0)}(x,y,\zeta \pm g/2) - \frac{\partial}{\partial y}(\zeta \pm g/2)\, Q_2^{(0)}(x,y,\zeta \pm g/2) + Q_3^{(1)}(x,y,\zeta \pm g/2) = 0 \text{ for } (x,y) \in D. \tag{83}$$

Now, integrating (82) from $Z = \zeta - g/2$ to $Z = \zeta + g/2$ and using (83) gives

$$(f^{(0)})^2\Phi^{(0)} + \frac{1}{g}\nabla\cdot\left(g(f^{(0)})^2\boldsymbol{Q}^{(0)}\right) = 0 \text{ for } (x,y) \in D. \tag{84}$$

A similar analysis on (50) and (53) gives

$$\nabla\cdot\left(g\boldsymbol{J}^{(0)}\right) = 0. \tag{85}$$

Finally, by continuity of magnetic field at the interface (68) and constancy of H_3 in Z (71), equation (62) becomes

$$\frac{\partial Q_2^{(0)}}{\partial x} - \frac{\partial Q_1^{(0)}}{\partial y} = h_3^{(0)}(x,y,0). \tag{86}$$

Summary. We now have a closed model for the leading-order problem as $\epsilon \to 0$ with Ξ and Λ fixed. Dropping the superscripts for clarity we have

$$\Xi^2\left(-\frac{\partial f}{\partial t} + \frac{1}{g}\nabla\cdot(g\nabla f)\right) = f^3 - a(\boldsymbol{x})f + \Xi^2 f|\boldsymbol{Q}|^2 \quad \text{for} \quad (x,y) \in D, \tag{87}$$

$$f^2\Phi + \frac{1}{g}\nabla\cdot\left(gf^2\boldsymbol{Q}\right) = 0 \quad \text{for} \quad (x,y) \in D, \tag{88}$$

$$\frac{\partial Q_2}{\partial x} - \frac{\partial Q_1}{\partial y} = h_3(x,y,0) \quad \text{for} \quad (x,y) \in D, \tag{89}$$

$$Q_3 = 0, \tag{90}$$

$$\nabla \cdot (g\boldsymbol{J}) = 0, \tag{91}$$

$$\boldsymbol{J} = -\Xi^2\sigma\left(\frac{\partial \boldsymbol{Q}}{\partial t} + \nabla\Phi\right) - f^2\boldsymbol{Q} \quad \text{for} \quad (x,y) \in D, \tag{92}$$

$$J_3 = 0, \tag{93}$$

$$[\boldsymbol{e}_z \wedge \boldsymbol{h}] = \frac{g}{\Lambda}\boldsymbol{J} \quad \text{for} \quad (x,y) \in D, \tag{94}$$

$$[\boldsymbol{e}_z \cdot \boldsymbol{h}] = 0 \quad \text{for} \quad (x,y) \in D, \tag{95}$$

$$\operatorname{curl} \boldsymbol{h} = \boldsymbol{J}_{ext}, \tag{96}$$

$$\operatorname{div} \boldsymbol{h} = 0, \tag{97}$$

$$\boldsymbol{h} \to \boldsymbol{h}_{ext} \quad \text{as} \quad |\boldsymbol{x}| \to \infty. \tag{98}$$

If g is constant, then (87)–(93) are simply the two-dimensional Ginzburg–Landau equations. The big difference in the thin-film case is that h_3 is not related to $\boldsymbol{J}$ through $\boldsymbol{J} = \operatorname{curl}(0,0,h_3)$ but through the exterior problem (94)–(98).

We note that there is an equivalent complex formulation

$$\Xi^2\left(-\frac{\partial\Psi}{\partial t} - i\Psi\phi + \frac{1}{g}(\nabla - i\boldsymbol{A})\cdot g\,(\nabla - i\boldsymbol{A})\,\Psi\right) = \Psi(|\Psi|^2 - a(\boldsymbol{x})) \text{ for } (x,y) \in D, \tag{99}$$

$$\frac{\partial A_2}{\partial x} - \frac{\partial A_1}{\partial y} = h_3(x,y,0) \quad \text{for} \quad (x,y) \in D, \tag{100}$$

$$A_3 = 0, \tag{101}$$

$$[\boldsymbol{e}_z \wedge \boldsymbol{h}] = \frac{g}{\Lambda}\boldsymbol{J} \quad \text{for} \quad (x,y) \in D, \tag{102}$$

$$[\boldsymbol{e}_z \cdot \boldsymbol{h}] = 0 \quad \text{for} \quad (x,y) \in D, \tag{103}$$

$$\nabla \cdot (g\boldsymbol{J}) = 0, \tag{104}$$

$$-\Xi^2\sigma\left(\frac{\partial \boldsymbol{A}}{\partial t} + \nabla\phi\right) + \frac{1}{2}(\Psi^*\nabla\Psi - \Psi\nabla\Psi^*) - |\Psi|^2\boldsymbol{A} = \boldsymbol{J} \quad \text{for} \quad (x,y) \in D, \tag{105}$$

$$J_3 = 0, \tag{106}$$

$$\operatorname{curl} \boldsymbol{h} = \boldsymbol{J}_{ext}, \tag{107}$$

$$\operatorname{div} \boldsymbol{h} = 0, \tag{108}$$

$$\boldsymbol{h} \to \boldsymbol{h}_{ext} \quad \text{as} \quad |\boldsymbol{x}| \to \infty, \tag{109}$$

where

$$\Psi = fe^{i\chi}, \qquad \boldsymbol{Q} = \boldsymbol{A} - \nabla\chi, \qquad \Phi = \phi + \frac{\partial\chi}{\partial t}, \tag{110}$$

and χ is arbitrary. To these equations we must add the boundary conditions (5) (or equivalently (31)–(32)) and (33) on the lateral edges of the film, giving

$$\boldsymbol{\nu} \cdot (\nabla\Psi - i\boldsymbol{A}\Psi) = 0 \qquad \text{on } \partial D, \tag{111}$$

$$\boldsymbol{J} \cdot \boldsymbol{\nu} = 0 \quad \text{on } \partial D, \tag{112}$$

where $\boldsymbol{\nu}$ is the unit outward normal to ∂D.

Before proceeding let us first examine the flow of information in the equations, and in particular whether we have the right number of equations and unknowns. As we have already mentioned, equations (102)–(103) and (107)–(109) form a standard problem in magnetostatics, namely to determine the magnetic field generated by a surface current sheet. Thus these equations determine $\boldsymbol{h}$ if $\boldsymbol{J}$ is given. Having determined $\boldsymbol{h}$, equations (100)–(101) then determine $\boldsymbol{A}$ up to a gradient, which we can fix due to the gauge invariance property (110). With $\boldsymbol{A}$ given, equations (99) and (104) (using (105)) with the boundary conditions (111)–(112) form a closed system for Ψ and ϕ. Finally, the loop is closed by (105), which gives $\boldsymbol{J}$ as a function of Ψ, ϕ, and $\boldsymbol{A}$.

2.2.1. Thin strip geometry. Suppose our thin film is a thin strip of width $2l$, i.e., $D = (x, y) : -l < x < l$, and that $\boldsymbol{J}_{ext}$ and $\boldsymbol{h}_{ext}$ are such that $\boldsymbol{h} = (h_1, 0, h_3)$, $\boldsymbol{J} = (0, J_2, 0)$, with h_1, h_3, and J_2 independent of y. Then we can write the relationship between h_3 and J_2 explicitly using a Hilbert transform as follows. We first write $\boldsymbol{h} = \boldsymbol{h}_{app} + \boldsymbol{h}'$, where $\boldsymbol{h}_{app}$ is the total applied magnetic field (generated by $\boldsymbol{J}_{ext}$ and $\boldsymbol{h}_{ext}$) so that $\boldsymbol{h}'$ satisfies the homogeneous version of (107) and tends to zero at infinity. Then (107) and (108) imply that we can write

$$\boldsymbol{h}' = \nabla w, \tag{113}$$

$$\nabla^2 w = 0, \tag{114}$$

where ∇ is the two-dimensional gradient in x and z, with

$$\left[\frac{\partial w}{\partial x}\right] = \frac{gJ_2}{\Lambda}, \tag{115}$$

$$\left[\frac{\partial w}{\partial z}\right] = 0, \tag{116}$$

$$w \to 0 \quad \text{as } x^2 + z^2 \to \infty. \tag{117}$$

Since the solution to (114)–(117) will be symmetric, so that

$$w(x, z) = -w(x, -z), \tag{118}$$

we may consider the problem in the upper half-plane only, giving

$$\nabla^2 w = 0 \quad \text{in } z > 0, \tag{119}$$

$$\frac{\partial w}{\partial x} = \frac{gJ_2}{2\Lambda} \quad \text{on } z = 0, \tag{120}$$

$$\frac{\partial w}{\partial z} = h_3(x, 0) - h_{app,3}(x, 0) \quad \text{on } z = 0, \tag{121}$$

$$w \to 0 \quad \text{as } x^2 + z^2 \to \infty. \tag{122}$$

Then

$$\left.\frac{\partial w}{\partial z}\right|_{z=0} = \mathcal{H}\left(\left.\frac{\partial w}{\partial x}\right|_{z=0}\right) \equiv \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{1}{\bar{x} - x}\frac{\partial w}{\partial x}(\bar{x}, 0)\, d\bar{x}, \tag{123}$$

so that

$$h_3(x, 0) - h_{app,3}(x, 0) = \frac{1}{\pi}\int_{-l}^{l} \frac{gJ_2(\bar{x})}{2\Lambda(\bar{x} - x)}\, d\bar{x}. \tag{124}$$

This equation replaces (102)–(103) and (107)–(109) in (99)–(112).

2.2.2. The limit $\Lambda \to \infty$ with Ξ fixed. There are three key parameters left in the model, namely Λ, Ξ, and σ. We have so far considered the canonical scaling

in which all the equations remain coupled in the thin-film limit. The thin-film limit of the Ginzburg–Landau equations considered in [8] corresponds to the limit $\Lambda \to \infty$ with Ξ fixed, since they consider the limit $\epsilon \to 0$ with fixed λ, ξ, L, and κ. As $\Lambda \to \infty$ we see from (94) that the current in the film is insufficient to affect the magnetic field at leading order, and the problems for $\boldsymbol{A}$ and Ψ decouple leading to a great simplification. At leading order the applied magnetic field simply passes straight through the film, so that

$$\boldsymbol{h} = \boldsymbol{h}_{app}, \qquad \boldsymbol{A} = (0, xH_{3,app}, 0),$$

where $\boldsymbol{h}_{app}$ is the total applied magnetic field (that generated by the external current plus that imposed at infinity), which is the solution to (107)–(109). Equations (99) and (104) can then be solved for Ψ and ϕ since $\boldsymbol{A}$ is known. Note that the factor Ξ^2 may be removed from these equations by a rescaling of $\boldsymbol{x}$ and $\boldsymbol{A}$, as is done in [8]. The current in the film is then determined from (105). The correction to the magnetic field can then be calculated from (102)–(103) and (107)–(109). Thus in this expansion for large Λ at each stage we have to solve the standard magnetostatics problem of determining the magnetic field due to a known surface current sheet and then a problem for Ψ and ϕ in which $\boldsymbol{A}$ is known.

The numerical solutions in [8] show that vortices are attracted to the minima of $g(x)$, that is, the thin parts of the film. Figure 2 shows a typical calculation, performed on a square film with side of (nondimensional) length 20 ($L = \xi$ is chosen as the unit of length, so dimensionally the side of the square is 20ξ), with applied field perpendicular to the film and of (nondimensional) strength 0.5. In Figure 2(a) the solution for a film of constant thickness is shown. (Contours show the level curves of the modulus of the order parameter.) Ten squares of size $\xi \times \xi$ were then chosen randomly, and the thickness of the film on the squares were reduced to half that of the remaining film; the distribution of the thin regions is shown in Figure 2(b). Figure 2(c) shows the position of the vortices for the variable thickness film with the same magnetic field strength as Figure 2(a). Figure 2(d) shows Figures 2(b) and 2(c) superimposed, so that the position of the vortices can be compared to the position of the thin regions of the film.

Numerical solutions of the time-dependent version of this reduced model are also performed in [1] to study vortex nucleation at boundaries in the presence of applied magnetic fields and currents.

The limit $\Lambda \to 0$ *with* Ξ *fixed.* The alternative limit is $\Lambda \to 0$ with Ξ fixed. In this case, from (94) and (92), we must rescale $\boldsymbol{J}$, $\boldsymbol{Q}$, and Φ with Λ. Then to leading order (89) gives

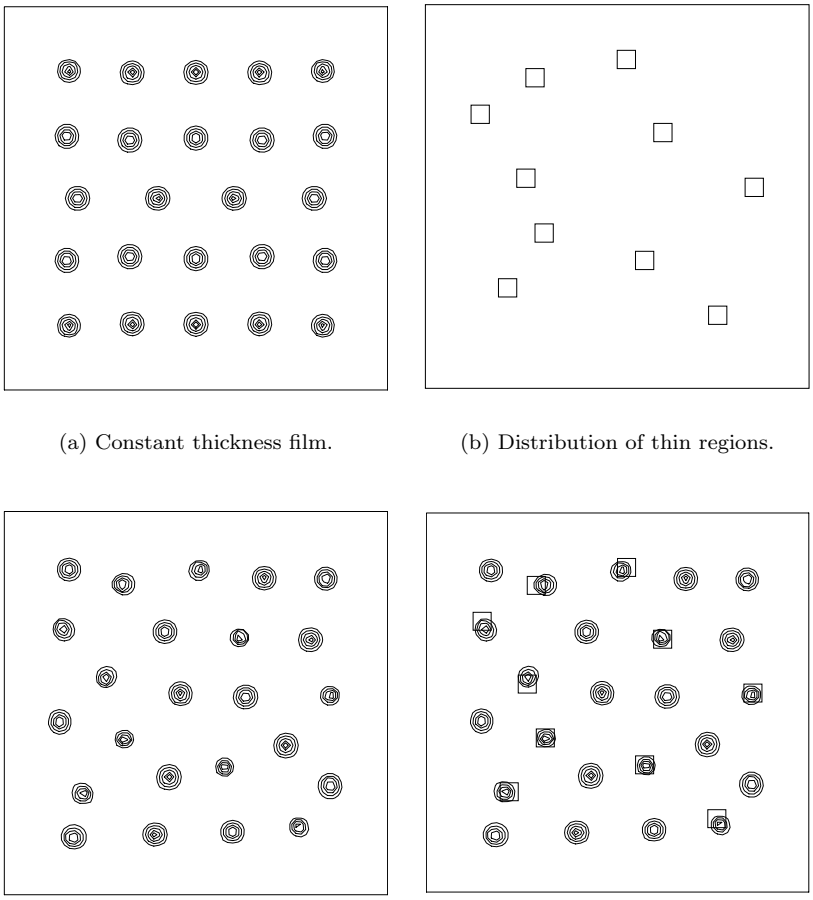
$$h_3(x, y, 0) = 0 \text{ for } (x, y) \in D. \tag{125}$$

Equations (96)–(98) can then be solved with this boundary condition for $\boldsymbol{h}$.

Equation (125) is valid away from vortices. Near each vortex an inner problem must be solved in which length is rescaled with Λ, which couples the problems for $\boldsymbol{h}$ and $\boldsymbol{J}$ together again. The limit $\Lambda \to 0$ will be much more interesting when we go on to consider vortex density models in section 4.

3. London models.

3.1. The London limit of the thin-film Ginzburg–Landau model. From (87) we see that the vortex core radius is of order Ξ in these units. The London model, corresponding to vanishing core radius, therefore corresponds to the limit



(a) Constant thickness film.

(b) Distribution of thin regions.

(c) Variable thickness film.

(d) Figs. 2(b) and 2(c) superimposed.

FIG. 2. *Level curves of the magnitude of the order parameter for superconducting samples having sides equal to* 20 *coherence lengths* [S. J. Chapman, Q. Du, and M. D. Gunzburger, *A model for variable thickness superconducting thin films,* Z. Angew. Math. Phys., 47 (1996), pp. 410–431]. Reprinted with permission.

$\Xi \to 0$. Since we are taking this limit after having let $\epsilon \to 0$ we are in the parameter regime $\epsilon \ll \Xi \ll 1$. We will consider the alternative regime $\Xi \ll \epsilon \ll 1$ in section 3.3.

In the limit $\Xi \to 0$ with Λ fixed we see that (87) implies $f^2 = a(\boldsymbol{x})$ except at vortices. Then from (92) we have

$$\boldsymbol{J} = -a\boldsymbol{Q}, \tag{126}$$

so that (89) gives

$$\boldsymbol{e}_z \cdot \operatorname{curl}\left(\frac{\boldsymbol{J}}{a}\right) + h_3(x, y, 0) = 0 \tag{127}$$

except at vortices.

Suppose there is a vortex at the origin. Near the vortex we rescale $r = (x^2 + y^2)^{1/2} = \Xi R$ and expand

$$f = f^{(0)} + \cdots, \tag{128}$$

$$\boldsymbol{Q} = \frac{1}{\Xi}\boldsymbol{Q}^{(0)} + \cdots. \tag{129}$$

Then at leading order in the core region

$$\boldsymbol{Q}^{(0)} = -\frac{1}{R}\boldsymbol{e}_\theta, \tag{130}$$

$$\frac{\partial^2 f^{(0)}}{\partial R^2} + \frac{1}{R}\frac{\partial f^{(0)}}{\partial R} - \frac{f^{(0)}}{R^2} = (f^{(0)})^3 - a(0)f^{(0)}. \tag{131}$$

Writing

$$f^{(0)} = \sqrt{a(0)}\bar{f}(\rho), \qquad \rho = \sqrt{a(0)}R,$$

as in [12] we find $\bar{f}$ satisfies

$$\bar{f}'' + \frac{\bar{f}'}{\rho} - \frac{\bar{f}}{\rho^2} = \bar{f}^3 - \bar{f}, \tag{132}$$

$$\bar{f}(0) = 0, \tag{133}$$

$$\bar{f} \to 1 \quad \text{as } \rho \to \infty, \tag{134}$$

where $' \equiv d/d\rho$. Hence $\bar{f}$ is simply the solution corresponding to a two-dimensional rectilinear isotropic vortex, whose existence and uniqueness have been shown in [13]. From (130) and (126) we see the matching condition on the outer solution $\boldsymbol{J}$ as a vortex is approached is therefore

$$\boldsymbol{J} \sim \frac{a(0)}{r}\boldsymbol{e}_\theta \text{ as } r \to 0.$$

We may combine this matching condition succinctly with the London equation (127) by writing

$$\boldsymbol{e}_z \cdot \operatorname{curl}\left(\frac{\boldsymbol{J}}{a(\boldsymbol{x})}\right) + h_3(x, y, 0) = 2\pi\delta(\boldsymbol{x}), \tag{135}$$

where $\boldsymbol{x} = (x, y)$ and $\delta(\boldsymbol{x})$ is the two-dimensional Dirac δ-function. Thus we have reduced the nonlinear but regular Ginzburg–Landau equation to a linear but singular London equation. The great advantage of the linearity of (135) is that if we have many vortices located at the positions $\boldsymbol{x}_n$, $n = 0, \ldots, N$, we may simply add up their contributions to the electric current and magnetic field by the principle of superposition to give

$$\boldsymbol{e}_z \cdot \operatorname{curl}\left(\frac{\boldsymbol{J}}{a(\boldsymbol{x})}\right) + h_3(x, y, 0) = 2\pi\sum_{n=0}^{N}\delta(\boldsymbol{x} - \boldsymbol{x}_n). \tag{136}$$

All that remains is to determine a law of motion for each vortex, so that we can determine the evolution of the positions $\boldsymbol{x}_n$. To do this we need to proceed further down the asymptotic expansions. The analysis parallels exactly that in [11, 12, 10].

Asymptotically expanding the solution of (136), (91) in inner coordinates as $r \to 0$ as in [10] we find

$$\boldsymbol{Q} = -\frac{1}{a}\boldsymbol{J} \sim -\frac{1}{\Xi R}\boldsymbol{e}_\theta - \frac{|\log \Xi|}{2(ag)} \begin{pmatrix} (ag)_y \\ -(ag)_x \end{pmatrix} + \cdots, \tag{137}$$

where the second term is evaluated at $r = 0$ (and is therefore constant). Thus the expansion in the inner region must proceed as

$$f = f^{(0)} + \Xi |\log \Xi| f^{(1)} + \cdots, \tag{138}$$

$$\boldsymbol{Q} = \frac{1}{\Xi}\boldsymbol{Q}^{(0)} + |\log \Xi| \boldsymbol{Q}^{(1)} + \cdots, \tag{139}$$

$$\Phi = \frac{|\log \Xi|}{\Xi}\Phi^{(0)} + \cdots, \tag{140}$$

$$\boldsymbol{v} = |\log \Xi| \boldsymbol{v}^{(0)} + \cdots, \tag{141}$$

where $\boldsymbol{v}$ is the velocity of the vortex at the origin. Equating coefficients of $|\log \Xi|$ in (87), (88), (91), and (92) we find

$$\begin{aligned} \boldsymbol{v}^{(0)} \cdot \boldsymbol{e}_r \frac{df^{(0)}}{dR} + \frac{\partial^2 f^{(1)}}{\partial R^2} + \frac{1}{R}\frac{\partial f^{(1)}}{\partial R} - \frac{f^{(1)}}{R^2} + \frac{1}{R^2}\frac{\partial^2 f^{(1)}}{\partial \theta^2} \\ = 3(f^{(0)})^2 f^{(1)} - a(0) f^{(1)} - \frac{2f^{(0)}}{R}\boldsymbol{Q}^{(1)} \cdot \boldsymbol{e}_\theta, \end{aligned} \tag{142}$$

$$\begin{aligned} (f^{(0)})^2 \Phi^{(0)} - \frac{1}{R}\frac{\partial}{\partial \theta}\left(\frac{2f^{(0)} f^{(1)}}{R}\right) \\ + \frac{1}{R}\frac{\partial}{\partial R}\left(R(f^{(0)})^2 \boldsymbol{Q}^{(1)} \cdot \boldsymbol{e}_r\right) + \frac{1}{R}\frac{\partial}{\partial \theta}\left((f^{(0)})^2 \boldsymbol{Q}^{(1)} \cdot \boldsymbol{e}_\theta\right) = 0, \end{aligned} \tag{143}$$

$$\frac{1}{R}\frac{\partial}{\partial R}\left(R\boldsymbol{Q}^{(1)} \cdot \boldsymbol{e}_\theta\right) - \frac{1}{R}\frac{\partial}{\partial \theta}\left(\boldsymbol{Q}^{(1)} \cdot \boldsymbol{e}_r\right) = 0, \tag{144}$$

$$\begin{aligned} -\frac{1}{R}\frac{\partial}{\partial \theta}\left(\frac{2f^{(0)} f^{(1)}}{R}\right) + \frac{1}{R}\frac{\partial}{\partial R}\left(R(f^{(0)})^2 \boldsymbol{Q}^{(0)} \cdot \boldsymbol{e}_r\right) + \frac{1}{R}\frac{\partial}{\partial \theta}\left(R(f^{(0)})^2 \boldsymbol{Q}^{(0)} \cdot \boldsymbol{e}_\theta\right) \\ + \sigma\left(\frac{1}{R}\frac{\partial}{\partial R}\left(R\frac{\partial \Theta^{(0)}}{\partial R}\right) + \frac{1}{R^2}\frac{\partial^2 \Theta^{(0)}}{\partial \theta^2}\right) = 0. \end{aligned} \tag{145}$$

From (144) and (145) we see that

$$(f^{(0)})^2 \Phi^{(0)} = \sigma\left(\frac{1}{R}\frac{\partial}{\partial R}\left(R\frac{\partial \Theta^{(0)}}{\partial R}\right) + \frac{1}{R^2}\frac{\partial^2 \Theta^{(0)}}{\partial \theta^2}\right). \tag{146}$$

As $R \to 0$ we have the boundary condition [11]

$$\Phi^{(0)} \sim -\frac{1}{R}\boldsymbol{v}^{(0)} \cdot \boldsymbol{e}_\theta. \tag{147}$$

From (143) we may write $\boldsymbol{Q}^{(1)}$ in terms of a scalar potential ψ as

$$\boldsymbol{Q}^{(1)} \cdot \boldsymbol{e}_\theta = \frac{1}{R}\frac{\partial \psi}{\partial \theta}, \tag{148}$$

$$\boldsymbol{Q}^{(1)} \cdot \boldsymbol{e}_r = \frac{\partial \psi}{\partial R}. \tag{149}$$

To match with the outer behavior (137) we require

$$\psi \sim \text{constant} \times R\sin(\theta - \alpha)$$

for some constant angle α. Thus we make the ansatz [12]

$$\psi = a(0)^{-1/2}\phi(\rho)\sin(\theta - \alpha), \tag{150}$$

$$\Phi^{(0)} = Ua(0)^{1/2}\eta(\rho)\sin(\theta - \alpha), \tag{151}$$

$$f^{(1)} = F(\rho)\cos(\theta - \alpha), \tag{152}$$

$$\boldsymbol{v}^{(0)} = U(\cos(\theta - \alpha)\boldsymbol{e}_r - \sin(\theta - \alpha)\boldsymbol{e}_\theta) = U(\cos\alpha\boldsymbol{e}_x + \sin\alpha\boldsymbol{e}_y). \tag{153}$$

The system (142), (145), (146) then reduces to

$$\frac{1}{\rho}(\rho F')' - 3\bar{f}^2F + F = -U\bar{f}' - \frac{2\bar{f}\phi}{\rho^2} + \frac{2F}{\rho^2}, \tag{154}$$

$$\frac{1}{\rho}\left(\rho\bar{f}^2\phi'\right)' + \sigma U\left(\eta' + \frac{\eta}{\rho}\right)' - \frac{\bar{f}^2\phi}{\rho^2} + \frac{2\bar{f}F}{\rho^2} = 0, \tag{155}$$

$$\eta'' + \frac{1}{\rho}\eta' - \frac{1}{\rho^2}\eta - \frac{{f^{(0)}}^2\eta}{\sigma} = 0, \tag{156}$$

where $' \equiv d/d\rho$. Equation (147) gives the boundary condition

$$\eta \sim \frac{1}{\rho} \qquad \text{as } \rho \to 0 \tag{157}$$

on (156). Matching with the outer solution gives the second condition:

$$\rho\eta \to 0 \qquad \text{as } \rho \to \infty. \tag{158}$$

This gives a well-posed problem for η, and numerical solutions of (156) with boundary conditions (157) and (158) have been given by Peres and Rubinstein [25].

We consider (154) and (155). Noting that the derivative of the leading-order solution $\bar{f}$ satisfies the homogeneous version of (154) we see that there will be a solution if and only if a certain solvability condition is satisfied. To derive this condition we multiply (154) by $\rho\bar{f}'$, the derivative of (131) by $\rho f^{(1)}$, and subtract to obtain

$$\rho U(\bar{f}')^2 + \left(\rho F'\bar{f}' - \rho F\bar{f}''\right)' = -\frac{2\bar{f}'\bar{f}\phi}{\rho} + \frac{2\bar{f}F}{\rho^2}. \tag{159}$$

Using (155) to eliminate the final F, integrating over $(0, \infty)$ and using the asymptotic behavior of the $\bar{f}$ and $\Theta^{(0)}$ at 0 and ∞ we find

$$\lim_{\rho\to\infty}\left(\phi' + \frac{\phi}{\rho}\right) = -U\beta, \tag{160}$$

where

$$\beta = \int_0^\infty \rho(\bar{f}')^2 d\rho + \int_0^\infty \bar{f}^2\eta d\rho. \tag{161}$$

Hence

$$\lim_{\rho\to\infty}\boldsymbol{Q}^{(1)} = \frac{U\beta}{2}\left(\sin\alpha\boldsymbol{e}_x - \cos\alpha\boldsymbol{e}_y\right). \tag{162}$$

Matching with the outer solution (137) using (153) gives

$$\boldsymbol{v}^{(0)} = -\frac{1}{\beta}\nabla \log(ag), \tag{163}$$

so that to leading order

$$\boldsymbol{v} = -\frac{|\log \Xi|}{\beta}\nabla \log(ag). \tag{164}$$

We see that vortices are attracted to the minima of ag. If we proceed to higher orders we find that the first two terms in the vortex velocity law give

$$\boldsymbol{v} = -\frac{|\log \Xi|}{\beta}\nabla \log(ag) + \frac{2}{\beta a}\boldsymbol{J} \wedge \boldsymbol{e}_z, \tag{165}$$

where $\boldsymbol{J}$ is the regular part of the current density. The second term here becomes dominant either when ag is constant, so that the first term vanishes, or when the background current density $\boldsymbol{J}$ is large, which will be the case when we allow the vortex separation ν/L to tend to zero to arrive at vortex-density models in section 4.[1]

Summary. The leading-order London limit of the thin-film Ginzburg–Landau model is (dropping the superscripts for clarity)

$$\operatorname{curl} \boldsymbol{h} = \boldsymbol{J}_{ext}, \tag{166}$$

$$\operatorname{div} \boldsymbol{h} = 0, \tag{167}$$

with

$$[\boldsymbol{e}_z \wedge \boldsymbol{h}] = \frac{g}{\Lambda}\boldsymbol{J} \quad \text{for} \quad (x, y) \in \Omega, \tag{168}$$

$$[\boldsymbol{e}_z \cdot \boldsymbol{h}] = 0, \tag{169}$$

$$\boldsymbol{h}^{(0)} \to \boldsymbol{h}_{ext} \quad \text{as} \quad |\boldsymbol{r}| \to \infty, \tag{170}$$

for the magnetic field outside the superconductor, and

$$\boldsymbol{e}_z \cdot \operatorname{curl}\left(\frac{\boldsymbol{J}}{a}\right) + h_3(x, y, 0) = 2\pi \sum_n \delta(\boldsymbol{x} - \boldsymbol{x}_n), \tag{171}$$

$$\operatorname{div}(g\boldsymbol{J}) = 0, \tag{172}$$

with

$$\boldsymbol{J} \cdot \boldsymbol{\nu} = 0 \quad \text{on} \quad \partial D, \tag{173}$$

for the electric current inside the superconductor, with the vortices moving according to the law

$$\dot{\boldsymbol{x}}_n = -\frac{|\log \Xi|}{\beta}\nabla \log(ag) + \frac{2}{\beta a}\boldsymbol{J} \wedge \boldsymbol{e}_z. \tag{174}$$

[1] Strictly speaking the law of motion needs to be rederived when the terms in (165) switch order so that the second one dominates, but the analysis proceeds in exactly the same way and (165) still holds.

The key equation here is (171); as we have already said, all the other equations are simply the thin-film versions of Maxwell's equations. We note that there are two source terms in (171) for the current. The first is the sum of δ-functions due to the vortices, as expected. The second is due to the applied magnetic field, which acts as a negative distributed vorticity. This term is due to the Meissner effect, by which a superconductor attempts to exclude a magnetic field from its interior. The current generated by this term in (171) is an attempt to shield the superconductor from the applied field and will result in a lower magnetic field inside the superconductor than the applied field $\boldsymbol{h}_{ext}$.

Note that if we instead consider vortices with negative winding number (corresponding to a magnetic field in the $-\boldsymbol{e}_z$-direction rather than the $\boldsymbol{e}_z$-direction), then (171) and (174) are modified to

$$\boldsymbol{e}_z \cdot \operatorname{curl}\left(\frac{\boldsymbol{J}}{a}\right) + h_3(x,y,0) = -2\pi \sum_n \delta(\boldsymbol{x} - \boldsymbol{x}_n), \tag{175}$$

$$\dot{\boldsymbol{x}}_n = -\frac{|\log \Xi|}{\beta} \nabla \log(ag) - \frac{2}{\beta a} \boldsymbol{J} \wedge \boldsymbol{e}_z. \tag{176}$$

Note also that by writing the equations in terms of the total current in the film $g\boldsymbol{J}$ (rather than the average current $\boldsymbol{J}$) the solution depends only on the product ag. This shows how variations in film thickness play exactly the same role as variations in the equilibrium density of superconducting electrons.

These equations were derived in [10, 21] for the case $a = 1$ and in [16] for the case of a strip geometry $D = (x, y) : -l < x < l$ in the absence of vortices, with $a = g = 1$. In the strip case (166)–(170) may be replaced by

$$h_3(x,0) - h_{app,3}(x,0) = \frac{1}{\pi} \int_{-l}^{l} \frac{g J_2(\bar{x})}{2\Lambda(\bar{x} - x)}\, d\bar{x}, \tag{177}$$

as in section 2.2.1.

The increased Ginzburg–Landau parameter. Note that from (94) the lengthscale for the decay of the magnetic field away from a vortex is Λ, which is the nondimensional version of Pearl's effective screening length for thin films [24]. Thus the London limit we have considered is valid so long as $\Xi \ll \Lambda$. Thus in a thin film the relevant parameter in determining the type of superconductor is not the Ginzburg–Landau parameter κ but the increased Ginzburg–Landau parameter

$$\kappa_{\text{eff}} = \frac{\Lambda}{\Xi} = \frac{\lambda_{\text{eff}}}{\xi} = \frac{\lambda^2}{\xi d} = \frac{\lambda \kappa}{d}.$$

This explains why thin films of even type-I superconducting material develop vortex solutions similar to bulk type-II superconductors once the thickness becomes smaller than the penetration depth, as was first observed by Tinkham [30].

3.2. The bulk London model. Let us now consider the opposite sequence of limits, namely we first let $\Xi \to 0$ keeping λ and ϵ fixed (i.e., we let $\kappa \to \infty$), and then we let $\epsilon \to 0$, that is, we are in the parameter regime $\Xi \ll \epsilon \ll 1$.

We have seen that the vortex cores have vanishingly small radius in the limit $\Xi \to 0$ and (25) gives $f = a(\boldsymbol{x})$ except at these isolated vortex lines. Then from (27)

$$\frac{\lambda^2}{L^2} \operatorname{curl}^2 \boldsymbol{Q} + a\boldsymbol{Q} = \boldsymbol{0}, \tag{178}$$

or, taking the curl,

$$\frac{\lambda^2}{L^2}\mathrm{curl}\left(\frac{1}{a}\mathrm{curl}\boldsymbol{H}\right)+\boldsymbol{H}=\boldsymbol{0}, \tag{179}$$

away from vortices. By matching a local asymptotic analysis of an individual vortex core with this outer expansion away from vortices, as in section 3.1 and as detailed in [11], it may be shown that the presence of vortices leads to δ-function singularities on the right-hand side of (179), so that

$$\frac{\lambda^2}{L^2}\mathrm{curl}\left(\frac{1}{a}\mathrm{curl}\boldsymbol{H}\right)+\boldsymbol{H}=2\pi\sum_n \boldsymbol{\delta}_{\Gamma_n}, \tag{180}$$

where

$$\boldsymbol{\delta}_\Gamma(\boldsymbol{x})=\int_\Gamma \delta(x-\bar{x})\delta(y-\bar{y})\delta(z-\bar{z})\,\boldsymbol{d\bar{x}}. \tag{181}$$

Proceeding to first order in this asymptotic matching and applying a solvability condition gives the law of motion of superconducting vortices as [11, 12]

$$\boldsymbol{v}=\frac{|\log\Xi|}{\beta}C\boldsymbol{n}-\frac{|\log\Xi|}{\beta}\nabla\log a+\frac{2}{a\beta}\boldsymbol{J}\wedge\boldsymbol{t}, \tag{182}$$

where $\boldsymbol{J}$ is the background current, C is the curvature of the vortex line, and β is given by (161) as before.

It has been shown in [10] that (182), (180) imply that vortices must meet the boundary $\partial\Omega$ normally, since if this is not so an infinite current density is produced. (This can also be thought of as an infinite curvature of the vortex and its image, which is clearly incompatible with (182).)

Equations (180), (182) must be coupled with Maxwell's equations (6)–(10) outside the film, along with continuity conditions on $\boldsymbol{H}$ across $\partial\Omega$.

3.3. Thin film of the London equations. As in section 2, we assume that the vortex separation is of the same order as the lateral dimension of the film as $\epsilon\to 0$, that is, the film contains a finite number of vortices in the limit. We write (180) as

$$\boldsymbol{J}=\epsilon\Lambda\,\mathrm{curl}\,\boldsymbol{H}, \tag{183}$$

$$\mathrm{curl}\left(\frac{\boldsymbol{J}}{a}\right)+\boldsymbol{H}=2\pi\sum_n \boldsymbol{\delta}_{\Gamma_n}, \tag{184}$$

where our scaling of $\boldsymbol{J}$, consistent with section 2, is such that terms in (184) are balanced. Of course, in the canonical case both (183) and (184) will be balanced, and it is only when we are considering limiting cases that the scaling of $\boldsymbol{J}$ is important. As in section 2 we will find that the canonical scaling is to have $\Lambda=\lambda^2/dL$ of order one. Since the vortices must meet the upper and lower surfaces of the film normally, they must lie in the z-direction to leading order. Thus the equations to be satisfied are

$$\mathrm{curl}\,\boldsymbol{j}+\boldsymbol{H}=2\pi\sum_n \delta(\boldsymbol{x}-\boldsymbol{x}_n)\boldsymbol{e}_z+\cdots \qquad \text{in } \Omega, \tag{185}$$

$$a\boldsymbol{j}=\Lambda\epsilon\,\mathrm{curl}\,\boldsymbol{H} \qquad \text{in } \Omega, \tag{186}$$

$$\operatorname{div} \boldsymbol{H} = 0 \qquad \text{in } \Omega, \tag{187}$$

$$\operatorname{div} (a\boldsymbol{j}) = 0 \qquad \text{in } \Omega, \tag{188}$$

$$\operatorname{curl} \boldsymbol{h} = \boldsymbol{J}_{ext} \qquad \text{outside } \Omega, \tag{189}$$

$$\operatorname{div} \boldsymbol{h} = 0 \qquad \text{outside } \Omega, \tag{190}$$

with

$$\boldsymbol{h} = \boldsymbol{H} \quad \text{on } \partial\Omega, \tag{191}$$

$$\boldsymbol{j} \cdot \boldsymbol{n} = 0 \quad \text{on } \partial\Omega, \tag{192}$$

and

$$\boldsymbol{h} \to \boldsymbol{h}_{ext} \quad \text{as} \quad |\boldsymbol{r}| \to \infty, \tag{193}$$

where $\boldsymbol{n}$ is the normal to $\partial\Omega$ and we have written $\boldsymbol{J} = a\boldsymbol{j}$.

As usual, we start by considering the problem inside the film. Rescaling $z = \epsilon Z$, equation (185) becomes

$$-\frac{1}{\epsilon}\frac{\partial j_2}{\partial Z} + \frac{\partial j_3}{\partial y} + H_1 = \cdots, \tag{194}$$

$$\frac{1}{\epsilon}\frac{\partial j_1}{\partial Z} - \frac{\partial j_3}{\partial x} + H_2 = \cdots, \tag{195}$$

$$\frac{\partial j_2}{\partial x} - \frac{\partial j_1}{\partial y} + H_3 = 2\pi \sum_n \delta(\hat{\boldsymbol{x}} - \hat{\boldsymbol{x}}_n) + \cdots. \tag{196}$$

We expand all quantities asymptotically in powers of ϵ as

$$\boldsymbol{h} \sim \boldsymbol{h}^{(0)} + \epsilon \boldsymbol{h}^{(1)} + \epsilon^2 \boldsymbol{h}^{(2)} + \cdots, \tag{197}$$

etc. Substituting these expansions into (194)–(196) and equating powers of ϵ gives, at leading order,

$$\frac{\partial j_1^{(0)}}{\partial Z} = \frac{\partial j_2^{(0)}}{\partial Z} = 0, \tag{198}$$

$$\frac{\partial j_2^{(0)}}{\partial x} - \frac{\partial j_1^{(0)}}{\partial y} + H_3^{(0)} = 2\pi \sum_n \delta(\hat{\boldsymbol{x}} - \hat{\boldsymbol{x}}_n). \tag{199}$$

Thus, as before, $j_1^{(0)}$ and $j_2^{(0)}$ are constant in Z. The thin-film analysis of (186)–(193) then proceeds exactly as in section 2.2 to give [10]

$$j_3^{(0)} = 0, \tag{200}$$

$$\left[\boldsymbol{e}_z \wedge \boldsymbol{h}^{(0)}\right] = \frac{ag}{\Lambda}\boldsymbol{j}^{(0)} \quad \text{for} \quad (x, y) \in D, \tag{201}$$

$$\left[\boldsymbol{e}_z \cdot \boldsymbol{h}^{(0)}\right] = 0 \quad \text{for} \quad (x, y) \in D. \tag{202}$$

Similarly the thin-film analysis of (188) gives

$$\operatorname{div} (ag\,\boldsymbol{j}^{(0)}) = 0. \tag{203}$$

Finally we need to determine the law of motion for the vortices. An asymptotic analysis in [10] for constant a matching this outer model with the inner model in the vicinity of the vortex cores gives the law of motion

$$\boldsymbol{v} = \frac{2}{\beta}\boldsymbol{J} \wedge \boldsymbol{e}_z - \frac{|\log \Xi|}{\beta}\nabla \log g. \tag{204}$$

If a variable a is included in this analysis, the result is

$$\boldsymbol{v} = \frac{2}{a\beta}\boldsymbol{J} \wedge \boldsymbol{e}_z - \frac{|\log \Xi|}{\beta}\nabla \log(ag). \tag{205}$$

It is shown in [10] that the motion due to variations in film thickness is consistent with the motion of the vortex under (182) with the curvature it must have if it forms an arc of a circle meeting the upper and lower surfaces of the film normally.

Thus we find that the thin-film version of the London model corresponds exactly to the London limit of the thin-film Ginzburg–Landau model (166)–(174).

3.4. Limiting cases. In the limit $\Lambda \to \infty$ the magnetic field generated by the current in the film gives a negligible contribution to (171), and $h_3(x, y, 0)$ is simply replaced by $h_{app,3}(x, y, 0)$, the third component of the total applied magnetic field. The problem for the electric current is then decoupled from the problem for the magnetic field, which as usual reduces to the standard problem in electromagnetism of calculating the magnetic field due to a current sheet once the electric current has been found.

For Λ small we must rescale the electric current with Λ. Then the current term in (171) is negligible, and this equation reads

$$h_3(x, y, 0) = 2\pi \sum_n \delta(\boldsymbol{x} - \boldsymbol{x}_n) \tag{206}$$

to leading order. Hence the normal component of the magnetic field is zero on $z = 0$, except at the positions of the vortices. Near each vortex an inner problem must be solved, in which length is rescaled with Λ, so that the first term in (171) is relevant again. Thus, effectively, the relevant horizontal lengthscale is no longer L but the effective penetration depth λ_{eff}, and the vortices are completely isolated. This last limit is much more interesting when the vortices are more closely separated, so that the sum of δ-functions can be averaged into a vortex density or vorticity.

4. Vortex-density models. So far we have been considering the limit $\epsilon \to 0$ with the vortex separation of the same order as L, so that there are a finite number of vortices in the film in the limit. In this section we will consider the limit in which $\nu/L \to 0$ where ν is the vortex separation, so that the individual vortices are replaced by a vortex density.

As before we now have two limiting processes ($\epsilon \to 0$ and $\nu/L \to 0$), and we can choose to perform them in either order. We will consider first the vortex-density limit of the thin-film London model, that is, we first let $\epsilon \to 0$ and then let $\nu/L \to 0$, so that $\epsilon \ll \nu/L \ll 1$. In section 4.3 we will consider the alternative regime $\nu/L \ll \epsilon \ll 1$.

4.1. Averaging the thin-film London model. We consider the limit in which the vortex separation $\nu/L \to 0$. In this case it is clear from (171) that both $\boldsymbol{h}$ and $\boldsymbol{J}$ need to be rescaled with L^2/ν^2.

We formally define the vortex density as

$$\omega(\boldsymbol{x}) = \lim_{\eta\to 0} \frac{2\pi\nu^2}{\eta^2 L^2} \int_{|x|<\eta/2,|y|<\eta/2} \sum_i \delta(\boldsymbol{x}' - \boldsymbol{x}_n)\, d\boldsymbol{x}', \tag{207}$$

where the integration is over a square of dimension η, and $\nu/L \ll \eta \ll 1$ as $\nu/L \to 0$, so that the size of the square is tending to zero, but each square contains many vortices. The prefactor ν^2/L^2 ensures that the limit is order one. Now locally averaging (166)–(173) by formally integrating over the same square of side η, we find the electric current inside the superconductor satisfies

$$\boldsymbol{e}_z \cdot \operatorname{curl}\left(\frac{\boldsymbol{J}}{a}\right) + h_3(x,y,0) = \omega, \tag{208}$$

$$\operatorname{div}(g\boldsymbol{J}) = 0, \tag{209}$$

with

$$\boldsymbol{J}\cdot\boldsymbol{\nu} = 0 \quad \text{on} \quad \partial D, \tag{210}$$

while the problem for the magnetic field outside the superconductor,

$$\operatorname{curl}\boldsymbol{h} = \boldsymbol{J}_{ext}, \tag{211}$$

$$\operatorname{div}\boldsymbol{h} = 0, \tag{212}$$

with

$$[\boldsymbol{e}_z \wedge \boldsymbol{h}] = \frac{g}{\Lambda}\boldsymbol{J} \quad \text{for} \quad (x,y) \in D, \tag{213}$$

$$[\boldsymbol{e}_z \cdot \boldsymbol{h}] = 0, \tag{214}$$

$$\boldsymbol{h} \to \boldsymbol{h}_{ext} \quad \text{as} \quad |\boldsymbol{r}| \to \infty, \tag{215}$$

is unchanged. In the strip geometry, with a and g constant, equation (208) was derived by Larkin and Ovchinnikov [22]. As usual, in this geometry (211)–(215) can be replaced by (124). However, we now have another dependent variable, ω, so we need another equation. In passing from isolated vortices to a vortex density we have lost the property of vortex conservation, which is automatic when we are tracking individual vortices. Thus the equation we need to add is a conservation law for the vortices. This law takes the usual form, namely

$$\frac{\partial\omega}{\partial t} + \operatorname{div}(\omega\boldsymbol{v}) = 0, \tag{216}$$

where $\boldsymbol{v}$ is the velocity of the vortices, which is given by (174). Remembering that we have rescaled the electric current with L^2/ν^2, we also now rescale time with $2L^2/\beta\nu^2$ to give

$$\boldsymbol{v} = \frac{1}{a}\boldsymbol{J}\wedge\boldsymbol{e}_z - \frac{\nu^2|\log\Xi|}{2L^2}\nabla\log(ag). \tag{217}$$

Now the pinning effect of variations in ag has been weakened due to the increased current density in the film; if the lengthscale for variations in ag is order L, then the first term is likely to dominate. In section 5 we will consider the case when the

lengthscale ε for variations in a is also small, so that pinning becomes important again.

If we perform the same analysis in a region of vortices with negative winding number, we find $\omega < 0$ and (217) becomes

$$\boldsymbol{v} = -\frac{1}{a}\boldsymbol{J} \wedge \boldsymbol{e}_z - \frac{\nu^2 |\log \Xi|}{2L^2} \nabla \log(ag). \tag{218}$$

We can combine these two results in the single vortex velocity law

$$\boldsymbol{v} = \frac{\operatorname{sign}(\omega)}{a}\boldsymbol{J} \wedge \boldsymbol{e}_z - \frac{\nu^2 |\log \Xi|}{2L^2} \nabla \log(ag). \tag{219}$$

Equation (216) introduces one real characteristic into the model, so that we need to give an extra boundary condition (on ω) whenever this characteristic points into the domain, i.e., whenever $\boldsymbol{v} \cdot \boldsymbol{\nu} < 0$. This condition will typically relate either the magnitude of the vorticity or the flux of vorticity through the boundary to the local current density. In the two-dimensional case it is found that vortices will not be nucleated at the boundary until the current density reaches a critical value. In the thin-film case the details of vortex nucleation will depend on the local shape of the film near the boundary ∂D, and an inner boundary layer problem needs to be solved on a lateral lengthscale of order d, the film thickness. However, it is natural to assume still that vortices will not be nucleated until the current density at the boundary reaches a critical value, J_{nucl} say.

4.2. The bulk vortex-density model. Let us now consider the parameter regime $\nu/L \ll \epsilon \ll 1$; that is, we first consider the limit $\nu/L \to 0$ to obtain the bulk vortex-density model and then consider the thin-film limit of it by letting $\epsilon \to 0$.

In three dimensions a similar formal averaging gives the vortex-density model as [6]

$$\frac{\lambda^2}{L^2} \operatorname{curl} \left(\frac{1}{a} \operatorname{curl} \boldsymbol{H} \right) + \boldsymbol{H} = \boldsymbol{\omega} \quad \text{in } \Omega, \tag{220}$$

$$\operatorname{div} \boldsymbol{H} = 0, \tag{221}$$

with Maxwell's equations (6)–(10) as usual outside Ω, with continuity of $\boldsymbol{H}$ across $\partial\Omega$ and

$$\boldsymbol{J} \cdot \boldsymbol{n} = 0 \quad \text{on } \partial\Omega, \tag{222}$$

where $\boldsymbol{\omega}$ is now a vector vortex density or vorticity given by

$$\boldsymbol{\omega}(\boldsymbol{x}) = \lim_{\eta \to 0} \frac{2\pi\nu^2}{\eta^2 L^2} \int_{|x|<\eta/2, |y|<\eta/2} \sum_i \boldsymbol{\delta}_{\Gamma_n}(\boldsymbol{x}')\, d\boldsymbol{x}'.$$

The law of conservation of vortices in three dimensions is [6, 5]

$$\frac{\partial \boldsymbol{\omega}}{\partial t} + \operatorname{curl} (\boldsymbol{\omega} \wedge \boldsymbol{v}) = 0, \tag{223}$$

where the vortex velocity $\boldsymbol{v}$ is given by

$$\boldsymbol{v} = \frac{1}{a}\boldsymbol{J} \wedge \hat{\boldsymbol{\omega}} + \frac{\nu^2 |\log \Xi|}{2L^2} \left(-\nabla \log a + \operatorname{curl} \hat{\boldsymbol{\omega}} \wedge \hat{\boldsymbol{\omega}} \right), \tag{224}$$

and $\hat{\boldsymbol{\omega}}$ is the unit vector in the direction of $\boldsymbol{\omega}$. The term $\operatorname{curl} \hat{\boldsymbol{\omega}} \wedge \hat{\boldsymbol{\omega}}$ here is just the curvature term rewritten. If the limits $\Xi \to 0$ and $\nu/L \to 0$ are such that $\nu^2 |\log \Xi|/L^2 \to 0$, then the last term is of lower order and can be neglected. In the general three-dimensional situation neglecting the self-induced curvature term leads to the possibility of a short-wavelength large–growth-rate instability on vortex lines if $\boldsymbol{J} \cdot \boldsymbol{\omega} \neq 0$, as shown in [26] and discussed in [6]. However, in the thin-film limit it is safe to neglect the self-induced term since the current will lie in the (x, y)-plane while the vorticity will be normal to it.

As in the thin-film vortex-density model, equation (223) has introduced one real characteristic, and we need to give a boundary condition on $\boldsymbol{\omega}$ whenever $\boldsymbol{v} \cdot \boldsymbol{n} < 0$. Now, an analysis in [15] indicates that on such an inflow boundary either $\boldsymbol{\omega} \cdot \boldsymbol{n} = 0$ (which is the case in the thin-film model and also in the two-dimensional model with axial symmetry) or $\boldsymbol{\omega} \wedge \boldsymbol{n} = \boldsymbol{0}$. In the first case the flux of vorticity through the boundary will be related to the local current density, while in the second no extra condition needs to be given. In a general three-dimensional situation the position of the switch between these two types of behavior is unknown, and a boundary layer calculation may be necessary in its vicinity. However, in our thin film scenario we are fortunate that the top and bottom of the film will have vortices passing through them and be such that $\boldsymbol{\omega} \wedge \boldsymbol{n} = \boldsymbol{0}$ if they are inflow, while we will see that the sides of the film will have $\boldsymbol{\omega} \cdot \boldsymbol{n} = 0$ to leading order, and therefore these are the boundaries through which new vortices will pass.

4.3. Thin film of the vortex-density model. Let us now consider the thin-film limit $\epsilon \to 0$ of the bulk vortex-density model (220)–(224), corresponding to the parameter regime $\nu/L \ll \epsilon \ll 1$. We begin as usual by writing (220) as

$$\operatorname{curl}\left(\frac{\boldsymbol{J}}{a}\right) + \boldsymbol{H} = \boldsymbol{\omega} \quad \text{in } \Omega, \tag{225}$$

$$\boldsymbol{J} = \Lambda\epsilon \operatorname{curl} \boldsymbol{H} \quad \text{in } \Omega. \tag{226}$$

Expanding all quantities in powers of ϵ the analysis proceeds exactly as in section 3.3 giving

$$\operatorname{curl} \boldsymbol{h} = \boldsymbol{J}_{ext}, \tag{227}$$

$$\operatorname{div} \boldsymbol{h} = 0, \tag{228}$$

with

$$[\boldsymbol{e}_z \wedge \boldsymbol{h}] = \frac{g}{\Lambda}\boldsymbol{J} \quad \text{for} \quad (x, y) \in D, \tag{229}$$

$$[\boldsymbol{e}_z \cdot \boldsymbol{h}] = 0, \tag{230}$$

$$\boldsymbol{h} \to \boldsymbol{h}_{ext} \quad \text{as} \quad |\boldsymbol{r}| \to \infty, \tag{231}$$

for the magnetic field outside the superconductor, and

$$\boldsymbol{e}_z \cdot \operatorname{curl}\left(\frac{\boldsymbol{J}}{a}\right) + h_3(x, y, 0) = \omega_3, \tag{232}$$

$$\operatorname{div}(g\boldsymbol{J}) = 0, \tag{233}$$

with

$$\boldsymbol{J} \cdot \boldsymbol{\nu} = 0 \quad \text{on} \quad \partial D, \tag{234}$$

for the electric current inside the superconductor. The only additional work is to check that the three-dimensional vector conservation law (223) reduces to the two-dimensional scalar conservation law (216) in the thin-film limit. To do that we need to show that $\boldsymbol{\omega}$ lies in the z-direction to leading order. Since one of the top or bottom sides of the film is an inflow boundary, and on that inflow boundary $\boldsymbol{\omega} \wedge \boldsymbol{n} = \boldsymbol{0}$, to leading order we must have $\boldsymbol{\omega} \wedge \boldsymbol{e}_z = 0$ there. To be sure that $\omega_1 = \omega_2 = 0$ everywhere we must check that they cannot vary rapidly in z, i.e., they are independent of Z to leading order. Writing (223) in component form in the film gives

$$\frac{\partial \omega_1}{\partial t} + \frac{\partial}{\partial y}(\omega_1 v_2 - \omega_2 v_1) - \frac{1}{\epsilon}\frac{\partial}{\partial Z}(\omega_3 v_1 - \omega_1 v_3) = 0, \tag{235}$$

$$\frac{\partial \omega_2}{\partial t} - \frac{\partial}{\partial x}(\omega_1 v_2 - \omega_2 v_1) + \frac{1}{\epsilon}\frac{\partial}{\partial Z}(\omega_2 v_3 - \omega_3 v_2) = 0, \tag{236}$$

$$\frac{\partial \omega_3}{\partial t} + \frac{\partial}{\partial x}(\omega_3 v_1 - \omega_1 v_3) - \frac{\partial}{\partial y}(\omega_2 v_3 - \omega_3 v_2) = 0, \tag{237}$$

where

$$a v_1 = J_2 \hat{\omega}_3, \tag{238}$$

$$a v_2 = -J_1 \hat{\omega}_3, \tag{239}$$

$$a v_3 = J_1 \hat{\omega}_2 - J_2 \hat{\omega}_1. \tag{240}$$

Thus at leading order in ϵ

$$\omega_3 v_1 - \omega_1 v_3 = \text{ independent of } Z, \tag{241}$$

$$\omega_2 v_3 - \omega_3 v_2 = \text{ independent of } Z. \tag{242}$$

Now (225) implies div $\boldsymbol{\omega} = 0$, which in the film gives ω_3 independent of Z to leading order. Then eliminating $\boldsymbol{v}$ from (241)–(242) gives two equations for ω_1 and ω_2 in terms of ω_3, so that ω_1 and ω_2, and therefore $\boldsymbol{v}$, are also independent of Z to leading order.

Finally (237) then gives

$$\frac{\partial \omega_3}{\partial t} + \frac{\partial}{\partial x}(\omega_3 v_1) + \frac{\partial}{\partial y}(\omega_3 v_2) = 0, \tag{243}$$

as required.

4.4. Limiting cases. In the limit in which $\Lambda \to \infty$, the problem for the electric current becomes

$$\boldsymbol{e}_z \cdot \text{curl}\left(\frac{\boldsymbol{J}}{a}\right) + h_{app,3}(x, y, 0) = \omega, \tag{244}$$

$$\frac{\partial \omega}{\partial t} + \text{div}\,(\omega \boldsymbol{v}) = 0, \tag{245}$$

$$\boldsymbol{v} = \frac{\text{sign}(\omega)}{a} \boldsymbol{J} \wedge \boldsymbol{e}_z, \tag{246}$$

$$\text{div}\,(g\boldsymbol{J}) = 0, \tag{247}$$

with

$$\boldsymbol{J} \cdot \boldsymbol{\nu} = 0 \quad \text{on} \quad \partial\Omega. \tag{248}$$

Once the current has been found, the magnetic field may be found from (211)–(215) in the usual way.

In the limit in which $\Lambda \to 0$, we must scale the electric current with Λ to give

$$\operatorname{curl} \boldsymbol{h} = \boldsymbol{J}_{ext}, \tag{249}$$

$$\operatorname{div} \boldsymbol{h} = 0, \tag{250}$$

with

$$[\boldsymbol{e}_z \wedge \boldsymbol{h}] = g\boldsymbol{J} \quad \text{for} \quad (x, y) \in \Omega, \tag{251}$$

$$[\boldsymbol{e}_z \cdot \boldsymbol{h}] = 0, \tag{252}$$

$$\boldsymbol{h}^{(0)} \to \boldsymbol{h}_{ext} \quad \text{as} \quad |\boldsymbol{r}| \to \infty, \tag{253}$$

$$h_3(x, y, 0) = \omega. \tag{254}$$

The problems for the magnetic field and electric current again simplify, but in this case (249)–(250), (252)–(254) are first solved for the magnetic field, and then (251) is used to determine the current in the superconductor. Having found the current, (216) and (219) are then solved for the vortex density, which in turn feeds into (254). Thus the problems remain coupled in this limit.

However, in taking the limit $\nu/L \to 0$ and $\Lambda \to 0$ there is a constraint on relative magnitudes of these two parameters for the model (216), (219), (249)–(254) to be valid. As we let $\Lambda \to 0$ we are weakening the size of the mean-field current over the local perturbation to this current due to neighboring individual vortices. We have seen that the mean-field current is $O(\Lambda L^2/\nu^2)$, while influence of the current due to a single vortex on its neighbors is $O(L/\nu)$. Thus for the dynamics to be dominated by the mean field as in (219) we need $\nu/L \ll \Lambda$, i.e., $\nu \ll \lambda_{\text{eff}}$. If this is not the case, then the vortex motion will be dominated by local forces and they will form a strong lattice rather than the vortex liquid we are supposing. Note that for bulk superconductors the equivalent condition is $\nu < \lambda^2/L$; the thin film condition is much less stringent since the right-hand side is increased by the aspect ratio L/d.

In the strip geometry (216), (219), (249)–(254) reduce to an interesting singular integral equation. Suppose that $D = (x, y) : -l < x < l$ and that $h_{app,2} = 0$ as usual. In the strip geometry vortex conservation (245) gives

$$\frac{\partial \omega}{\partial t} + \frac{\partial}{\partial x}\left(\frac{J_2|\omega|}{a}\right) = 0, \tag{255}$$

while (124) and (254) give

$$\omega - h_{app,3} = \frac{1}{\pi} \fint_{-l}^{l} \frac{gJ_2(\bar{x})}{2(\bar{x} - x)}\, d\bar{x}, \quad -l < x < l. \tag{256}$$

Hence

$$\frac{\partial}{\partial t}\left(h_{app,3} + \frac{1}{\pi} \fint_{-l}^{l} \frac{gJ_2(\bar{x})}{2(\bar{x} - x)}\, d\bar{x}\right) + \frac{\partial}{\partial x}\left(\frac{J_2}{a}\left|h_{app,3} + \frac{1}{\pi} \fint_{-l}^{l} \frac{gJ_2(\bar{x})}{2(\bar{x} - x)}\, d\bar{x}\right|\right) = 0, \quad -l < x < l. \tag{257}$$

If $l = \infty$, we may invert (256) to give

$$\frac{gJ_2}{2} = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{\omega(\bar{x}) - h_{app,3}}{x - \bar{x}}\, d\bar{x}, \tag{258}$$

and hence

$$\frac{\partial \omega}{\partial t} + \frac{\partial}{\partial x}\left(\frac{2|\omega|}{ag}\left[\frac{1}{\pi}\int_{-\infty}^{\infty} \frac{\omega(\bar{x}) - h_{app,3}}{x - \bar{x}}\, d\bar{x}\right]\right) = 0. \tag{259}$$

This inversion is not so easy if l is finite, since J_2 depends on h_3 outside the domain D, which cannot be simply written in terms of ω.

5. Critical-state models. So far we have been considering the limit in which the separation of vortices $\nu/L \to 0$ but in which the lengthscale for variations in a, namely ε, remains fixed. Let us now consider the limit in which $\varepsilon/L \to 0$ also, so that the pinning potential is also homogenized. Such a limit will lead to critical state models, in which vortices do not move until the forcing current exceeds a critical value.

5.1. Critical-state limit of the vortex-density thin-film model. Suppose we now allow a in (208)–(217) to vary rapidly. If we look locally near a single vortex which is in the deepest local well and consider the effect of applying a current $\boldsymbol{J}$ to it, we see that if the current is not sufficient to cause the vortex to leave its local well, then it will move up the side of the well until the attraction from the potential balances the applied current (see Figure 3). In this case the vortex will have moved a distance of order ε. Now, if we increase the current $\boldsymbol{J}$, then at some point it will be sufficient to cause the vortex to leave its local well and it will jump into the next well. However, since the vortex was in the deepest local well the current will also be sufficient to cause the vortex to leave the neighboring well. It will continue in this way until it has moved an order-one distance, until either the local well depth has increased, or until the local current density has decreased sufficiently to catch the vortex. Hence, in the limit that $\varepsilon/L \to 0$, we arrive at a stick-slip mobility law: if the local current density is less that a critical value, J_c say (which may be a function of position), then the vortex does not move, while if the local current density is greater than J_c, then the vortex will move but at a reduced speed due to moving through the potential wells. If the distribution of well depth is nonuniform, so that there are a range of depths locally, then it is possible to see how the critical current in this stick-slip model may depend on ω. If we add more and more vortices locally, we have to make use of shallower and shallower wells, thus reducing the critical current required to start some of the vortices moving.

The critical current density in these scalings will be of order $(\nu^2/L\varepsilon)|\log \Xi|$.

We have described the interaction of a single vortex with a rapidly varying potential, which is the case $\varepsilon \ll \nu$. A similar scenario exists in the complementary situation $\nu \ll \varepsilon$ with the single vortex replaced by a "pool" of vorticity, as shown in Figure 4. Applying a forcing term now corresponds to "tilting" the potential, and again there will be a critical current at which the well can no longer hold the pool. In this case it is clear that the critical current will depend on the vortex density.

Thus, after homogenizing the pinning potential the law of motion (219) is modified to be

$$\boldsymbol{v} = \operatorname{sign}(\omega) F(|\boldsymbol{J}|; \boldsymbol{x}, \omega)\hat{\boldsymbol{J}} \wedge \boldsymbol{e}_z, \tag{260}$$

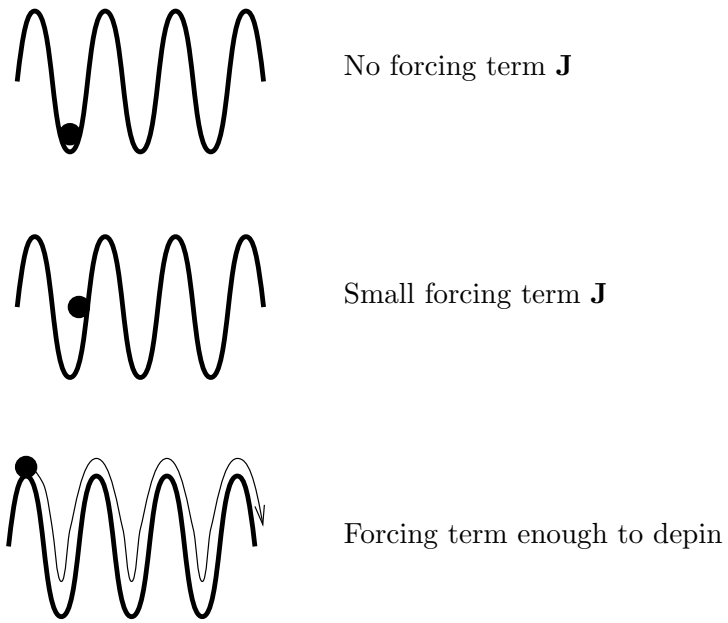


FIG. 3. *Pinning of a single vortex by a rapidly varying pinning potential.*

where $\hat{\boldsymbol{J}}$ is the unit vector in the direction of $\boldsymbol{J}$ and $F(J) = 0$ for $J < J_c(\boldsymbol{x}, \omega)$ (see Figure 5). The value of J_c and the exact form of the velocity law for J above J_c depend on the nature of the pinning potential. We do not delve into the details of this correspondence here, but note that for a single vortex moving in a simple sinusoidal potential with uniform well depth, F approaches zero as $(J - J_c)^{1/2}$. Note also that (208) becomes

$$\boldsymbol{e}_z \cdot \operatorname{curl}\left(\frac{1}{\hat{a}(\boldsymbol{x})}\boldsymbol{J}\right) + h_3(x, y, 0) = \omega, \tag{261}$$

where $\hat{a}$ is the effective equilibrium density of superconducting electrons, which is not simply the local average of a but must be determined through a multiple scales analysis.

Since the critical state corresponds to simply changing the law of motion, the thin-film limit of the bulk critical-state model leads to the same set of equations.

We note that it is quite common in the literature to use a vortex velocity law in which the velocity depends on the electric current via a power law, either with or without a critical current. In [9] a law of this type is used to derive the current-voltage characteristics of thin strips from the underlying vortex velocity law. Even more common (especially when $\Lambda = 0$) is to model the superconductor as a nonlinear conductor by assuming a nonlinear (typically power-law) relationship between the electric field

$$\boldsymbol{E} = \Lambda\frac{\partial \boldsymbol{J}}{\partial t} + \omega \boldsymbol{e}_z \wedge \boldsymbol{v} = \Lambda\frac{\partial \boldsymbol{J}}{\partial t} + |\omega| F(|\boldsymbol{J}|; \boldsymbol{x}, \omega)\hat{\boldsymbol{J}}$$

and the electric current. When $\Lambda = 0$ this corresponds to choosing $F(|\boldsymbol{J}|; \boldsymbol{x}, \omega) \propto |\boldsymbol{J}|^n/|\omega|$. Typically the power chosen is large, $n = 9$ [28] or even $n = 19$ [29]. When n is large these power laws approximate to a stick-slip law of the form (260), but with

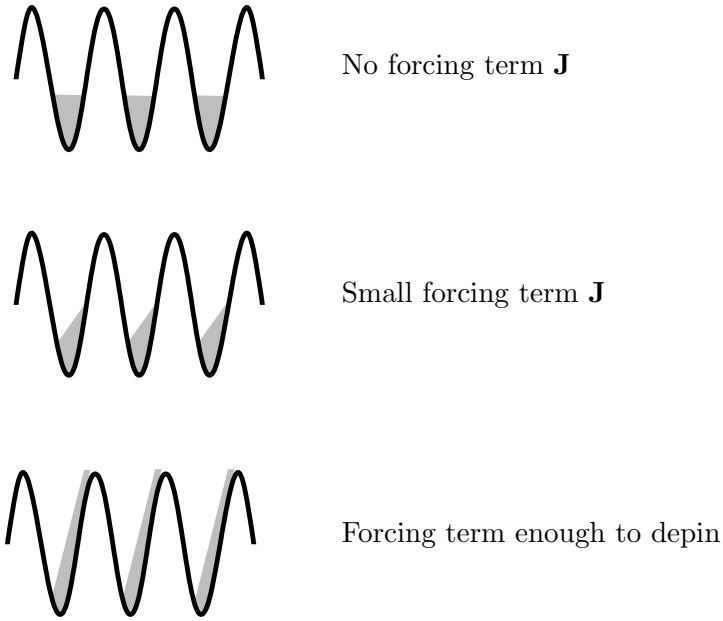


FIG. 4. *Pinning of a vortex pool by a rapidly varying pinning potential.*

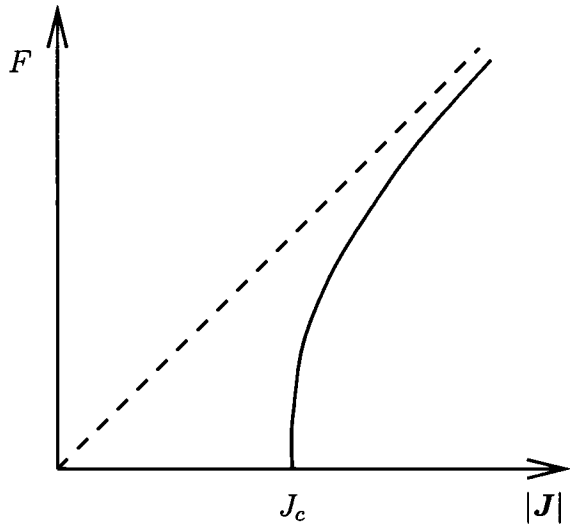


FIG. 5. *Velocity law with stick-slip pinning.*

an infinite vortex mobility once the critical current is exceeded (so that the graph in Figure 5 is vertical). Such an infinite mobility also approximates an arbitrary stick-slip law (260) when the applied magnetic field is slowly varying.

When the mobility of unpinned vortices is infinite the velocity $\boldsymbol{v}$ lies in the direction of $\text{sign}(\omega)\boldsymbol{J} \wedge \boldsymbol{e}_z$, but its magnitude is determined from the constraint that $|\boldsymbol{J}^{(0)}| \leq J_c$, so that

$$\boldsymbol{v} = \text{sign}(\omega) m \boldsymbol{J} \wedge \boldsymbol{e}_z, \tag{262}$$

$$|\boldsymbol{J}| \le J_c \text{ if } \omega \neq 0, \tag{263}$$

$$m \ge 0, \tag{264}$$

$$m\,(|\boldsymbol{J}| - J_c) = 0. \tag{265}$$

These equations are coupled with the law of vortex conservation,

$$\frac{\partial \omega}{\partial t} + \operatorname{div}(\omega \boldsymbol{v}) = 0, \tag{266}$$

and the usual equations for the magnetic field and electric current, namely

$$\operatorname{curl} \boldsymbol{h} = \boldsymbol{J}_{ext}, \tag{267}$$

$$\operatorname{div} \boldsymbol{h} = 0, \tag{268}$$

with

$$[\boldsymbol{e}_z \wedge \boldsymbol{h}] = \frac{g}{\Lambda}\boldsymbol{J} \quad \text{for} \quad (x,y) \in D, \tag{269}$$

$$[\boldsymbol{e}_z \cdot \boldsymbol{h}] = 0, \tag{270}$$

$$\boldsymbol{h} \to \boldsymbol{h}_{ext} \quad \text{as} \quad |\boldsymbol{r}| \to \infty, \tag{271}$$

and

$$\boldsymbol{e}_z \cdot \operatorname{curl}\left(\frac{\boldsymbol{J}}{\hat{a}}\right) + h_3(x,y,0) = \omega, \tag{272}$$

$$\operatorname{div}(g\boldsymbol{J}) = 0, \tag{273}$$

with

$$\boldsymbol{J} \cdot \boldsymbol{\nu} = 0 \quad \text{on} \quad \partial D. \tag{274}$$

However, so far $\omega = 0$ is a solution for all applied magnetic fields. What we are missing is the condition of nucleation of vortices at the boundary ∂D. Vortices will be nucleated once the current reaches the nucleation value J_{nucl}, but if $J_{nucl} < J_c$, they will not move until the current reaches this higher depinning value. Thus in addition to (263) we must also impose

$$|\boldsymbol{J}| \le J_c \quad \text{on } \partial D. \tag{275}$$

Now in three dimensions in regions where $\omega = 0$, taking the curl of (225) shows that $|\boldsymbol{J}|$ takes its maximum value on the boundary. If this is true for the thin-film model (272) also, then we can replace (262)–(265) with

$$\omega \boldsymbol{v} = m\boldsymbol{J} \wedge \boldsymbol{e}_z, \tag{276}$$

$$|\boldsymbol{J}| \le J_c, \qquad m \ge 0, \qquad m\,(|\boldsymbol{J}| - J_c) = 0. \tag{277}$$

Streamfunction formulation. Using (273) we may introduce a streamfunction ψ such that

$$\boldsymbol{J} = \frac{1}{g}\,(\psi_y, -\psi_x, 0)\,. \tag{278}$$

Then $\boldsymbol{v}$ is in the direction of $-\nabla\psi$, so we may set

$$\omega \boldsymbol{v} = -m\nabla\psi, \tag{279}$$

giving

$$\frac{\partial \omega}{\partial t} = \operatorname{div}(m\nabla\psi), \tag{280}$$

$$-\nabla\cdot\left(\frac{1}{\hat{a}g}\nabla\psi\right) + h_3(x,y,0) = \omega, \tag{281}$$

with the constraints

$$|\nabla\psi| \le gJ_c, \qquad m \ge 0, \qquad m\,(|\nabla\psi| - gJ_c) = 0. \tag{282}$$

These equations are coupled to (267)–(271) for Λ order one.

In a virgin sample in an increasing magnetic field there will be a region in the interior, D_1 say, which vortices have not yet reached, and a region around the boundary, $D_2 = D\backslash D_1$ say, in which the current is equal to the critical value.

In a strip $-l < x < l$, with $\hat{a} = 1$ and J_c constant, and with D_1 given by $-s < x < s$, this gives

$$\frac{\partial J_2}{\partial x} + h_3(x,0) = 0 \quad \text{for } |x| < s, \tag{283}$$

$$J_2 = J_c \quad \text{for } -l < x < -s, \tag{284}$$

$$J_2 = -J_c \quad \text{for } s < x < l, \tag{285}$$

with J_2 continuous at $x = \pm s$, and

$$\omega = 0 \quad \text{for } |x| < s, \tag{286}$$

$$\omega = h_3(x,0) \quad \text{for } s < |x| < l, \tag{287}$$

where

$$h_3(x,0) - h_{app,3}(x,0) = \frac{1}{\pi}\int_{-l}^{l} \frac{gJ_2(\bar{x})}{2\Lambda(\bar{x}-x)}\,d\bar{x}. \tag{288}$$

5.2. Limiting cases. As $\Lambda \to \infty$ the problems for the current decouples from the problem for the magnetic field, since, as usual, $h_3(x,y,0)$ in (281) is simply replaced by $h_{app,3}(x,y,0)$. Then we have

$$\frac{\partial \omega}{\partial t} = \operatorname{div}(m\nabla\psi), \tag{289}$$

$$-\nabla\cdot\left(\frac{1}{\hat{a}g}\nabla\psi\right) + h_{app,3} = \omega, \tag{290}$$

with the constraints

$$|\nabla\psi| \le gJ_c, \qquad m \ge 0, \qquad m\,(|\nabla\psi| - gJ_c) = 0. \tag{291}$$

In this limit with $g = 1$ the free-boundary problem (283)–(285) gives

$$J_2 = J_c \quad \text{for } -L < x < -s, \tag{292}$$

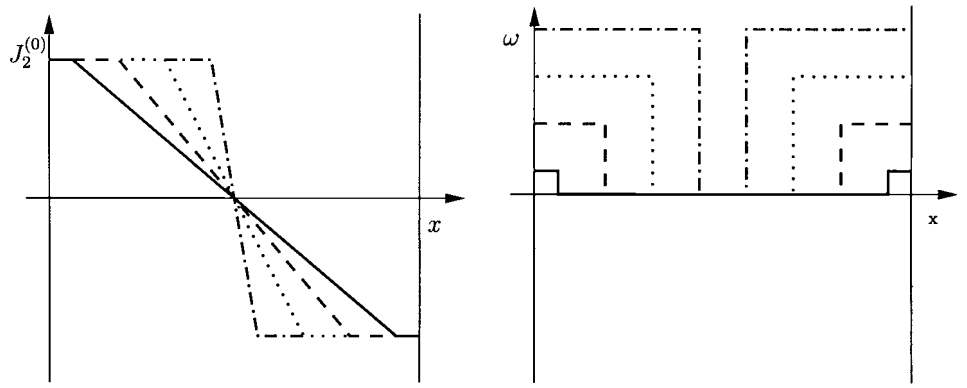
$$J_2 = -h_{app,3}x \quad \text{for } -s < x < s, \tag{293}$$

$$J_2 = -J_c \quad \text{for } s < x < L, \tag{294}$$

$$s = \frac{J_c}{h_{app,3}}, \tag{295}$$

with

$$\omega = 0 \quad \text{for } |x| < s, \tag{296}$$



(a) (b)

FIG. 6. *Solutions of the thin-film critical-state model in a thin strip in the limit* $\Lambda \to \infty$.

$$\omega = h_{app,3}(x, y, 0) \quad \text{for } s < |x| < l. \tag{297}$$

These solutions are plotted in Figure 6 for an increasing applied magnetic field $h_{app,3}$.

For comparison, the corresponding solutions for a two-dimensional critical-state model in a slab are

$$J_2 = J_c \quad \text{for } -L < x < -s, \tag{298}$$

$$J_2 = 0 \quad \text{for } -s < x < s, \tag{299}$$

$$J_2 = -J_c \quad \text{for } s < x < L, \tag{300}$$

$$s = l - \frac{h_{app,3}}{J_c}, \tag{301}$$

with

$$\omega = 0 \quad \text{for } |x| < s, \tag{302}$$

$$\omega = J_c(x - s) \quad \text{for } s < x < l, \tag{303}$$

$$\omega = -J_c(x + s) \quad \text{for } -l < x < -s, \tag{304}$$

which are shown in Figure 7.

In the second limit, $\Lambda \to 0$, we must again scale $\boldsymbol{J}$ with Λ to give

$$\operatorname{curl} \boldsymbol{h} = \boldsymbol{J}_{ext}, \tag{305}$$

$$\operatorname{div} \boldsymbol{h} = 0, \tag{306}$$

with

$$[\boldsymbol{e}_z \wedge \boldsymbol{h}] = (\psi_y, -\psi_x, 0) \quad \text{for} \quad (x, y) \in D, \tag{307}$$

$$[\boldsymbol{e}_z \cdot \boldsymbol{h}] = 0, \tag{308}$$

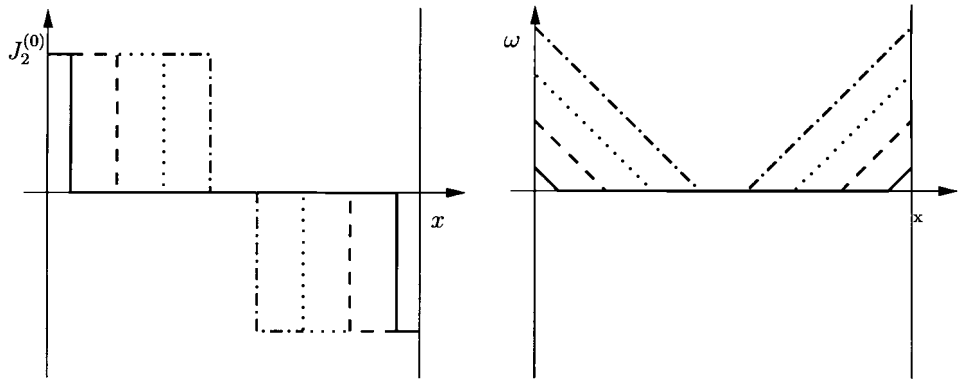
$$\boldsymbol{h} \to \boldsymbol{h}_{ext} \quad \text{as} \quad |\boldsymbol{r}| \to \infty, \tag{309}$$

$$h_3(x, y, 0) = \omega. \tag{310}$$

$$\frac{\partial \omega}{\partial t} = \operatorname{div}(m \nabla \psi), \tag{311}$$

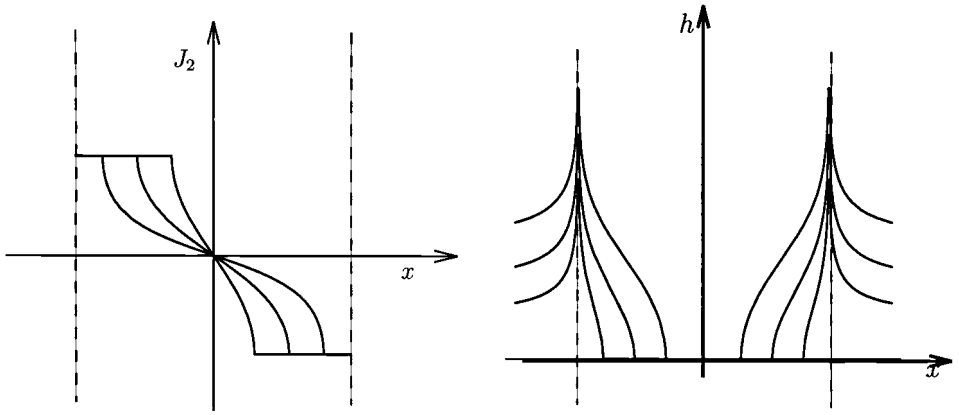
with the constraints

$$|\nabla \psi| \le g J_c, \qquad m \ge 0, \qquad m\,(|\nabla \psi| - g J_c) = 0. \tag{312}$$



(a) (b)

FIG. 7. *Solutions of the two-dimensional critical-state model in a slab.*



(a) (b)

FIG. 8. *Solutions of the thin-film critical-state model in a thin strip in the limit* $\Lambda \to 0$.

In this case with $g = 1$ the free-boundary problem (283)–(285) becomes

$$h_3(x,0) = h_{app,3} + \frac{1}{\pi}\int_{-l}^{l} \frac{J_2(\bar{x})}{2(\bar{x}-x)}\, d\bar{x} = 0 \quad \text{for } |x| < s, \tag{313}$$

$$J_2 = J_c \quad \text{for } -l < x < -s, \tag{314}$$

$$J_2 = -J_c \quad \text{for } s < x < l. \tag{315}$$

This problem has been considered previously in [4, 10, 29, 31]. Solutions are plotted in Figure 8. Note that the magnetic field tends to infinity as $x \to \pm l$; there is an inner region of width Λ in which the problems for $\boldsymbol{J}$ and $\boldsymbol{h}$ are coupled again and which regularizes the magnetic field.

In both [4] and [31] field and current profiles are calculated for superconducting strips in an applied field, with an applied current, and with both an applied field and

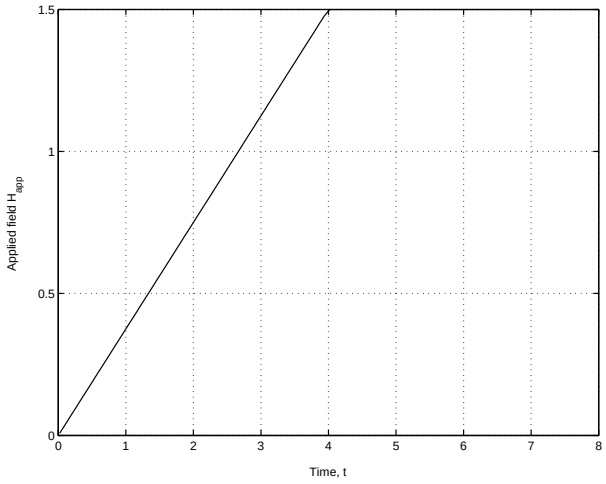


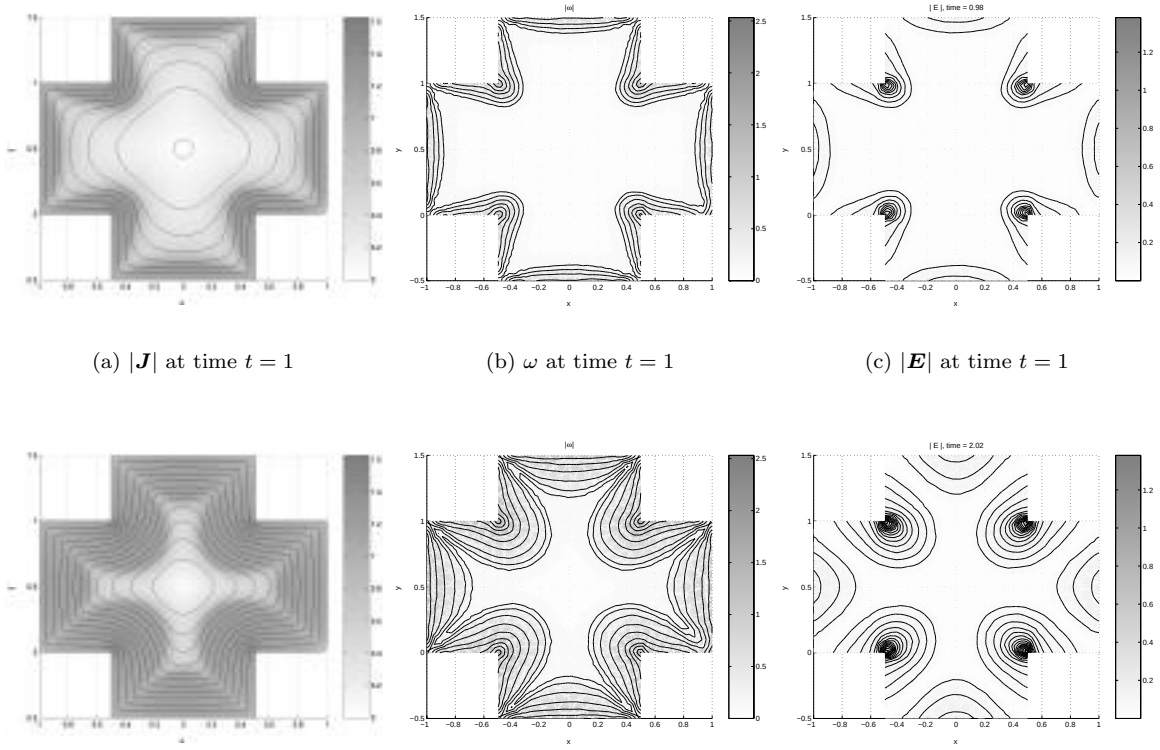
FIG. 9. *Variation of the applied magnetic field with time for the simulations shown in Figures* 10 *and* 11.

an applied transport current. In [29] the theoretical profiles are compared with experimentally measured profiles of the magnetic field distribution in a superconducting slab in an applied magnetic field, and the agreement is shown to be good. Schuster et al. also consider experimentally the effects of inhomogeneous pinning. If the critical current is lower in the center of a sample than it is at the outside, there can be a sudden jump in the free boundary as the applied magnetic field reaches a critical value. This effect is also examined in [20], where its implications for the flux flow regime under an applied current are considered.

In [29] the profiles of magnetic field distribution on a square film in an applied field are measured experimentally, and in [28] the profiles on a film in the shape of a cross are measured experimentally and calculated theoretically, using a power-law dependence of $\boldsymbol{E}$ on $\boldsymbol{J}$. In [20] field and current distributions for general two-dimensional films are calculated numerically in both the critical state and flux flow regimes (using general vortex velocity laws as well as general nonlinear Ohm's laws), both with an applied magnetic field and a transport current. This work also includes the heating effect of the electric field coupled with an equation for the evolution of temperature, with temperature dependent critical fields and vortex mobilities, but a discussion of this extension is beyond the scope of the present paper.

The solutions for a cross in an increasing applied magnetic field are reproduced in Figure 10. The field is ramped up from zero to 1.5 over 4 time units and then held fixed until the sample reaches steady state, as illustrated in Figure 9. Figure 10 shows the current, vorticity, and electric field at times $t = 1$, 2, 3, and 4. Figure 11 shows the current and vorticity in the steady state which is reached at about $t = 8$ (the electric field is zero in steady state). In Figures 10(a), (d), (g), (j) and 11(a) the shading shows the magnitude of the current density and the lines show the direction of current flow. In the other figures the lines are contour lines of constant ω or $|\boldsymbol{E}|$.

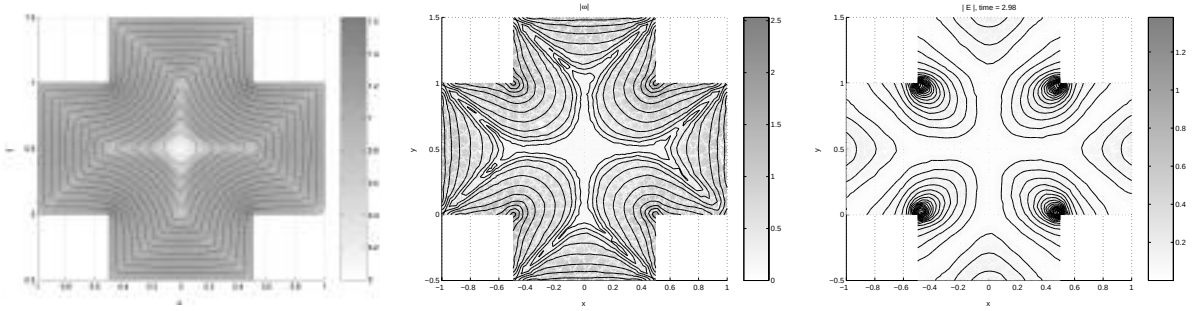
The vortices enter first at the reentrant corners of the cross, and this is where the electric field is largest. In the thermal problem the heating at these points may significantly affect the solution.



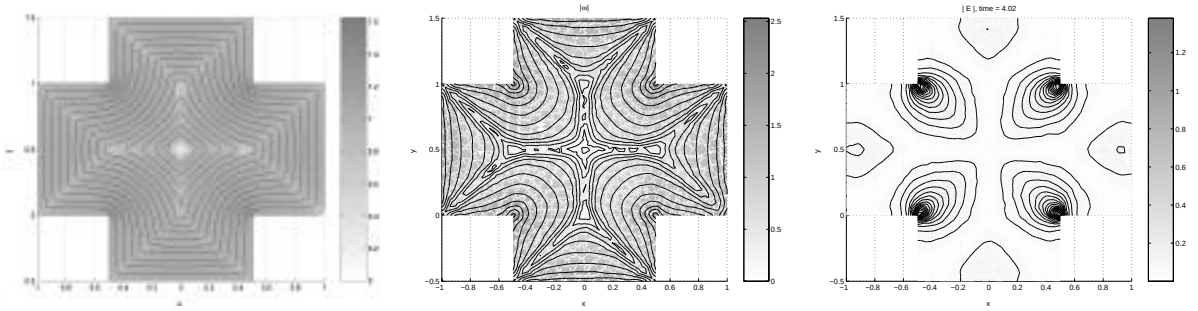
(a) $|\boldsymbol{J}|$ at time $t = 1$ (b) ω at time $t = 1$ (c) $|\boldsymbol{E}|$ at time $t = 1$

(d) $|\boldsymbol{J}|$ at time $t = 2$ (e) ω at time $t = 2$ (f) $|\boldsymbol{E}|$ at time $t = 2$

FIG. 10. *See caption on following page.*

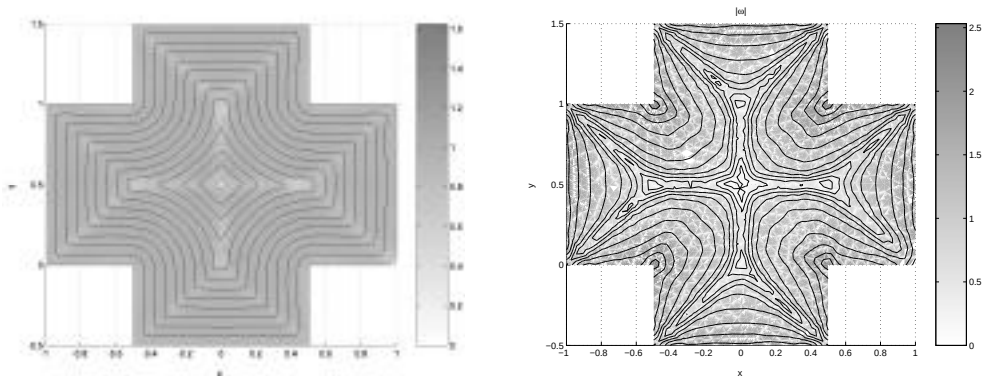


(g) $|\boldsymbol{J}|$ at time $t = 3$ (h) ω at time $t = 3$ (i) $|\boldsymbol{E}|$ at time $t = 3$



(j) $|\boldsymbol{J}|$ at time $t = 4$ (k) ω at time $t = 4$ (l) $|\boldsymbol{E}|$ at time $t = 4$

FIG. 10. *Solutions of the thin-film critical-state model of a cross in an applied magnetic field in the limit* $\Lambda \to 0$. *The magnetic field is ramped up from zero to* 1.5 *over* 4 *time units. Numerical solution due to A. D. Grief* [20].



(a) $|\boldsymbol{J}|$ in steady state

(b) ω in steady state

FIG. 11. *Steady state of the thin-film critical-state model of a cross in an applied magnetic field of strength* 1.5 *in the limit* $\Lambda \to 0$. *Numerical solution due to A. D. Grief* [20].

6. Conclusion. We have developed a hierarchy of models of superconducting thin films, allowing for a spatially varying film thickness and a spatially varying equilibrium density of superconducting electrons.

We began with the Ginzburg–Landau model, where we found that the canonical scaling was to let $\epsilon = d/L \to 0$ with $\Xi = \xi/L$ and $\Lambda = \lambda_{\text{eff}}/L = \lambda^2/dL$ fixed, where λ_{eff} is the effective screening length. This corresponds to letting the Ginzburg–Landau parameter $\kappa = \lambda/\xi$ tend to zero like $\epsilon^{1/2}$. The London limit of this model, in which vortices appear as δ-function singularities, corresponds to the limit in which the increased Ginzburg–Landau parameter $\kappa_{\text{eff}} = \Lambda/\Xi = \lambda\kappa/d \to \infty$, in contrast to bulk superconductors, where it corresponds to the limit $\kappa \to \infty$. This explains why thin films of even type-I superconductors will exhibit vortex solutions when the thickness becomes smaller than the penetration depth.

Following on from the London limit we considered the situation in which the vortex separation tends to zero and the vortices can be averaged to produce a vortex density.

Finally we considered critical state models, in which the pinning potential is homogenized to give stick/slip mobility laws.

In each case we found that a key parameter is $\Lambda = \lambda^2/Ld$. If this parameter is order one, corresponding to a lateral film dimension of the same order as the effective screening length λ_{eff}, then the problems for the electric current in the film and the magnetic field outside it are coupled. If this parameter is large, then the applied magnetic field passes straight through the film to leading order, and the problem for the current inside the film is decoupled. If Λ is small, then the model also simplifies, since then the vorticity in the film determines the magnetic field, which subsequently determines the current in the film.

Our hierarchy contains one or two models which have appeared elsewhere in the literature. The thin-film limit of the Ginzburg–Landau equations was considered in [8], but in the limit $d/L \to 0$ with ξ and λ fixed. Thus the model they obtained corresponds to the limit $\Lambda \to \infty$.

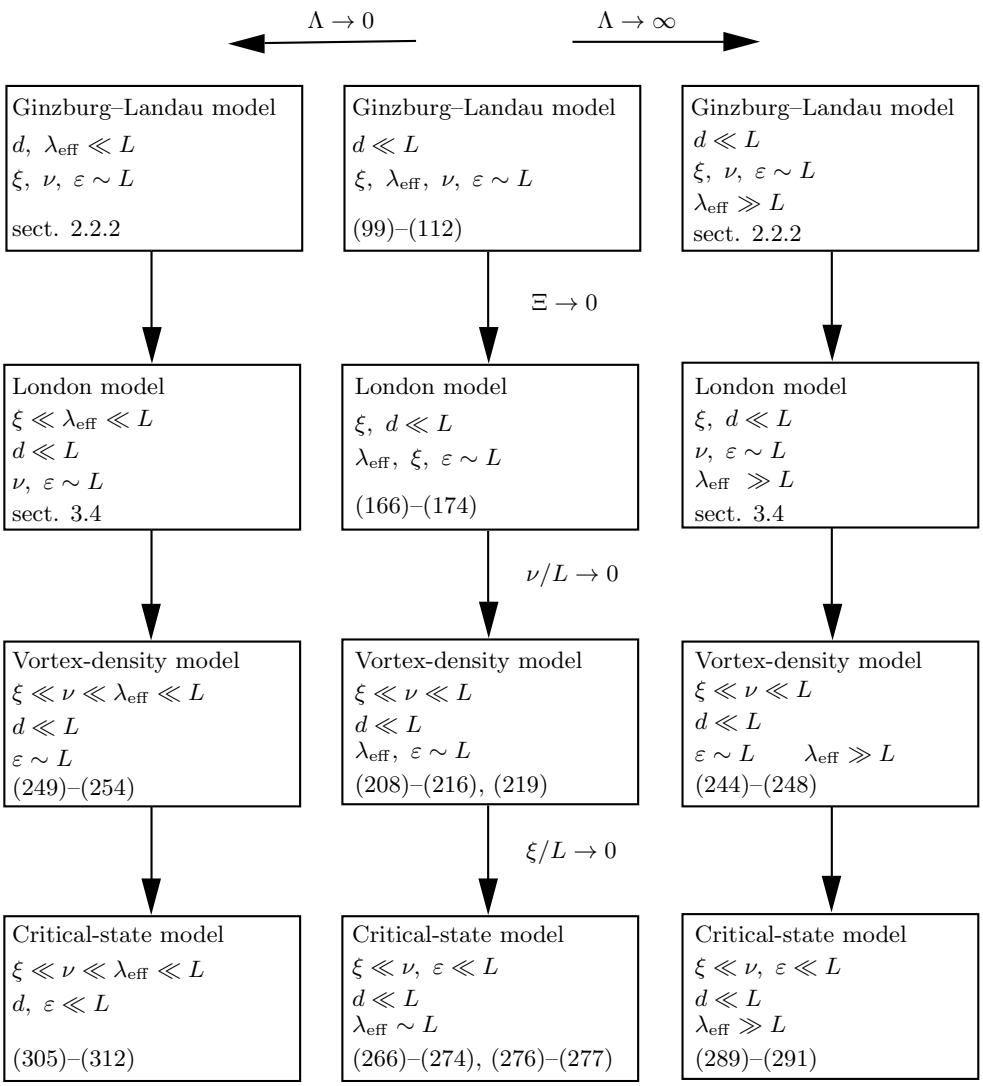


FIG. 12. *The hierarchy of models for superconducting thin films. Recall that* $\Xi = \xi/L$ *and* $\Lambda = \lambda_{\text{eff}}/L$. *Numbers refer to equation numbers in the text. The lengthscales are defined in Table* 1.

The London model for uniform strips in the absence of vortices was considered in [16] and independently for variable thickness films in two lateral dimensions with vortices in [10, 21].

The thin-film critical-state model has been considered in [29] and [31]. However, because they took the thin-film limit of the bulk critical state model, in which λ had already been set to zero, their model corresponds to the $\Lambda \to 0$ limit.

A summary of the hierarchy of models we have derived is shown in Figure 12, with a summary of the definitions of the various lengthscales in Table 1.

Acknowledgments. The first author would like to thank Andrew Grief for many useful discussions, for identifying errors in a preliminary version of this manuscript,

TABLE 1
Definition of the lengthscales in Figure 12.

λ	penetration depth
ξ	coherence length
d	typical film thickness
L	typical film width
ν	vortex separation
ε	lengthscale of pinning potential
$\lambda_{\text{eff}} = \lambda^2/d$	effective penetration depth

and for kindly allowing us to present some of his numerical solutions of thin-film models in Figures 10 and 11.

REFERENCES

[1] I. ARANSON, M. GLITTERMAN, AND B. Y. SHAPIRO, *Onset of vortices in thin superconducting strips and wires*, Phys. Rev. B, 51 (1995), pp. 3092–3096.
[2] J. BARDEEN, L. N. COOPER, AND J. R. SCHRIEFFER, *Theory of superconductivity*, Phys. Rev. (2), 108 (1957), pp. 1175–1204.
[3] M. S. BERGER AND Y. Y. CHEN, *Symmetric vortices for the Ginzburg-Landau equations of superconductivity and the nonlinear desingularization phenomenon*, J. Funct. Anal., 82 (1989), pp. 259–295.
[4] E. H. BRANDT AND M. INDENBOM, *Type-*II *superconductor strip with current in a perpendicular magnetic field*, Phys. Rev. B, 48 (1993), pp. 12893–12906.
[5] S. J. CHAPMAN, *A mean-field model of superconducting vortices in three dimensions*, SIAM J. Appl. Math., 55 (1995), pp. 1259–1274.
[6] S. J. CHAPMAN, *A hierarchy of models for type-*II *superconductors*, SIAM Rev., 42 (2000), pp. 555–598.
[7] S. J. CHAPMAN, Q. DU, AND M. D. GUNZBURGER, *A Ginzburg-Landau-type model of superconducting/normal junctions including Josephson junctions*, European J. Appl. Math., 6 (1995), pp. 97–114.
[8] S. J. CHAPMAN, Q. DU, AND M. D. GUNZBURGER, *A model for variable thickness superconducting thin films*, Z. Angew. Math. Phys., 47 (1996), pp. 410–431.
[9] S. J. CHAPMAN, A. D. GRIEF, S. D. HOWISON, M. D. MCCULLOCH, D. DEW-HUGHES, J. MOORE, AND C. M. GROVENOR, *Vortex velocity laws to* I-V *data for flat superconductors*, IEEE Trans. Appl. Supercon., 11 (2001), pp. 3943–3946.
[10] S. J. CHAPMAN AND D. R. HERON, *The motion of superconducting vortices in thin films of varying thickness*, SIAM J. Appl. Math., 58 (1998), pp. 1808–1825.
[11] S. J. CHAPMAN AND G. RICHARDSON, *Motion of vortices in type* II *superconductors*, SIAM J. Appl. Math., 55 (1995), pp. 1275–1296.
[12] S. J. CHAPMAN AND G. RICHARDSON, *Vortex pinning by inhomogeneities in type-*II *superconductors*, Phys. D, 108 (1997), pp. 397–407.
[13] X. CHEN, C. M. ELLIOTT, AND Q. TANG, *Shooting method for vortex solutions of a complex-valued Ginzburg-Landau equation*, Proc. Roy. Soc. Edinburgh Sect. A, 124 (1994), pp. 1075–1088.
[14] J. DEANG, Q. DU, M. D. GUNZBURGER, AND J. PETERSON, *Vortices in superconductors: Modelling and computer simulations*, Philos. Trans. Roy. Soc. London Ser. A, 355 (1997), pp. 1957–1968.
[15] K. DECKELNICK, C. M. ELLIOTT, AND G. RICHARDSON, *Long time asymptotics for forced curvature flow with applications to the motion of a superconducting vortex*, Nonlinearity, 10 (1997), pp. 655–678.
[16] A. T. DORSEY, *Linear response of thin superconductors in perpendicular magnetic fields: An asymptotic analysis*, Phys. Rev. B, 51 (1995), pp. 15329–15343.
[17] V. L. GINZBURG AND L. D. LANDAU, *On the theory of superconductivity*, JETP, 20 (1950), pp. 1064–1082.
[18] L. P. GOR'KOV, *Microscopic derivation of the Ginzburg-Landau equations in the theory of superconductivity*, Soviet Phys. JETP, 9 (1959), pp. 1364–1367.

[19] L. P. Gor'kov and G. M. Éliashberg, *Generalisation of the Ginzburg-Landau equations for non-stationary problems in the case of alloys with paramagnetic impurities*, Soviet Phys. JETP, 27 (1968), pp. 328–334.
[20] A. D. Grief, *Superconducting Thin Films*, Ph.D. thesis, University of Oxford, Oxford, UK, 2003.
[21] D. R. Heron, *Mathematical Methods in Superconductivity*, Ph.D. thesis, University of Oxford, Oxford, UK, 1995.
[22] A. I. Larkin and Y. N. Ovchinnikov, *Influence of inhomogeneities on superconductor properties*, Soviet Phys. JETP, 34 (1972), pp. 651–655.
[23] K. Likharev, *Superconducting weak links*, Rev. Modern Phys., 51 (1979), pp. 101–159.
[24] J. Pearl, *Current distribution in superconducting films carrying quantized fluxoids*, Appl. Phys. Lett., 5 (1964), pp. 65–66.
[25] L. Peres and J. Rubinstein, *Vortex dynamics in $U(1)$ Ginzburg-Landau models*, Phys. D, 64 (1993), pp. 299–309.
[26] G. Richardson, *Instability of a superconducting line vortex*, Phys. D, 110 (1997), pp. 139–153.
[27] A. Schmid, *A time dependent Ginzburg-Landau equation and its application to the problem of resistivity in the mixed state*, Physik der Kondensierten Materie, 5 (1966), p. 302.
[28] T. Schuster, H. Kuhn, and E. H. Brandt, *Flux penetration into flat superconductors of arbitrary shape: Patterns of magnetic and electric fields and current*, Phys. Rev. B, 54 (1996), pp. 3514–3524.
[29] T. Schuster, H. Kuhn, E. H. Brandt, M. Indenbom, M. Koblischka, and M. Konczykowski, *Flux motion in thin superconductors with inhomogeneous pinning*, Phys. Rev. B, 50 (1994), p. 16684.
[30] M. Tinkham, *Effect of fluxoid quantization on transitions of superconducting films*, Phys. Rev., 129 (1963), p. 2413.
[31] E. Zeldov, J. R. Clem, M. McElfresh, and M. Darwin, *Magnetization and transport currents in thin superconducting films*, Phys. Rev. B, 49 (1994), p. 9802.

SIAM J. APPL. MATH.
Vol. 63, No. 6, pp. 2128–2154

IMAGE SEGMENTATION USING ACTIVE CONTOURS: CALCULUS OF VARIATIONS OR SHAPE GRADIENTS?*

GILLES AUBERT†, MICHEL BARLAUD‡, OLIVIER FAUGERAS§, AND STÉPHANIE JEHAN-BESSON‡

Abstract. We consider the problem of segmenting an image through the minimization of an energy criterion involving region and boundary functionals. We show that one can go from one class to the other by solving Poisson's or Helmholtz's equation with well-chosen boundary conditions. Using this equivalence, we study the case of a large class of region functionals by standard methods of the calculus of variations and derive the corresponding Euler–Lagrange equations. We revisit this problem using the notion of a shape derivative and show that the same equations can be elegantly derived without going through the unnatural step of converting the region integrals into boundary integrals. We also define a larger class of region functionals based on the estimation and comparison to a prototype of the probability density distribution of image features and show how the shape derivative tool allows us to easily compute the corresponding Gâteaux derivatives and Euler–Lagrange equations. Finally we apply this new functional to the problem of regions segmentation in sequences of color images. We briefly describe our numerical scheme and show some experimental results.

Key words. image segmentation, region segmentation, active contours, active regions, image statistics, region functionals, boundary functionals, calculus of variations, shape optimization, shape gradient, Euler–Lagrange equations, Gâteaux derivative, Parzen window estimation, level set methods

AMS subject classifications. 35, 35K, 49, 49Q10, 68U10

DOI. 10.1137/S0036139902408928

1. Introduction.

Many problems in image processing, such as segmentation, tracking, or classification, can be cast in the framework of optimization theory, e.g., as the minimization of some energy measure. The energy is often some combination of region or boundary functionals. The minimization is usually not trivial, and many methods have been developed to reach an optimum which may be only local.

We address here the problem of the optimization of region or boundary functionals with the method of active contours. Active contours have been introduced by Kass, Witkin and Terzopoulos [34] and were originally boundary methods. Snakes [34], balloons [10], or geodesic active contours [4] are driven towards the edges of an image through the minimization of a boundary integral of functions of features depending on edges. Active contours driven by region functionals in addition to boundary functionals have appeared later. Introduced by [11] and [43], they have been further developed in [52, 5, 9, 38, 39, 40, 41, 21, 51] and [31, 33]. In effect, the use of active contours for the optimization of a criterion including both region and boundary functionals appears to be really powerful.

In general, features of the image region to be segmented, tracked, etc., ... are embedded in region functionals while the boundary functional allows smoothness and

*Received by the editors May 29, 2002; accepted for publication (in revised form) January 29, 2003; published electronically September 17, 2003.
http://www.siam.org/journals/siap/63-6/40892.html
†Laboratoire Dieudonné, CNRS-UNSA, Parc Valrose, 06108 Nice Cedex 2, France (gaubert@math.unice.fr).
‡Laboratoire I3S, CNRS-UNSA, 2000 route des Lucioles, 06903 Sophia Antipolis, France (barlaud@i3s.unice.fr, jehan@i3s.unice.fr).
§INRIA Sophia Antipolis, 2004 route des Lucioles, BP 93 06902, Sophia-Antipolis Cedex, France (faugeras@sophia.inria.fr).

regularity of the region boundary. The basic principle is to construct a parabolic partial differential equation (PDE) from the energy criterion, e.g., by computing some sort of Euler–Lagrange equations; this PDE changes the shape of the current curve according to some velocity field which can be thought of as a direction of descent of the energy criterion. Given a closed curve enclosing an initial region one then computes the solution of this PDE for this initial condition. The corresponding family of curves decreases the energy criterion and converges toward a (local) minimum of the criterion hopefully corresponding to the objects to be segmented. To compute such a PDE, several methods have been proposed.

Some authors do not compute the theoretical expression of the velocity field (basically the gradient of the energy criterion) but choose a deformation of the curve that will make the criterion decrease [5, 9] (they compute a direction of descent). Other authors [52, 39, 41] compute the theoretical expression of the velocity vector from the Euler–Lagrange equations. The computation is performed in three main steps. First, region integrals representing region functionals are transformed into boundary integrals using the Green–Riemann theorem. Second, the corresponding Euler–Lagrange equations are derived and used to define a dynamic scheme to evolve the initial region. Another alternative is to keep the region formulation to compute the gradient of the energy criterion with respect to the region boundary instead of reducing region integrals to boundary integrals. In [21], the authors propose computing the derivative of the criterion while taking into account the discontinuities across the contour. In [31, 33] the computation of the evolution equation is achieved through shape derivation principles.

This computation becomes more involved when global information about regions is present in the energy criterion, the so-called region-dependent case. It happens, for example, when statistical features of a region such as, for example, the mean or the variance of the intensity are involved in the minimization. This case has been studied in [6, 7, 20, 21, 31, 33, 51]. In [31, 33] the authors show that the minimization of functionals involving region-dependent features induces additional terms in the evolution equation of the active contour that are important in practice. These additional terms are easily computed thanks to shape derivation tools.

In this article we clarify the relationships between the boundary and region functionals that arise naturally in several image processing tasks. We show in section 3 that one can go from one to the other by solving Poisson's equation with Dirichlet conditions or Helmholtz's equation with Neumann conditions.

We then concentrate on the problem of finding local minima of a large class of region functionals. In section 4 we first transform them into boundary functionals and apply methods from the calculus of variations to compute the corresponding Gâteaux derivatives and construct a velocity field on the region boundary. This field defines a PDE whose solution for a given initial boundary condition defines a one-parameter family of regions which, in practice, converges towards a local minimum of the functional. The problem of the existence and uniqueness of a solution to this PDE is not addressed in this article.

We next change our point of view and rederive the same equations in a simpler and more natural way, i.e., without going through the trouble of turning region integrals into boundary integrals; this is achieved in section 5 by applying shape derivation methods [49, 22]. This line of approach has already been followed in [46] in his work on the estimation of the optical flow.

We then turn in section 6 our attention to a new class of region-based functionals

by considering histograms of image features. The shape derivation tools allow us to easily derive the velocity field that defines the evolution of the region boundary.

Section 7 is devoted to an application of the previous methods to the problem of region segmentation with a given color histogram in a sequence of images. Our experimental results show that the technique indeed has some interesting potentials.

2. Problem statement. In many image processing problems, the issue is to find a set of image regions that minimize a given error criterion. This criterion is often a combination of region and boundary functionals.

A local minimizer for such a criterion including both region and boundary functionals is usually hard to compute. This is mostly due to the fact that the set of image regions, i.e., the set of regular open domains in $\mathbb{R}^n$ (whose boundary is a closed, C^2 manifold), does not have a structure of vector space, preventing us from using in a straightforward fashion gradient descent methods. In order to circumvent this difficulty, calculus of variations and shape optimization techniques can be brought to bear on the problem. The basic idea is to use them in order to derive a PDE that will drive the boundary of an initial region toward a local minimum of the error criterion. The key point is to compute the velocity vector at each point of the boundary at each time instant. In this paper we propose a framework for achieving these goals in a number of practically important cases.

To fix ideas in the two-dimensional case, the evolving boundary, or active contour, is modeled by a parametric curve $\Gamma(s,\tau) = (x_1(s,\tau), x_2(s,\tau))$, where s may be its arc-length and τ is an evolution parameter—the time. The active contour is then driven by the following PDE:

$$\Gamma_\tau \stackrel{def}{=} \frac{\partial \Gamma(s,\tau)}{\partial \tau} = \mathbf{v} \qquad \text{with} \quad \Gamma(\tau = 0) = \Gamma_0, \tag{2.1}$$

where Γ_0 is an initial curve defined by the user and $\mathbf{v}$ the velocity vector of $\Gamma(s,\tau)$. This velocity is the unknown that must be derived from the error criterion so that the solution $\Gamma(.,\tau)$ converges towards a curve achieving a local minimum and thus, hopefully, towards the boundary of the object to segment as $\tau \to \infty$.

2.1. Boundary and region functionals. Let us now define more precisely the region and boundary functionals. Let $\mathcal{U}$ be a class of domains (open, regular bounded sets, i.e., C^2) of $\mathbb{R}^n$ and Ω an element of $\mathcal{U}$ of boundary $\partial\Omega$, which we sometimes denote Γ. A boundary functional, J_b, may be expressed as a boundary integral of some scalar function g of image features:

$$J_b(\partial\Omega) = \int_{\partial\Omega} g(\mathbf{x})\, d\mathbf{a}(\mathbf{x}), \tag{2.2}$$

where $\partial\Omega$ is the boundary of the region and $d\mathbf{a}$ its area element. The derivation of this boundary functional is classical [4, 35] and leads to the following velocity vector:

$$\mathbf{v}_b = [g(\mathbf{x})\kappa - \nabla g(\mathbf{x}) \cdot \mathbf{N}]\, \mathbf{N},$$

where $\mathbf{N}$ is the inward unit normal vector of Γ and κ its mean curvature. The idea is to use a local parametrization of Γ to reduce (2.2) to a standard problem in the calculus of variations.

A region functional, J_r, may be expressed as an integral in a domain Ω of $\mathcal{U}$ of some function f of some region features:

$$J_r(\Omega) = \int_\Omega f(\mathbf{x}, \Omega) d\mathbf{x}. \tag{2.3}$$

In that case, the computation of the velocity vector for (2.1) is not as easy. We propose comparing two main approaches. The first approach is based upon the idea of transforming all functionals into boundary functionals, thereby reducing (through a local parametrization of the boundary) the problem of minimization to a standard problem in the calculus of variations from which the computation of the Gâteaux derivatives follows. The second approach is based upon the use of shape derivation tools. In a sense it is more natural since it keeps the region representation.

Note that the scalar function f in (2.3) is generally region-dependent. This is important since this dependency on the region must be taken into account when searching for a local minimum of the functional, as discussed in later sections.

Also note that we could have added a dependency of g on $\partial\Omega$, i.e., write $g(\mathbf{x}, \partial\Omega)$ in (2.2), to keep the symmetry with the region functional. This is not necessary since the results in section 4.2, in particular Theorem 4.6, do in fact provide an answer for this case.

2.2. Examples of such optimization problems in image processing. An image is represented by its intensity $I(\mathbf{x})$ defined on some region of $\mathbb{R}^n$.

Active contours were originally introduced to search for minima of boundary functionals. In [4, 35], the function g is a function of the magnitude of the image gradient through a strictly decreasing function $\varphi : [0, +\infty[\to \mathbb{R}^+$ such that $\varphi(r) \to 0$ as $r \to +\infty$. Hence $g(\mathbf{x}) = \varphi(|\nabla I(\mathbf{x})|)$. The minimization amounts to the minimization of the length of a curve in a Riemannian space. Local minima are obtained via the steepest descent method.

Region functionals have also been introduced. The region information is embedded in the function f of (2.3). These functionals have been used for many applications such as moving objects detection [38, 40, 30, 32], image segmentation [5, 21, 7, 39, 40, 51], or classification [52, 44, 45, 41]. For example, people have used statistical features such as the mean or the variance of a region Ω:

$$\begin{cases} \mu_\Omega & = \frac{1}{|\Omega|} \int_\Omega I(\mathbf{x}) d\mathbf{x} \quad \text{with} \quad |\Omega| = \int_\Omega d\mathbf{x}, \\ \sigma^2_\Omega & = \frac{1}{|\Omega|} \int_\Omega (I(\mathbf{x}) - \mu_\Omega)^2 d\mathbf{x}. \end{cases}$$

We use these two examples to motivate the introduction of a general region functional

$$J_r(\Omega) = \int_\Omega f(\mathbf{x}, G_1(\Omega), G_2(\Omega), \ldots, G_m(\Omega))\, d\mathbf{x}, \tag{2.4}$$

where the functionals G_i are given by

$$G_i(\Omega) = \int_\Omega H_i(\mathbf{x}, \Omega)\, d\mathbf{x}, \quad i = 1 \ldots m. \tag{2.5}$$

As shown in this equation, the function H_i is itself region-dependent; more precisely,

$$H_i(\mathbf{x}, \Omega) \stackrel{def}{=} H_i(\mathbf{x}, K_{i1}(\Omega), K_{i2}(\Omega), \ldots, K_{il_i}(\Omega)), \tag{2.6}$$

where

$$K_{ij}(\Omega) = \int_\Omega L_{ij}(\mathbf{x})\ d\mathbf{x}, \quad j = 1 \ldots l_i \quad i = 1 \ldots m. \tag{2.7}$$

Note that we have stopped the process at the second level but it could conceivably continue. We have chosen this special case of dependency because it often arises in applications, as shown in the next two sections. The various methods that we develop can be extended in a fairly straightforward fashion to more complicated situations if needed; see, for example, section 6.

2.3. An example involving the mean. Let us choose

$$f(\mathbf{x}, \Omega) = \varrho(I(\mathbf{x}) - \mu_\Omega), \tag{2.8}$$

where $\varrho : \mathbf{R} \to \mathbf{R}^+$ is a positive function of class C^1. f is region-dependent. This is an example where the process described in the previous section stops at the first level:

$$J(\Omega) \;=\; \int_\Omega f(\mathbf{x}, \Omega)\, d\mathbf{x} \;=\; \int_\Omega \varrho(I(\mathbf{x}) \,-\, \mu_\Omega)\, d\mathbf{x} \;=\; \int_\Omega \varrho\left(I(\mathbf{x}) - \frac{G_1(\Omega)}{G_2(\Omega))}\right) d\mathbf{x},$$

where

$$G_1(\Omega) = \int_\Omega H_1(\mathbf{x}, \Omega)\, d\mathbf{x} \quad \text{with} \quad H_1(\mathbf{x}, \Omega) = I(\mathbf{x})\,,$$
$$G_2(\Omega) = \int_\Omega H_2(\mathbf{x}, \Omega)\, d\mathbf{x} \quad \text{with} \quad H_2(\mathbf{x}, \Omega) = 1\,.$$

In this case, the functions H_i, $i = 1, 2$, do not depend on the region Ω, $l_1 = l_2 = 0$, and $K_{ij}(\mathbf{x}) = 0$ for all i, j.

2.4. An example involving the variance. Let us take an example where we stop the process at the second level. Consider the case where the function f is a function of the variance given by

$$f(\mathbf{x}, \Omega) = \varrho(\sigma^2_\Omega). \tag{2.9}$$

$\varrho : \mathbf{R}^+ \to \mathbf{R}^+$ is of class C^1. We write

$$J(\Omega) = \int_\Omega f(\mathbf{x}, \Omega)\, d\mathbf{x} = \int_\Omega \varrho(\sigma^2_\Omega)\, d\mathbf{x} = \int_\Omega \varrho\left(\frac{G_1(\Omega)}{G_2(\Omega))}\right) d\mathbf{x}\,.$$

Therefore we have

$$G_1(\Omega) = \int_\Omega H_1(\mathbf{x}, \Omega)\, d\mathbf{x}\,, \quad H_1(\mathbf{x}, \Omega) = (I(\mathbf{x}) - \mu_\Omega)^2\,,$$
$$G_2(\Omega) = \int_\Omega H_2(\mathbf{x}, \Omega)\, d\mathbf{x}\,, \quad H_2(\mathbf{x}, \Omega) = 1\,,$$

with

$$H_1(\mathbf{x}, \Omega) = \left(I(\mathbf{x}) - \frac{K_{11}}{K_{12}}\right)^2\,, \quad l_1 = 2,$$
$$H_2(\mathbf{x}, \Omega) = 1\,, \quad l_2 = 0\,,$$

and finally

$$K_{11}(\Omega) = \int_\Omega I(\mathbf{x})\, d\mathbf{x}\,, \quad L_{11}(\mathbf{x}) = I(\mathbf{x})\,,$$
$$K_{12}(\Omega) = \int_\Omega d\mathbf{x}\,, \quad L_{12}(\mathbf{x}) = 1\,.$$

3. Expression of region functionals as boundary functionals and vice versa. In this section, we show that a region functional may always be expressed as a boundary functional and vice versa.

3.1. Tranformation of region functionals into boundary functionals. Consider the region functional (2.3); the next proposition shows that, under some reasonable assumptions on the function f, it can always be turned into a boundary functional (2.2).

PROPOSITION 3.1. *Let Ω be a bounded open set with regular boundary $\partial\Omega$. Let $f : \overline{\Omega} \to \mathbb{R}$ be a continuous function and u be the unique solution of Poisson's equation:*

$$\left\{ \begin{array}{rcll} -\Delta u & = & f & in \quad \Omega, \\ u_{|\partial\Omega} & = & 0. & \end{array} \right.$$

We have the following equality:

$$\int_\Omega f(\mathbf{x}, \Omega)\, d\mathbf{x} = \int_{\partial\Omega} \nabla u \cdot \mathbf{N}\, d\mathbf{a}(\mathbf{x}),$$

where $\mathbf{N}$ is the inside pointing unit normal to $\partial\Omega$ and $d\mathbf{a}(\mathbf{x})$ its area element.

Proof. Because of our assumptions, Poisson's equation has a unique classical, i.e., C^2, solution in $\overline{\Omega}$ [2, 25], and we have

$$\int_\Omega f(\mathbf{x}, \Omega)\, d\mathbf{x} = -\int_\Omega \Delta u\, d\mathbf{x} = \int_{\partial\Omega} \nabla u \cdot \mathbf{N}\, d\mathbf{a}(\mathbf{x}),$$

the last equality being a consequence of the Green–Riemann theorem. □

A region functional can always be expressed as a boundary functional, via the solution of Poisson's equation with Dirichlet conditions.

3.2. Tranformation of boundary functionals into region functionals. The converse of Proposition 3.1 is also true. Let us consider the boundary functional (2.2).

PROPOSITION 3.2. *Let Ω be a bounded open set with regular boundary $\partial\Omega$. Let u be the unique solution of Helmholtz's equation:*

$$\left\{ \begin{array}{rcll} -\Delta u + u & = & 0 & in \quad \Omega, \\ \frac{\partial u}{\partial \mathbf{N}}{}_{|\partial\Omega} & = & -g. & \end{array} \right.$$

Then we have the following equality:

$$\int_{\partial\Omega} g(\mathbf{x})\, d\mathbf{a}(\mathbf{x}) = \int_\Omega u(\mathbf{x}, \Omega)\, d\mathbf{x},$$

where $d\mathbf{a}(\mathbf{x})$ is the area element of $\partial\Omega$.

Proof. Because of our assumptions, Helmholtz's equation has a unique classical, i.e., C^2, solution in $\overline{\Omega}$ [42, 13, 14, 15, 16, 17, 18], and we have

$$\int_\Omega u\, d\mathbf{x} = \int_\Omega \Delta u\, d\mathbf{x} = -\int_{\partial\Omega} \nabla u \cdot \mathbf{N}\, d\mathbf{a}(\mathbf{x}),$$

the last equality being a consequence of the Green–Riemann theorem. Therefore

$$\int_\Omega u\, d\mathbf{x} = -\int_{\partial\Omega} \frac{\partial u}{\partial \mathbf{N}}\, d\mathbf{a}(\mathbf{x}) = \int_{\partial\Omega} g(\mathbf{x})\, d\mathbf{a}(\mathbf{x}). \qquad \square$$

A boundary functional can always be expressed as a region functional, via the solution of Helmholtz's equation with Neumann boundary conditions.

4. Computation of the evolution equation using a boundary approach. Originally, the derivation of region functionals has been performed by using the Green–Riemann theorem to transform region functionals into boundary functionals and then by computing the Euler–Lagrange equations. In this section, we recall the principles of the derivation and we explicitly take into account the case of region-dependent features when computing the Gâteaux derivative. Region functionals are transformed into boundary functionals by using Proposition 3.2. The region functional to minimize is (2.3).

The computation of a velocity field for the evolution of the boundary in order to reach a minimum of the error criterion proceeds in three main steps:

1. Tranformation of the region functionals into boundary functionals.
2. Computation of the Gâteaux derivatives of the boundary functionals.
3. Construction of a velocity field for the evolution of the boundary.

The first step can always be performed as it has been proven in Proposition 3.1.

The computation of an optimal velocity field is carried out for region-independent features first, i.e., when the function f does not depend on Ω. We then consider the more general case where f has some region dependency. We derive our results in the two-dimensional case; the generalization to any dimension is tedious but straightforward.

4.1. Region-independent features. In this part, we detail the three steps for region-independent features. We do it for two-dimensional images ($n = 2$) to keep notation simple, but the results hold in any dimension greater than 2.

We parameterize $\partial\Omega$ through the C^2 function $\Gamma : [0,1] \to \mathbb{R}^2$ such that when p varies from 0 to 1 we go once around $\partial\Omega$ counterclockwise. The unit tangent vector to $\partial\Omega$ is the vector $\Gamma'(p)/|\Gamma'(p)|$, where $'$ indicates the derivative with respect to the parameter p. The inside pointing normal $\mathbf{N}$ is the vector $\Gamma'^{\perp}(p)/|\Gamma'(p)|$. The vector $\Gamma'^{\perp}$ is obtained by rotating Γ' by 90 degrees counterclockwise; hence if $\Gamma' = [\Gamma'_1, \Gamma'_2]^T$, $\Gamma'^{\perp} = [-\Gamma'_2, \Gamma'_1]^T$.

4.1.1. Transformation of region functionals into boundary functionals. The following proposition is a straightforward consequence of Proposition 3.1

PROPOSITION 4.1. *If f satisfies the hypotheses of Proposition* 3.1*, the functional* (2.3)*,*

$$J_r(\Omega) = \int_\Omega f(\mathbf{x})\, d\mathbf{x},$$

is equal to

$$\Phi(\Gamma) = \int_0^1 \left(u_{x_2}(\Gamma(p))\Gamma'_1(p) - u_{x_1}(\Gamma(p))\Gamma'_2(p)\right)\, dp \stackrel{def}{=} \int_0^1 \varphi(\Gamma(p), \Gamma'(p))\, dp, \tag{4.1}$$

where $\Gamma = \partial\Omega$ and u is the unique classical solution of

$$\left\{ \begin{array}{rcll} -\Delta u & = & f & in \quad \Omega, \\ u_{|\partial\Omega} & = & 0. & \end{array} \right.$$

Therefore minimizing (2.3) *with respect to Ω is equivalent to minimizing* (4.1) *with respect to Γ.*

Proof. According to Proposition 3.1, we have

$$\int_\Omega f(\mathbf{x})\, d\mathbf{x} = -\int_\Omega \Delta u\, d\mathbf{x} = \int_{\partial\Omega} \nabla u \cdot \mathbf{N}\, d\mathbf{a}(\mathbf{x}),$$

and since $d\mathbf{a}(\mathbf{x}) = |\Gamma'(p)|dp$, the result follows. ∎

4.1.2. Computation of the Gâteaux derivative. According to Proposition 4.1, minimizing (2.3) with respect to Ω is equivalent to minimizing (4.1) with respect to Γ. Thus, we have to compute the Gâteaux derivative of the functional Φ.

THEOREM 4.2. *The Gâteaux derivative in the direction γ of the functional Φ is*

$$\langle \Phi'(\Gamma), \gamma \rangle = -\int_0^1 f(\Gamma(p))\,(\Gamma'^{\perp}(p) \cdot \gamma(p))\,dp.$$

Proof. Let $\gamma : [0,1] \to \mathbf{R}^2$ be a C^2 parametrization of an arbitrary closed curve. The Gâteaux derivative of $\Phi(\Gamma)$ in the direction γ noted $\langle \Phi'(\Gamma), \gamma \rangle >$ is defined by

$$\langle \Phi'(\Gamma), \gamma \rangle = \lim_{\tau \to 0} \frac{\Phi(\Gamma + \tau\gamma) - \Phi(\Gamma)}{\tau}.$$

We have

$$\lim_{\tau \to 0} \frac{\Phi(\Gamma + \tau\gamma) - \Phi(\Gamma)}{\tau} = \int_0^1 (\varphi_\Gamma(\Gamma(p), \Gamma'(p))\gamma(p) + \varphi_{\Gamma'}(\Gamma(p), \Gamma'(p))\gamma'(p))\ dp,$$

where $\varphi_\Gamma = \frac{\partial \varphi}{\partial \Gamma}(\Gamma, \Gamma')$. Integrating by parts, we obtain the following expression for the Gâteaux derivative:

$$\langle \Phi'(\Gamma), \gamma \rangle = \int_0^1 \left(\varphi_\Gamma(\Gamma(p), \Gamma'(p)) - \frac{d}{dp}\varphi_{\Gamma'}(\Gamma(p), \Gamma'(p)) \right) \cdot \gamma(p)\,dp.$$

We then explicitly compute the derivative of φ with respect to Γ using (4.1),

$$\varphi_\Gamma = \nabla u_{x_2}(\Gamma(p))\Gamma_1'(p) - \nabla u_{x_1}(\Gamma(p))\Gamma_2'(p),$$

and with respect to Γ',

$$\varphi_{\Gamma'} = [u_{x_2}, -u_{x_1}]^T.$$

Therefore

$$\frac{d}{dp}\varphi_{\Gamma'} = [\nabla u_{x_2} \cdot \Gamma', -\nabla u_{x_1} \cdot \Gamma']^T.$$

Putting everything together we obtain

$$\varphi_\Gamma - \frac{d}{dp}\varphi_{\Gamma'} = \Delta u \Gamma'^{\perp} = -f\,\Gamma'^{\perp}$$

thanks to Proposition 4.1. ∎

The Euler–Lagrange equations associated with the Gâteaux derivative are thus given by

$$\varphi_\Gamma - \frac{d}{dp}\varphi_{\Gamma'} = -f(\Gamma(p))\Gamma'^{\perp}.$$

An interesting point to note is that the intermediary function u disappears.

4.1.3. Construction of an optimal velocity vector for the evolution of an active contour. In order to find a local extremum of the criterion (4.1), we evolve a curve using the steepest descent method, starting from an initial curve defined by the user. Thus, we obtain the following evolution equation:

$$\frac{\partial \Gamma}{\partial \tau} = f(\Gamma)\mathbf{N} \quad \text{with} \quad \Gamma(\tau = 0) = \Gamma_0. \tag{4.2}$$

This is the classical result [52, 38, 40, 51] when f has no region dependency. Let us now consider the more general case where the function f has some region dependency.

4.2. General case. Let us now derive the evolution equation in the general case. As in the previous case, we follow our three steps.

4.2.1. Transformation of the region functional into a boundary functional. In the following, to simplify the proofs and the notations, we take $m = 1$ and $l_1 = 1$ and drop the indexes. The equations for $m > 1$ and $l_i \geq 1$ are then given without proof.

Because of the form of (2.4)–(2.7), we have to go through three levels of transformations. We start with the first level and the following proposition.

PROPOSITION 4.3. *If L satisfies the assumptions of Proposition* 3.1*, the functional*

$$K(\Omega) = \int_\Omega L(\mathbf{x})\, d\mathbf{x}$$

is equal to

$$\begin{aligned}\Theta(\Gamma) &= \int_0^1 (u_{x_2}(\Gamma(p), L(\Gamma))\Gamma_1'(p) - u_{x_1}(\Gamma(p), L(\Gamma))\Gamma_2'(p))\ dp \\ &\stackrel{def}{=} \int_0^1 \theta(\Gamma(p), \Gamma'(p))\, dp,\end{aligned}$$

where $\Gamma = \partial\Omega$ and u is the unique classical solution of

$$\left\{ \begin{array}{rcll} -\Delta u & = & L & in \quad \Omega, \\ u_{|\partial\Omega} & = & 0. & \end{array} \right.$$

Proof. The proof is identical to that of Proposition 4.1. □

In the same manner, for the second level, we have the following proposition.

PROPOSITION 4.4. *If H satisfies the assumptions of Proposition* 3.1*, the functional*

$$G(\Omega) = \int_\Omega H(\mathbf{x}, K(\Omega))\, d\mathbf{x}$$

with $K(\Omega) = \int_\Omega L(\mathbf{x})d\mathbf{x}$ is equal to

$$\begin{aligned}\Psi(\Gamma) &= \int_0^1 (v_{x_2}(\Gamma(p), \Theta(\Gamma))\Gamma_1'(p) - v_{x_1}(\Gamma(p), \Theta(\Gamma))\Gamma_2'(p))\ dp \\ &\stackrel{def}{=} \int_0^1 \psi(\Gamma(p), \Gamma'(p), \Theta(\Gamma))\, dp,\end{aligned}$$

where $\Gamma = \partial\Omega$ and v is the unique classical solution of

$$\left\{ \begin{array}{rcll} -\Delta v & = & H & in \quad \Omega, \\ v_{|\partial\Omega} & = & 0. & \end{array} \right.$$

Θ is given by Proposition 4.3*.*

Proof. The proof is identical to that of Proposition 4.1. □

We finally reach the third and last level with the following proposition.

PROPOSITION 4.5. *If f satisfies the assumptions of Proposition* 3.1*, the functional*

$$J(\Omega) = \int_\Omega f(\mathbf{x}, G(\Omega))\, d\mathbf{x}\,, \tag{4.3}$$

with $G(\Omega) = \int_\Omega H(\mathbf{x}, K(\Omega))d\mathbf{x}$ *and* $K(\Omega) = \int_\Omega L(\mathbf{x})d\mathbf{x}$*, is equal to*

$$\begin{aligned}\Phi(\Gamma) &= \int_0^1 (w_{x_2}(\Gamma(p), \Psi(\Gamma))\Gamma_1'(p) - w_{x_1}(\Gamma(p), \Psi(\Gamma))\Gamma_2'(p))\, dp \\ &\stackrel{def}{=} \int_0^1 \varphi(\Gamma(p), \Gamma'(p), \Psi(\Gamma))\, dp,\end{aligned} \tag{4.4}$$

where $\Gamma = \partial\Omega$ *and* u *is the unique classical solution of*

$$\begin{cases} -\Delta w &= f \quad in \quad \Omega, \\ w_{|\partial\Omega} &= 0. \end{cases}$$

$\Psi(\Gamma)$ *is given by Proposition* 4.4. *Therefore minimizing* (4.3) *with respect to* Ω *is equivalent to minimizing* (4.4) *with respect to* Γ.

Proof. The proof is identical to that of Proposition 4.1. ∎

4.2.2. Computation of the Gâteaux derivative. According to Proposition 4.5, minimizing (4.3) with respect to Ω is equivalent to minimizing (4.4) with respect to Γ. Thus we compute the Gâteaux derivative of Φ given by (4.4).

THEOREM 4.6. *The Gâteaux derivative in the direction* γ *of the functional* Φ *defined in* (4.4) *is*

$$\begin{aligned}\langle \Phi'(\Gamma), \gamma\rangle = -\int_0^1 [&f(\Gamma(p), \Psi(\Gamma)) + AH(\Gamma(p), \Theta(\Gamma)) \\ &+ ABL(\Gamma(p))]\, q(p)\, dp,\end{aligned}$$

where

$$A = \int_\Omega f_G(x, G(\Omega))\, d\mathbf{x} \quad and \quad B = \int_\Omega H_K(x, K(\Omega))\, d\mathbf{x}$$

with $f_G = \frac{\partial f}{\partial G}$*, and* $q(p) = (\Gamma'^{\perp}(p) \cdot \gamma(p))$.

Proof. The Gâteaux derivative of $\Phi(\Gamma)$ in the direction γ denoted $\langle \Phi'(\Gamma), \gamma\rangle$ is given by

$$\langle \Phi'(\Gamma), \gamma\rangle = \lim_{\tau\to 0} \frac{\Phi(\Gamma + \tau\gamma) - \Phi(\Gamma)}{\tau}.$$

We have

$$\begin{aligned}&\lim_{\tau\to 0} \frac{\Phi(\Gamma + \tau\gamma) - \Phi(\Gamma)}{\tau} \\ &\quad = \int_0^1 (\varphi_\Gamma(\Gamma(p), \Gamma'(p), \Psi(\Gamma))\gamma(p) + \varphi_{\Gamma'}(\Gamma(p), \Gamma'(p), \Psi(\Gamma))\gamma'(p))\, dp \\ &\qquad + \int_0^1 \varphi_\Psi(\Gamma(p), \Gamma'(p), \Psi(\Gamma))\langle \Psi'(\Gamma), \gamma\rangle\, dp,\end{aligned}$$

where $\varphi_\Psi = \frac{\partial \varphi}{\partial \Psi}(\Gamma, \Gamma', \Psi)$. Integrating by parts, we obtain

$$\begin{aligned}\langle \Phi'(\Gamma), \gamma\rangle = &\int_0^1 \left[\varphi_\Gamma - \frac{d}{dp}\varphi_{\Gamma'}\right]\gamma(p)\, dp \\ &+ \int_0^1 \varphi_\Psi(\Gamma(p), \Gamma'(p), \Psi(\Gamma))\langle \Psi'(\Gamma), \gamma\rangle\, dp.\end{aligned} \tag{4.5}$$

According to Theorem 4.2, we obtain $\varphi_\Gamma - \frac{d}{dp}\varphi_{\Gamma'} = -f\,\Gamma'^{\perp}$. The Gâteaux derivative of $\Psi(\Gamma)$ in the direction γ is computed in the same manner and we find

$$\begin{aligned}\langle \Psi'(\Gamma), \gamma\rangle = &- \int_0^1 H(\Gamma(p), \Theta(\Gamma))\, q(p)\, dp \\ &+ \int_0^1 \psi_\Theta(\Gamma(p), \Gamma'(p), \Theta(\Gamma))\langle \Theta'(\Gamma), \gamma\rangle\, dp.\end{aligned}$$

According to Theorem 4.2, the Gâteaux derivative of $\Theta(\Gamma)$ in the direction γ is given by:

$$\langle \Theta'(\Gamma), \gamma\rangle = - \int_0^1 L(\Gamma(p))\, q(p)\, dp.$$

Putting all terms together in (4.5), we find the following expression for the derivative:

$$\begin{aligned}\langle \Phi'(\Gamma), \gamma\rangle = &- \int_0^1 f(\Gamma(p), \Psi(\Gamma))\, q(p)\, dp \\ &- \int_0^1 \varphi_\Psi(\Gamma(p), \Gamma'(p), \Psi(\Gamma))dp \int_0^1 H(\Gamma(p), \Theta(\Gamma))\, q(p)\, dp \\ &- \int_0^1 \varphi_\Psi(\Gamma(p), \Gamma'(p), \Psi(\Gamma))dp \int_0^1 \psi_\Theta(\Gamma(p), \Gamma'(p), \Theta(\Gamma))\, dp \int_0^1 L(\Gamma(p))\, q(p)\, dp.\end{aligned}$$

Using Propositions 4.4 and 4.5, we find that

$$\int_0^1 \varphi_\Psi(\Gamma(p), \Gamma'(p), \Psi(\Gamma))dp = \int_\Omega f_G(\mathbf{x}, G(\Omega))d\mathbf{x} \stackrel{def}{=} A.$$

Similarly, using Propositions 4.3 and 4.4, we obtain

$$\int_0^1 \psi_\Theta(\Gamma(p), \Gamma'(p), \Theta(\Gamma))dp = \int_\Omega H_K(\mathbf{x}, K(\Omega))d\mathbf{x} \stackrel{def}{=} B.$$

The equation of the derivative is obtained:

$$\langle \Phi'(\Gamma), \gamma\rangle = - \int_0^1 \left[f(\Gamma(p), \Psi(\Gamma)) + A\, H(\Gamma(p), \Theta(\Gamma)) + A\, B\, L(\Gamma(p))\right]\, q(p)\, dp. \qquad \square$$

The Euler–Lagrange equations associated with the Gâteaux derivative are given by

$$-\left[f(\Gamma(p), \Psi(\Gamma)) + AH(\Gamma(p), \Theta(\Gamma)) + ABL(\Gamma(p))\right]\Gamma'^{\perp} = 0.$$

Note again that the intermediate functions u, v, and w do not appear in this expression.

We can now state the general theorem for $m > 1$ and $l_i \geq 1$.

THEOREM 4.7. *The Gâteaux derivative in the direction γ of the functional J*

defined in (2.4) *is*

$$\langle \Phi'(\Gamma), \gamma \rangle = -\int_0^1 \left(f(\Gamma(p), G_1(\Gamma), \ldots, G_m(\Gamma)) + \sum_{i=1}^{m} A_i H_i(\Gamma(p), K_{i1}(\Gamma), \ldots, K_{il_i}(\Gamma)) + \sum_{i=1}^{m} A_i \left(\sum_{j=1}^{l_i} B_{ij} L_{ij}(\Gamma(p)) \right) \right) (\Gamma'^{\perp}(p) \cdot \gamma(p))\, dp,$$

where

$$A_i = \int_\Omega f_{G_i}(x, G_1(\Omega), \ldots, G_m(\Omega))\, d\mathbf{x}, \quad i = 1 \ldots m,$$

$$\textit{and} \quad B_{ij} = \int_\Omega H_{iK_{ij}}(x, K_{i1}(\Omega), \ldots, K_{il_i}(\Omega))\, d\mathbf{x}, \quad i = 1 \ldots m, \quad j = 1 \ldots l_i.$$

4.2.3. Construction of an optimal velocity vector for the evolution of an active contour. In the general case, according to Theorem 4.7 the steepest gradient descent method yields the following evolution equation for the active contour:

$$\frac{\partial \Gamma}{\partial \tau} = \left[f(\Gamma) + \sum_{i=1}^{m} A_i H_i(\Gamma) + \sum_{i=1}^{m} A_i \left(\sum_{j=1}^{l_i} B_{ij} L_{ij}(\Gamma) \right) \right] \mathbf{N} \tag{4.6}$$

with $\Gamma(\tau = 0) = \Gamma_0$. Compared with (4.2), some additional terms appear that come from the region dependency of the descriptors.

5. Computation of the derivative using shape derivation tools, or "how to keep a region formulation." In the previous part, region functionals were first transformed into boundary functionals for the computation of the derivative. This step is neither natural nor straightforward. Therefore, we propose another solution based on shape derivation tools [49, 22]. The region formulation is maintained for the computation and this provides a suitable way of obtaining the derivative of the criterion and the evolution equation of an active contour.

We perform three main steps:

1. Introduction of a dynamic scheme. Since the set of all image regions is not a vector space, it is difficult to compute the derivative of the criterion with respect to the domain Ω. To circumvent this problem, we apply a family of transformations T_τ, indexed by a real parameter $\tau \geq 0$, to Ω, and we denote $\Omega(\tau) = T_\tau(\Omega)$. The region functional becomes a function of τ, $J(\tau) \stackrel{def}{=} J(\Omega(\tau))$.
2. Derivation of the criterion based on shape derivation principles. The error criterion $J(\tau)$ is then derived with respect to τ using shape derivation principles.
3. Computation of the evolution equation from the derivative. From the derivative, we deduce the velocity field of the active contour that will make it evolve towards a local minimum of the error criterion.

5.1. Introduction of transformations. As it has already been pointed out, the optimization of the region functional (2.3) is difficult since the set of regular domains (regular open bounded sets) $\mathcal{U}$ of $\mathbb{R}^n$ does not have the structure of a vector space. Variations of a domain must then be defined in some way. Let us consider a reference domain $\Omega \in \mathcal{U}$ and the set $\hat{\mathcal{A}}$ of applications $T : \Omega \to \mathbb{R}^n$, which are at least as regular as homeomorphisms (i.e., one to one with T and T^{-1} continuous). We define

$$\hat{\mathcal{A}} = \left\{T \text{ one to one}, T, T^{-1} \in W^{1,\infty}(\Omega, \mathbb{R}^n)\right\},$$

where

$$W^{1,\infty}(\Omega, \mathbb{R}^n) = \{T : \Omega \to \mathbb{R}^n \text{ such that } T \in L^\infty(\Omega, \mathbb{R}^n) \text{ and } \partial_i T \in L^\infty(\Omega, \mathbb{R}^n),\ i = 1, \ldots, n\}.$$

Given a shape function $F : \mathcal{U} \to \mathbf{R}^+$ for $T \in \hat{\mathcal{A}}$, let us define $\hat{F}(T) = F(T(\Omega))$. The key point is that $W^{1,\infty}(\Omega, \mathbb{R}^n)$ is a Banach space. This allows us to define the notion of derivative with respect to the domain Ω as follows.

DEFINITION 5.1. *F is Gâteaux differentiable with respect to Ω if and only if $\hat{F}$ is Gâteaux differentiable with respect to T.*

In order to compute Gâteaux derivatives with respect to T we introduce a family of deformation $(T(\tau))_{\tau \geq 0}$ such that $T(\tau) \in \hat{\mathcal{A}}$ for $\tau \geq 0$, $T(0) = \mathrm{Id}$, and $T(.) \in C^1([0, A]; W^{1,\infty}(\Omega, \mathbb{R}^n))$, $A > 0$. From a practical point of view, there are many ways to construct such a family. The most famous one is the Hadamard deformation [27].

For a point $\mathbf{x} \in \Omega$, we denote

$$\mathbf{x}(\tau) = T(\tau, \mathbf{x}) \quad \text{with} \quad T(0, \mathbf{x}) = \mathbf{x},$$
$$\Omega(\tau) = T(\tau, \Omega) \quad \text{with} \quad T(0, \Omega) = \Omega.$$

Let us now define the velocity vector field $\mathbf{V}$ corresponding to $T(\tau)$ as

$$\mathbf{V}(\tau, \mathbf{x}) = \frac{\partial T}{\partial \tau}(\tau, \mathbf{x}) \quad \forall \mathbf{x} \in \Omega \ \ \forall \tau \geq 0.$$

5.2. Computation of the derivative using shape derivation tools. We now introduce three main definitions.

DEFINITION 5.2. *The Gâteaux derivative of $J(\Omega) = \int_\Omega f(\mathbf{x}, \Omega) d\mathbf{x}$ in the direction of $\mathbf{V}$, denoted $\langle J'(\Omega), \mathbf{V}\rangle$, is equal to*

$$\langle J'(\Omega), \mathbf{V}\rangle = \lim_{\tau \to 0} \frac{J(\Omega(\tau)) - J(\Omega)}{\tau}.$$

DEFINITION 5.3. *The material derivative of $f(\mathbf{x}, \Omega)$, denoted $f_m(\mathbf{x}, \Omega, \mathbf{V})$, is equal to*

$$f_m(\mathbf{x}, \Omega, \mathbf{V}) = \lim_{\tau \to 0} \frac{f(\mathbf{x}(\tau), \Omega(\tau)) - f(\mathbf{x}, \Omega)}{\tau}.$$

DEFINITION 5.4. *The shape derivative of $f(\mathbf{x}, \Omega)$, denoted $f_s(\mathbf{x}, \Omega, \boldsymbol{V})$, is equal to*

$$f_s(\mathbf{x}, \Omega, \mathbf{V}) = \lim_{\tau \to 0} \frac{f(\mathbf{x}, \Omega(\tau)) - f(\mathbf{x}, \Omega)}{\tau}.$$

5.2.1. Relation between the Gâteaux derivative and the shape derivative. The following theorem gives a relation between the Gâteaux derivative and the shape derivative for the region functional (2.3). The proof can be found in [49, 22], we provide an elementary one here for completeness.

THEOREM 5.5. *The Gâteaux derivative of the functional* $J(\Omega) = \int_\Omega f(\mathbf{x}, \Omega)\, d\mathbf{x}$ *in the direction of* $\mathbf{V}$ *is the following:*

$$\langle J'(\Omega), \mathbf{V}\rangle = \int_\Omega f_s(\mathbf{x}, \Omega, \mathbf{V}) d\mathbf{x} - \int_{\partial\Omega} f(\mathbf{x}, \Omega)(\mathbf{V}(\mathbf{x}) \cdot \boldsymbol{N}(\mathbf{x})) d\mathbf{a}(\mathbf{x}),$$

where $\boldsymbol{N}$ *is the unit inward normal to* $\partial\Omega$ *and* $d\mathbf{a}$ *its area element.*

Proof. As far as the computation of the derivative is concerned, only small deformations are relevant, and we thus consider a first order Taylor expansion of the transformation:

$$\begin{aligned} T(\tau, \mathbf{x}) &= T(0, \mathbf{x}) + \tau \frac{\partial T}{\partial \tau}(0, \mathbf{x}) \\ &= \mathbf{x} + \tau \mathbf{V}(\mathbf{x}), \end{aligned}$$

where $\mathbf{V}(\mathbf{x}) = \frac{\partial T}{\partial \tau}(0, \mathbf{x})$.

We obtain the following expressions for the material and the shape derivatives:

$$\begin{aligned} f_m(\mathbf{x}, \Omega, \mathbf{V}) &= \lim_{\tau \to 0} \frac{f(\mathbf{x} + \tau \mathbf{V}(\mathbf{x}), \Omega + \tau \mathbf{V}) - f(\mathbf{x}, \Omega)}{\tau}, \\ f_s(\mathbf{x}, \Omega, \mathbf{V}) &= \lim_{\tau \to 0} \frac{f(\mathbf{x}, \Omega + \tau \mathbf{V}) - f(\mathbf{x}, \Omega)}{\tau}. \end{aligned}$$

If we assume that $\lim_{\tau \to 0} \nabla f(\mathbf{x}, \Omega + \tau \mathbf{V}) = \nabla f(\mathbf{x}, \Omega)$, we can write

$$f_m(\mathbf{x}, \Omega, \mathbf{V}) = f_s(\mathbf{x}, \Omega, \mathbf{V}) + \nabla f(\mathbf{x}, \Omega) \cdot \mathbf{V}(\mathbf{x}). \tag{5.1}$$

We are now ready for the computation of the Gâteaux derivative of $J(\Omega)$ in the direction of $\mathbf{V}$. We have

$$\frac{J(\Omega(\tau)) - J(\Omega)}{\tau} = \frac{1}{\tau} \left[\int_{\Omega(\tau)} f(\mathbf{x}, \Omega(\tau)) d\mathbf{x} - \int_\Omega f(\mathbf{x}, \Omega) d\mathbf{x} \right]. \tag{5.2}$$

In the first integral, we make the change of variable $\mathbf{x} \to \mathbf{x} + \tau \mathbf{V}(\mathbf{x})$ which yields

$$\int_{\Omega(\tau)} f(\mathbf{x}, \Omega(\tau)) d\mathbf{x} = \int_\Omega f(\mathbf{x} + \tau \mathbf{V}(\mathbf{x}), \Omega + \tau \mathbf{V})| \det J_\tau(\mathbf{x})| d\mathbf{x},$$

where $J_\tau(\mathbf{x})$ is the Jacobian matrix,

$$J_\tau(\mathbf{x}) = \begin{pmatrix} 1 + \tau \frac{\partial V_1}{\partial x_1} & \cdots & \tau \frac{\partial V_1}{\partial x_n} \\ \vdots & \vdots & \vdots \\ \tau \frac{\partial V_n}{\partial x_1} & \cdots & 1 + \tau \frac{\partial V_n}{\partial x_n} \end{pmatrix},$$

$\mathbf{V}(\mathbf{x}) = [V_1(\mathbf{x}), \ldots, V_n(\mathbf{x})]^T$, and $\mathbf{x} = [x_1, \ldots, x_n]^T$. It follows that

$$\det J_\tau(\mathbf{x}) = 1 + \tau \mathrm{div}(\mathbf{V}(\mathbf{x})) + o(\tau).$$

This shows that if τ is small enough, $\det J_\tau(\mathbf{x}) > 0$ and

$$\lim_{\tau\to 0} \frac{\det J_\tau(\mathbf{x}) - 1}{\tau} = \operatorname{div}(\mathbf{V}(\mathbf{x})).$$

Equation (5.2) can now be rewritten as

$$\begin{aligned}\frac{J(\Omega(\tau)) - J(\Omega)}{\tau} &= \int_\Omega \frac{f(\mathbf{x}+\tau\mathbf{V}(\mathbf{x}), \Omega+\tau\mathbf{V}) - f(\mathbf{x},\Omega)}{\tau} \det(J_\tau(\mathbf{x}))d\mathbf{x} \\ &\quad - \int_\Omega f(\mathbf{x},\Omega)\ \frac{\det(J_\tau(\mathbf{x})) - 1}{\tau}\ d\mathbf{x} \stackrel{def}{=} I_1 - I_2.\end{aligned}$$

If τ goes to 0, using (5.1) and Definitions 5.3 and 5.4, we get

$$\begin{aligned}\lim_{\tau\to 0} I_1 &= \int_\Omega f_m(\mathbf{x},\Omega,\mathbf{V})d\mathbf{x} \\ &= \int_\Omega f_s(\mathbf{x},\Omega,\mathbf{V})d\mathbf{x} + \int_\Omega \nabla f(\mathbf{x},\Omega)\cdot\mathbf{V}(\mathbf{x})d\mathbf{x}, \\ \lim_{\tau\to 0} I_2 &= \int_\Omega f(\mathbf{x},\Omega)\operatorname{div}(\mathbf{V}(\mathbf{x}))d\mathbf{x}.\end{aligned}$$

We find that the Gâteaux derivative is given by

$$\begin{aligned}\langle J'(\Omega), \mathbf{V}\rangle &= \int_\Omega f_s(\mathbf{x},\Omega,\mathbf{V})d\mathbf{x} + \int_\Omega (\nabla f(\mathbf{x},\Omega)\cdot\mathbf{V}(\mathbf{x}) + f(\mathbf{x},\Omega)\operatorname{div}(\mathbf{V}(\mathbf{x})))d\mathbf{x} \\ &= \int_\Omega f_s(\mathbf{x},\Omega,\mathbf{V})d\mathbf{x} + \int_\Omega \operatorname{div}(f(\mathbf{x},\Omega)\ \mathbf{V}(\mathbf{x}))d\mathbf{x}.\end{aligned} \tag{5.3}$$

Applying the Green–Riemann theorem in (5.3), we finally obtain

$$\langle J'(\Omega), \mathbf{V}\rangle = \int_\Omega f_s(\mathbf{x},\Omega,\mathbf{V})d\mathbf{x} - \int_{\partial\Omega} f(\mathbf{x},\Omega)(\mathbf{V}(\mathbf{x})\cdot\mathbf{N}(\mathbf{x}))d\mathbf{a}(\mathbf{x}),$$

where $\mathbf{N}$ is the unit inward normal to $\partial\Omega$. □

Note that Theorem 5.5 provides a necessary condition for a domain $\hat{\Omega}$ to be an extremum of $J(\Omega)$:

$$\int_{\hat{\Omega}} f_s(\mathbf{x},\hat{\Omega},\mathbf{V})d\mathbf{x} - \int_{\partial\hat{\Omega}} f(\mathbf{x},\hat{\Omega})(\mathbf{V}(\mathbf{x})\cdot\mathbf{N}(\mathbf{x}))\, d\mathbf{a}(\mathbf{x}) = 0 \quad \forall\mathbf{V}.$$

5.3. Construction of the velocity vector of the active contour from the Gâteaux derivative. We now make good use of the previous tools to derive the velocity vector of the active contour for the same functionals as those which were considered in section 5. As expected we find the same results but in a way which, we feel, is more natural, since we do not have to turn a region integral into a boundary one, and simpler. The evolving region boundary $\partial\Omega$, denoted Γ, is modeled as an active contour: the user defines an initial curve $\Gamma_0 = \partial\Omega_0$ that evolves towards a local minimum of the region functional (2.3) according to a PDE that we will now derive.

5.3.1. Region-independent features. We first consider the simple case where the function f does not depend on Ω, i.e., $f = f(\mathbf{x})$. In that case, the shape derivative f_s is equal to zero and the Gâteaux derivative of J is simply (Theorem 5.5)

$$\langle J'(\Omega), \mathbf{V}\rangle = -\int_{\partial\Omega} f(\mathbf{x})(\mathbf{V}(\mathbf{x})\cdot\boldsymbol{N}(\mathbf{x}))d\mathbf{a}(\mathbf{x}).$$

This leads to the following evolution equation for region-independent descriptors:

$$\frac{\partial \Gamma}{\partial \tau} = f(\Gamma)\mathbf{N}$$

with $\Gamma(\tau = 0) = \Gamma_0$.

We notice that, as expected, the evolution equation is the same as (4.2) in section 4.

5.3.2. General case. Let us now tackle the same general case as in section 4.2, using the functional defined by (2.4)–(2.7). We similarly restrict the computation of the Gâteaux derivative of J to the case $m = 1$ and $l_i = 1$, state the result for $m > 1$ and $l_i \geq 1$, and drop the indexes.

THEOREM 5.6. *The Gâteaux derivative in the direction of* $\mathbf{V}$ *of the functional* J *defined in* (4.3) *is*

$$\langle J'(\Omega), \mathbf{V} \rangle = -\int_\Gamma \left(A\, B\, L(\mathbf{x}) + A\, H(\mathbf{x}, K(\Omega)) + f(\mathbf{x}, \Omega)\right) (\mathbf{V}(\mathbf{x}) \cdot \mathbf{N}(\mathbf{x})) d\mathbf{a}(\mathbf{x}),$$

where

$$A = \int_\Omega f_G(x, G(\Omega))\, d\mathbf{x} \quad \textit{and} \quad B = \int_\Omega H_K(x, K(\Omega))\, d\mathbf{x}.$$

Proof. According to Theorem 5.5, we have

$$\langle J'(\Omega), \mathbf{V} \rangle = \int_\Omega f_s \; d\mathbf{x} - \int_\Gamma f \; (\mathbf{V} \cdot \mathbf{N}) d\mathbf{a}(\mathbf{x}).$$

Let us first compute the shape derivative of f. From the chain rule we get

$$f_s(\mathbf{x}, \Omega, \mathbf{V}) = f_G(\mathbf{x}, G) \langle G'(\Omega), \mathbf{V} \rangle, \tag{5.4}$$

where f_G denotes the partial derivative of the function f with respect to its second argument.

Next we compute the Gâteaux derivative of G in the direction of $\mathbf{V}$. We again apply Theorem 5.5, and we get

$$\langle G'(\Omega), \mathbf{V} \rangle = \int_\Omega H_s \; d\mathbf{x} - \int_\Gamma H \; (\mathbf{V} \cdot \mathbf{N}) d\mathbf{a}(\mathbf{x}).$$

Plugging this into (5.4), we obtain

$$\int_\Omega f_s \, d\mathbf{x} = A \left(\int_\Omega H_s \; d\mathbf{x} - \int_\Gamma H (\mathbf{V} \cdot \mathbf{N}) d\mathbf{a}(\mathbf{x}) \right),$$

where

$$A = \int_\Omega f_G(\mathbf{x}, G(\Omega)) \; d\mathbf{x}.$$

We also compute the shape derivative of H through Theorem 5.5:

$$H_s(\mathbf{x}, \Omega, \mathbf{V}) = H_K(\mathbf{x}, K) \langle K'(\Omega), \mathbf{V} \rangle.$$

We continue with the Gâteaux derivative of K in the direction of $\mathbf{V}$:

$$\langle K'(\Omega), \mathbf{V}\rangle = \int_\Omega L_s \; d\mathbf{x} - \int_\Gamma L(\mathbf{x})(V(\mathbf{x}) \cdot \mathbf{N}(\mathbf{x}))d\mathbf{a}(\mathbf{x}).$$

Since L does not depend on Ω, we obtain $L_s = 0$ and we are done.

Putting all terms together, we obtain the complete expression of the Gâteaux derivative of J:

$$\langle J'(\Omega), \mathbf{V}\rangle = -\int_\Gamma \left(A\,B\,L(\mathbf{x}) + A\,H(\mathbf{x}, K(\Omega)) + f(\mathbf{x}, \Omega)\right)\,(\mathbf{V}(\mathbf{x}) \cdot \mathbf{N}(\mathbf{x}))d\mathbf{a}(\mathbf{x})$$

with $B = \int_\Omega H_K(\mathbf{x}, K)\, d\mathbf{x}$. □

The general case follows easily and is stated in the following theorem.

THEOREM 5.7. *The Gâteaux derivative in the direction of* $\mathbf{V}$ *of the functional* J *defined in* (2.4) *is*

$$\langle J'(\Omega), \mathbf{V}\rangle = -\int_\Gamma \left(\sum_{i=1}^{m} A_i \sum_{j=1}^{l_i}(B_{ij}\, L_{ij}(\mathbf{x})) + \sum_{i=1}^{m}(A_i\, H_i) + f\right)\,(\mathbf{V}(\mathbf{x})) \cdot \mathbf{N}(\mathbf{x})d\mathbf{a}(\mathbf{x}),$$

where

$$A_i = \int_\Omega f_{G_i}(x, G_1(\Omega), \ldots, G_m(\Omega))\, d\mathbf{x}, \quad i = 1 \ldots m,$$

$$\textit{and} \quad B_{ij} = \int_\Omega H_{iK_{ij}}(x, K_{i1}(\Omega), \ldots, K_{il_i}(\Omega))\, d\mathbf{x}, \quad i = 1 \ldots m, \quad j = 1 \ldots l_i.$$

From the Gâteaux derivative of J, we deduce the corresponding evolution equation:

$$\frac{\partial \Gamma}{\partial \tau} = \left[f(\Gamma) + \sum_{i=1}^{m} A_i H_i(\Gamma) + \sum_{i=1}^{m} A_i \left(\sum_{j=1}^{l_i} B_{ij} L_{ij}(\Gamma)\right)\right] \mathbf{N}, \tag{5.5}$$

which, as expected, is identical to (4.6). As far as the final result is concerned, the two methods of computation are equivalent.

5.4. Application. Let us now apply this method to the first example in section 2.3. The function f is given by (2.8). The corresponding functions G_i, H_i are given in section 2.3. We need the terms $A_j, j = 1, 2$:

$$\begin{cases} A_1 &= -\int_\Omega \frac{1}{G_2}\; \varrho'\left(I(\mathbf{x}) - \frac{G_1}{G_2}\right)\, d\mathbf{x} = \frac{-1}{|\Omega|}\int_\Omega \varrho'(I - \mu_\Omega)d\mathbf{x}, \\ A_2 &= \int_\Omega \frac{G_1}{(G_2)^2}\; \varrho'\left(I(\mathbf{x}) - \frac{G_1}{G_2}\right)\, d\mathbf{x} = \frac{\mu_\Omega}{|\Omega|}\int_\Omega \varrho'(I - \mu_\Omega)d\mathbf{x}. \end{cases}$$

Since the terms H_i are not region-dependent, the terms B_{ij} are equal to zero. The velocity vector of the active contour is then the following:

$$\frac{\partial \Gamma(\tau)}{\partial \tau} = \left[\; f - \frac{(I - \mu_\Omega)}{|\Omega|}\int_\Omega \varrho'(I - \mu_\Omega)d\mathbf{x}\;\right]\; \mathbf{N}.$$

In this example, the additional term coming from the region dependency of f is equal to $\frac{(I-\mu_\Omega)}{|\Omega|}\int_\Omega \varrho'(I-\mu_\Omega)d\mathbf{x}$. Note that in the particular case of $\varrho(r) = r^2$, this additional

term is equal to zero [6, 7, 20, 21]. However, in the general case, the additional term is not nul.

Let us apply this method to the second example in section 2.4. The function f is a function of the variance given by (2.9). The corresponding functions G_i, H_i, K_{ij}, and L_{ij} are also given in section 2.4. We need the terms $A_j, j = 1, 2$:

$$\left\{ \begin{array}{rcl} A_1 & = & \int_\Omega \frac{1}{G_2}\ \varrho'\left(\frac{G_1}{G_2}\right) d\mathbf{x} = \varrho'(\sigma^2_\Omega), \\ A_2 & = & -\int_\Omega \frac{G_1}{(G_2)^2}\ \varrho'\left(\frac{G_1}{G_2}\right) d\mathbf{x} = -\sigma^2_\Omega\ \varrho'(\sigma^2_\Omega). \end{array} \right.$$

The terms B_{ij} are given by

$$\left\{ \begin{array}{l} B_{11} = \int_\Omega H_{1K_{11}}(x, K_{11}, K_{12}) = -2\frac{1}{|\Omega|}\int_\Omega (I(\mathbf{x}) - \mu_\Omega) d\mathbf{x} = 0, \\ \\ B_{12} = \int_\Omega H_{1K_{12}}(x, K_{11}, K_{12}) = 2\frac{\mu_\Omega}{|\Omega|}\int_\Omega (I(\mathbf{x}) - \mu_\Omega) d\mathbf{x} = 0. \end{array} \right.$$

We can then compute the velocity vector of the active contour from (5.5) and we find

$$\frac{\partial \Gamma(\tau)}{\partial \tau} = [\ f + \varrho'(\sigma^2_\Omega)\left((I - \mu_\Omega)^2 - \sigma^2_\Omega\right)\ ]\ \mathbf{N}.$$

In this simple example, we notice that the dependency of the function on the region induces an additional term in the evolution equation compared with the evolution equation obtained in the case where the function is region independent (equation (4.2)). This additional term is $\varrho'(\sigma^2_\Omega)\left((I(\mathbf{x}) - \mu_\Omega)^2 - \sigma^2_\Omega\right)$. It must be included in order to reach a true minimum of the criterion as proved in [33].

6. Matching histograms. A natural way of generalizing the use of statistical image features such as the mean and the variance of the intensity for image segmentation is to consider the full probability distribution of the feature of interest within the region, e.g., intensity, color, texture, etc., ... It turns out that in attempting to do so, one is naturally led to extend the criterion (2.4) to the case where the function f depends on a *continuous* family of region criteria. Consider a function $\mathbf{h} : \mathbb{R}^n \to \mathbb{R}^m$ which describes the feature of interest. Suppose we have learnt the probability density function (pdf) of the feature $\mathbf{h}$ within the image region of interest, and let $q(\boldsymbol{\alpha})$ be this pdf. Given a region Ω, we can estimate the pdf of the feature $\mathbf{h}$ through the use of the Parzen method [24]: let $p : \mathbb{R}^m \to \mathbb{R}^+$ be the Parzen window, a smooth positive function whose integral is equal to 1. For the sake of simplicity but without loss of generality, we assume that p is an m-dimensional Gaussian with 0-mean and variance σ^2, we note

$$p(\boldsymbol{\alpha}) = g_\sigma(\boldsymbol{\alpha}) = \frac{1}{(2\pi\sigma^2)^{m/2}} \exp\left(-\frac{|\boldsymbol{\alpha}|^2}{2\sigma^2}\right),$$

and we define

$$\hat{q}(\boldsymbol{\alpha}, \Omega) = \frac{1}{K(\Omega)} \int_\Omega g_\sigma(\mathbf{h}(\mathbf{x}) - \boldsymbol{\alpha})\, d\mathbf{x},$$

where $\mathbf{h}(\mathbf{x})$ is the value of the feature of interest at the point $\mathbf{x}$ of Ω and K is a normalizing constant, in general depending of Ω, such that

$$\int_{\mathbb{R}^m} \hat{q}(\boldsymbol{\alpha}, \Omega)\, d\boldsymbol{\alpha} = 1.$$

Therefore

$$K(\Omega) = \int_{\Omega} \int_{\mathbb{R}^m} g_\sigma(\mathbf{h}(\mathbf{x}) - \boldsymbol{\alpha})\, d\boldsymbol{\alpha}\, d\mathbf{x} = \mid \Omega \mid .$$

We next assume that we have a function $\varphi : \mathbb{R}^+ \times \mathbb{R}^m \to \mathbb{R}^+$ which allows us to compare two pdfs. This function is small if the pdfs are similar and large otherwise. It allows us to introduce the following functional which represents the "distance" between the two histograms:

$$D(\Omega) = \int_{\mathbb{R}^m} \varphi(\hat{q}(\boldsymbol{\alpha}, \Omega), q(\boldsymbol{\alpha}))\, d\boldsymbol{\alpha}. \tag{6.1}$$

The distance can be the square of the L^2 norm when

$$\varphi(\hat{q}(\boldsymbol{\alpha}, \Omega), q(\boldsymbol{\alpha})) = (\hat{q}(\boldsymbol{\alpha}, \Omega) - q(\boldsymbol{\alpha}))^2,$$

or the commonly used Kullback–Leibler divergence when

$$\varphi(\hat{q}(\boldsymbol{\alpha}, \Omega), q(\boldsymbol{\alpha})) = \frac{1}{2}\left(q(\boldsymbol{\alpha}) \log \frac{q(\boldsymbol{\alpha})}{\hat{q}(\boldsymbol{\alpha}, \Omega)} + \hat{q}(\boldsymbol{\alpha}, \Omega) \log \frac{\hat{q}(\boldsymbol{\alpha}, \Omega)}{q(\boldsymbol{\alpha})} \right),$$

or the Hellinger distance when

$$\varphi(\hat{q}(\boldsymbol{\alpha}, \Omega), q(\boldsymbol{\alpha})) = (\sqrt{\hat{q}(\boldsymbol{\alpha}, \Omega)} - \sqrt{q(\boldsymbol{\alpha})})^2 ,$$

or the nonsymmetric chi-2 comparison function when

$$\varphi(\hat{q}(\boldsymbol{\alpha}, \Omega), q(\boldsymbol{\alpha})) = \frac{(\hat{q}(\boldsymbol{\alpha}, \Omega) - q(\boldsymbol{\alpha}))^2}{q(\boldsymbol{\alpha})}.$$

A further generalization of the previous case is to consider second order histograms which describe the probability of having the value $\boldsymbol{\alpha}_1$ at pixel $\mathbf{x}$ and the value $\boldsymbol{\alpha}_2$ at pixel $\mathbf{x} + \boldsymbol{\delta}$, where $\boldsymbol{\delta}$ is a fixed (usually small) vector of $\mathbf{R}^n$. This has been used very much in computer vision for analyzing textures [28, 29]. The corresponding pdf, denoted $q_{\boldsymbol{\delta}}(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)$, can be estimated with the same Parzen window technique. We define

$$\hat{q}_{\boldsymbol{\delta}}(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \Omega) = \frac{1}{K_{\boldsymbol{\delta}}(\Omega)} \int_{\Omega} g_\sigma(\mathbf{h}(\mathbf{x}) - \boldsymbol{\alpha}_1) g_\sigma(\mathbf{h}(\mathbf{x} + \boldsymbol{\delta}) - \boldsymbol{\alpha}_2)\, d\mathbf{x}.$$

The normalizing constant $K_{\boldsymbol{\delta}}(\Omega)$ is given by

$$K_{\boldsymbol{\delta}}(\Omega) = \int_{\Omega} \int_{\mathbb{R}^m \times \mathbb{R}^m} g_\sigma(\mathbf{h}(\mathbf{x}) - \boldsymbol{\alpha}_1) g_\sigma(\mathbf{h}(\mathbf{x} + \boldsymbol{\delta}) - \boldsymbol{\alpha}_2)\, d\boldsymbol{\alpha}_1\, d\boldsymbol{\alpha}_2\, d\mathbf{x} = \mid \Omega \mid .$$

We therefore define

$$D_{\boldsymbol{\delta}}(\Omega) = \int_{\mathbb{R}^m \times \mathbb{R}^m} \varphi(\hat{q}_{\boldsymbol{\delta}}(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \Omega), q_{\boldsymbol{\delta}}(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2))\, d\boldsymbol{\alpha}_1\, d\boldsymbol{\alpha}_2. \tag{6.2}$$

Using the tools developed in section 5, we compute the Gâteaux derivative of the functional D. We have the following theorem.

THEOREM 6.1. *The Gâteaux derivative in the direction* $\mathbf{V}$ *of the functional* D *defined in* (6.1) *is*

$$\langle D'(\Omega), \mathbf{V}\rangle = -\frac{1}{\mid \Omega \mid} \int_\Gamma \Big(\partial_1\varphi(\hat{q}(.,\Omega), q(.)) * g_\sigma(\mathbf{h}(\mathbf{x})) - C(\Omega)\Big)(\mathbf{V}\cdot\mathbf{N})d\mathbf{a}(\mathbf{x}),$$

where $C(\Omega) = \int_{\mathbb{R}^m} \partial_1\varphi(\hat{q}(\boldsymbol{\alpha},\Omega), q(\boldsymbol{\alpha}))\hat{q}(\boldsymbol{\alpha},\Omega)\, d\boldsymbol{\alpha}$.

Proof. By the definition of D we have

$$\langle D'(\Omega), \mathbf{V}\rangle = \int_{\mathbb{R}^m} \langle (\varphi(\hat{q}(\boldsymbol{\alpha},\Omega), q(\boldsymbol{\alpha})))', \mathbf{V}\rangle\, d\boldsymbol{\alpha}.$$

Let us compute the Gâteaux derivative of $\varphi(\hat{q}(\boldsymbol{\alpha},\Omega), q(\boldsymbol{\alpha}))$. We define

$$\varphi(\hat{q}(\boldsymbol{\alpha},\Omega), q(\boldsymbol{\alpha})) = f(G_1(\boldsymbol{\alpha},\Omega), G_2(\Omega)) = \varphi\left(\frac{G_1(\boldsymbol{\alpha},\Omega)}{G_2(\Omega)}, q(\boldsymbol{\alpha})\right),$$

where

$$G_1(\boldsymbol{\alpha},\Omega) = \int_\Omega g_\sigma(\mathbf{h}(\mathbf{x}) - \boldsymbol{\alpha})\, d\mathbf{x} \quad \text{with} \quad H_1(\boldsymbol{\alpha},\mathbf{x}) = g_\sigma(\mathbf{h}(\mathbf{x}) - \boldsymbol{\alpha}),$$
$$G_2(\Omega) = \mid \Omega \mid = \int_\Omega d\mathbf{x}.$$

We obtain

$$\begin{aligned}\langle f', \mathbf{V}\rangle &= f_{G_1}\langle G_1', \mathbf{V}\rangle + f_{G_2}\langle G_2', \mathbf{V}\rangle \\ &= \frac{\partial_1\varphi(\hat{q}(\boldsymbol{\alpha},\Omega), q(\boldsymbol{\alpha}))}{\mid \Omega \mid}\left(\langle G_1', \mathbf{V}\rangle - \hat{q}(\boldsymbol{\alpha},\Omega)\langle G_2', \mathbf{V}\rangle\right)\end{aligned}$$

and, using Theorem 5.5,

$$\langle f', \mathbf{V}\rangle = -\frac{\partial_1\varphi(\hat{q}(\boldsymbol{\alpha},\Omega), q(\boldsymbol{\alpha}))}{\mid \Omega \mid} \int_\Gamma \left(\, g_\sigma(\mathbf{h}(\mathbf{x}) - \boldsymbol{\alpha}) - \hat{q}(\boldsymbol{\alpha},\Omega)\right)(\mathbf{V}\cdot\mathbf{N})d\mathbf{a}(\mathbf{x}).$$

Plugging this result into the expression of $\langle D'(\Omega), \mathbf{V}\rangle$ and swapping the order of integration, we obtain

$$\begin{aligned}\langle D'(\Omega), \mathbf{V}\rangle = -\frac{1}{\mid \Omega \mid} \int_\Gamma \Big(\int_{\mathbb{R}^m} & g_\sigma(\mathbf{h}(\mathbf{x}) - \boldsymbol{\alpha})\partial_1\varphi(\hat{q}(\boldsymbol{\alpha},\Omega), q(\boldsymbol{\alpha}))\, d\boldsymbol{\alpha} \\ & - \int_{\mathbb{R}^m} \partial_1\varphi(\hat{q}(\boldsymbol{\alpha},\Omega), q(\boldsymbol{\alpha}))\hat{q}(\boldsymbol{\alpha},\Omega)\, d\boldsymbol{\alpha}\Big)(\mathbf{V}\cdot\mathbf{N})d\mathbf{a}(\mathbf{x}).\end{aligned}$$

The first integral on the right-hand side is the convolution $\partial_1\varphi(\hat{q}(.,\Omega), q(.)) * g_\sigma$ of the function $\partial_1\varphi(\hat{q}(.,\Omega), q(.)) : \mathbb{R}^m \to \mathbb{R}$ with the function g_σ. The final result is

$$\langle D'(\Omega), \mathbf{V}\rangle \;=\; -\frac{1}{\mid \Omega \mid} \int_\Gamma \Big(\partial_1\varphi(\hat{q}(.,\Omega), q(.)) \,*\, g_\sigma(\mathbf{h}(\mathbf{x})) \,-\, C(\Omega)\Big)(\mathbf{V} \,\cdot\, \mathbf{N})d\mathbf{a}(\mathbf{x}),$$

where $C(\Omega) = \int_{\mathbb{R}^m} \partial_1\varphi(\hat{q}(\boldsymbol{\alpha},\Omega), q(\boldsymbol{\alpha}))\hat{q}(\boldsymbol{\alpha},\Omega)\, d\boldsymbol{\alpha}$. ∎

This solves the question of first order histograms. For second order histograms we have the following theorem.

THEOREM 6.2. *The Gâteaux derivative in the direction* $\mathbf{V}$ *of the functional* $D_{\boldsymbol{\delta}}$ *defined in* (6.2) *is*

$$\langle D'_{\boldsymbol{\delta}}(\Omega), \mathbf{V}\rangle = -\frac{1}{\mid \Omega \mid} \int_\Gamma \Big(\partial_1 \varphi(\hat{q}_{\boldsymbol{\delta}}(.,.,\Omega), q(.,.)) * (g_\sigma \otimes g_\sigma)(\mathbf{h}(\mathbf{x}), \mathbf{h}(\mathbf{x}+\boldsymbol{\delta})) - C_{\boldsymbol{\delta}}(\Omega)\Big)(\mathbf{V}\cdot\mathbf{N}) d\mathbf{a}(\mathbf{x}),$$

where $C_{\boldsymbol{\delta}}(\Omega) = \int_{\mathbb{R}^m \times \mathbb{R}^m} \partial_1 \varphi(\hat{q}_{\boldsymbol{\delta}}(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \Omega), q(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)) \hat{q}_{\boldsymbol{\delta}}(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \Omega)\, d\boldsymbol{\alpha}_1\, d\boldsymbol{\alpha}_2$, *and* $g_\sigma \otimes g_\sigma(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2) = g_\sigma(\boldsymbol{\alpha}_1)\, g_\sigma(\boldsymbol{\alpha}_2)$.

Proof. The proof is identical to that of Theorem 6.1. □

7. Color histograms: Segmentation of regions in video sequences. This work has been motivated by [12, 8] where the tracking algorithms take advantage of statistical color distributions. We propose here to use active contours in order to fit exactly the shape of the object to be segmented. We consider a video sequence where each frame is represented by the color function $\mathbf{h} : \mathbb{R}^2 \to \mathbb{R}^2$. The color space used is (H, V), where H stands for the hue and V for the value.[1] The goal is to segment a reference region, given in the previous image of the sequence, into the current one by minimizing the distance between the reference histogram q of the region in the previous image and the estimated histogram $\hat{q}$ in the current frame. From an initial curve chosen by the user in the current frame, we want to make an active contour evolve towards the region in the current frame whose histogram is closest to the reference histogram of the previous frame.

In order to introduce a competition between the region of interest and the background region, we also consider the complement Ω^c of the region Ω of interest. They share the same boundary, Γ, but with normals pointing in opposite directions. We denote q^c the reference histogram of Ω^c, and we look for the region Ω which minimizes the following criterion:[2]

$$J(\Omega) = D(\Omega) + D(\Omega^c) + \lambda \int_\Gamma ds. \tag{7.1}$$

In this criterion, the first two terms are region functionals while the last one is a boundary functionals. The last one minimizes the curve length and is a regularization term weighted by the positive parameter λ. We have, of course,

$$D(\Omega) = \int_{\mathbb{R}^2} \varphi(\hat{q}(\boldsymbol{\alpha}, \Omega), q(\boldsymbol{\alpha}))\, d\boldsymbol{\alpha},$$
$$D(\Omega^c) = \int_{\mathbb{R}^2} \varphi(\hat{q}(\boldsymbol{\alpha}, \Omega^c), q(\boldsymbol{\alpha}))\, d\boldsymbol{\alpha}.$$

Computation of the Gâteaux derivative. A straightforward application of Theorem 6.1 yields

$$\langle D'(\Omega), \mathbf{V}\rangle = -\frac{1}{\mid \Omega \mid} \int_\Gamma \Big(\partial_1 \varphi(\hat{q}(.,\Omega), q(.)) * g_\sigma(\mathbf{h}(\mathbf{x})) - C(\Omega)\Big)(\mathbf{V}\cdot\mathbf{N}) ds$$

[1] We ignore the saturation to avoid the curse of dimensionality.

[2] The results are even better if we introduce the region area in the criterion by minimizing $D(\Omega)|\Omega| + D(\Omega^c)|\Omega^c| + \lambda \int_\Gamma ds$.

with

$$C(\Omega) = \int_{\mathbb{R}^m} \partial_1\varphi(\hat{q}(\boldsymbol{\alpha},\Omega), q(\boldsymbol{\alpha}))\hat{q}(\boldsymbol{\alpha},\Omega)\, d\boldsymbol{\alpha}\,. \tag{7.2}$$

Similar results hold for Ω^c:

$$\langle D'(\Omega^c), \mathbf{V}\rangle = \frac{1}{\mid \Omega^c \mid} \int_\Gamma \Big(\partial_1\varphi(\hat{q}(.,\Omega^c), q^c(.)) * g_\sigma(\mathbf{h}(\mathbf{x})) - C(\Omega^c)\Big)(\mathbf{V}\cdot\mathbf{N})ds\,,$$

with

$$C(\Omega^c) = \int_{\mathbb{R}^m} \partial_1\varphi(\hat{q}(\boldsymbol{\alpha},\Omega^c), q^c(\boldsymbol{\alpha}))\hat{q}(\boldsymbol{\alpha},\Omega^c)\, d\boldsymbol{\alpha}\,. \tag{7.3}$$

Computation of the evolution equation of an active contour. It is well known that the minimization of the curve length leads to the Euclidean curve shortening flow $\lambda\kappa$ [4, 35]. Then, from the previous derivatives, we can deduce the evolution of an active contour that will evolve towards a minimum of the criterion J_n defined in (7.1). We find the evolution equation

$$\frac{\partial\Gamma(\tau)}{\partial\tau} = F\mathbf{N} \tag{7.4}$$

with

$$\begin{aligned} F = \lambda\kappa &+ \frac{1}{\mid \Omega \mid}\Big(\partial_1\varphi(\hat{q}(.,\Omega), q(.)) * g_\sigma(\mathbf{h}(\mathbf{x})) - C(\Omega)\Big) \\ &- \frac{1}{\mid \Omega^c \mid}\Big(\partial_1\varphi(\hat{q}(.,\Omega^c), q^c(.)) * g_\sigma(\mathbf{h}(\mathbf{x})) - C(\Omega^c)\Big), \end{aligned}$$

where κ is the curvature of Γ and $C(\Omega)$, $C(\Omega^c)$ are given by (7.2) and (7.3), respectively.

Let us take the example of the Hellinger distance, where $\partial_1\varphi(r,\boldsymbol{\alpha}) = (\sqrt{r} - \sqrt{q(\boldsymbol{\alpha})})/\sqrt{r}$. We find for the velocity vector

$$\begin{aligned} F = \lambda\kappa &+ \frac{1}{\mid \Omega \mid}\left(\frac{(\sqrt{\hat{q}(.,\Omega)} - \sqrt{q(.)})}{\sqrt{\hat{q}(.,\Omega)}} * g_\sigma(\mathbf{h}(\mathbf{x})) - C(\Omega)\right) \\ &- \frac{1}{\mid \Omega^c \mid}\left(\frac{(\sqrt{\hat{q}(.,\Omega^c)} - \sqrt{q^c(.)})}{\sqrt{\hat{q}(.,\Omega^c)}} * g_\sigma(\mathbf{h}(\mathbf{x})) - C(\Omega^c)\right). \end{aligned}$$

And for the chi-2 comparison function where $\partial_1\varphi(r,\boldsymbol{\alpha}) = 2(r - q(\boldsymbol{\alpha}))/q(\boldsymbol{\alpha})$, we find

$$\begin{aligned} F = \lambda\kappa &+ \frac{2}{\mid \Omega \mid}\left(\frac{(\hat{q}(.,\Omega) - q(.))}{q(.)} * g_\sigma(\mathbf{h}(\mathbf{x})) - C(\Omega)\right) \\ &- \frac{2}{\mid \Omega^c \mid}\left(\frac{(\hat{q}(.,\Omega^c) - q^c(.))}{q^c(.)} * g_\sigma(\mathbf{h}(\mathbf{x})) - C(\Omega^c)\right). \end{aligned}$$

In the velocity, the convolution term allows us to compare locally the reference histogram to the current histogram.

7.1. Implementation. As far as the numerical implementation is concerned, we can model the active contour with either an explicit parameterization (Lagrangian formulation) or an implicit one (Eulerian formulation). See [23] for an interesting comparison between the two methods. Another interesting review may be found in [36].

Here, we use the level set method approach first proposed by Osher and Sethian [37] and applied this to active contours in [3]. The key idea of the level set method is to introduce an auxiliary function $U(\mathbf{x}, \tau)$ such that $\Gamma(\tau)$ is the zero level set of U. The function U is often chosen to be the signed distance function of $\Gamma(\tau)$ which satisfies

$$\Gamma(\tau) = \{\mathbf{x} \mid U(\mathbf{x}, \tau) = 0\} \quad \text{and} \quad |\nabla U| = 1.$$

This Eulerian formulation presents several advantages [47]. First, the curve U may break or merge as the function U evolves, and topological changes are thus easily handled. Second, the evolving function $U(\mathbf{x}, \tau)$ always remains a function allowing efficient numerical schemes. Third, the geometric properties of the curve, like the curvature κ and the normal vector field $\mathbf{N}$, can be estimated directly from the level set function:

$$\kappa = \operatorname{div}\left(\frac{\nabla U}{|\nabla U|}\right) \quad \text{and} \quad \mathbf{N} = -\frac{\nabla U}{|\nabla U|}.$$

The evolution equation (7.4) then becomes

$$\frac{\partial U(\tau)}{\partial \tau} = F|\nabla U|. \tag{7.5}$$

The velocity function F is computed only on the curve $\Gamma(\tau)$, but we can extend its expression to the whole image domain Ω. To implement the level set method, solutions must be found to circumvent problems coming from the fact that the signed distance function U is not a solution of the PDE (7.5); see [26] for details. In our work, the function U is re-initialized so that it remains a distance function. Details on the re-initialization equation are provided in [1, 19].

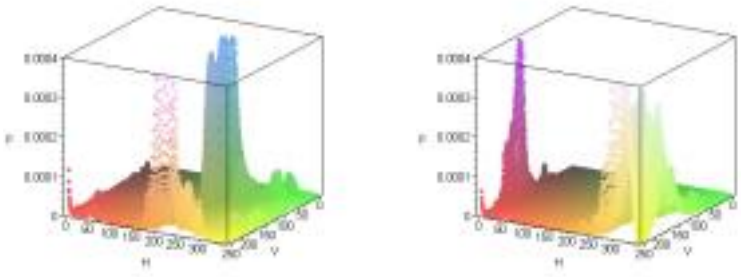
In order to improve numerical efficiency, we compute the equation in a narrow band enclosing the 0 level of the level set function [47, 48]. We also use multiresolution techniques by making the active contour evolve first in a low resolution image. The final contour obtained for this reduced image is then used as an initial curve for the real size image. Another possibility for increasing efficiency would be the use of accurate operator splitting (AOS) schemes [50].

7.2. Experimental results. Experimental results have been obtained on the sequence "Erik" from the European group COST211. Experiments are conducted using the chi-2 comparison function with $\varphi(r, \alpha) = (r - q(\boldsymbol{\alpha}))^2/q(\boldsymbol{\alpha})$ and $\partial_1\varphi(r, \boldsymbol{\alpha}) = 2(r - q(\boldsymbol{\alpha}))/q(\boldsymbol{\alpha})$.

The region of interest is the face. We assume that it has been segmented in the first image as shown in Figure 1(a). The first two reference histograms are computed. The two reference histograms are also given Figure 1(b) for the background reference histogram q^c and Figure 1(c) for the object reference histogram, q. For a given region Ω and for a point $\boldsymbol{\alpha} = [\alpha_1, \alpha_2]^T$, the function $\hat{q}(\boldsymbol{\alpha}, \Omega)$ represents the probability to obtain $H(\mathbf{x}) = \alpha_1$ and $V(\mathbf{x}) = \alpha_2$ for $\mathbf{x}$ belonging to the region Ω.

Then, using the two reference histograms of the previous frame, we make the active contour evolve using (7.4) in the current frame. The initial curve is chosen

a. Reference segmentation



b. Background reference histogram c. Object reference histogram

FIG. 1. *The reference segmentation of the previous frame* (a)*, the corresponding background reference histogram* q^c (b)*, and the corresponding object reference histogram* q (c).

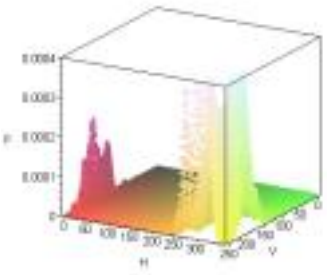
to be a circle. The evolution of the active contour in the current frame is shown in Figure 2. We notice that the final contour in Figure 2(c) nicely describes the region of interest, and the face is accurately segmented. We can also visualize the evolution of the object histogram, $\hat{q}(\boldsymbol{\alpha}, \Omega)$, during the propagation of the active contour. The final object histogram given in Figure 2(d) can be compared to the reference object histogram in Figure 1(c), showing an efficient minimization of the distance between the two histograms.

8. Conclusion. In this article we have clarified the relationships between the boundary and region functionals that arise naturally in several image processing tasks. We have shown that one can go from one to the other by solving Poisson's equation with Dirichlet conditions or Helmholtz's equation with Neumann conditions.

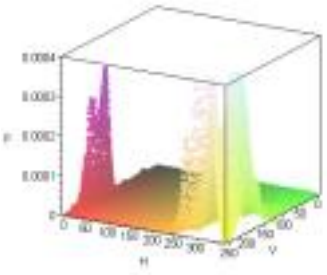
We have then concentrated on the problem of finding local minima of a large class of region functionals. By first transforming them into boundary functionals and applying methods from the calculus of variations we have computed the corresponding Gâteaux derivatives and constructed a velocity field on the region boundary. This field defines a PDE whose solution, for a given initial boundary, generates a one-parameter family of regions which, in practice, converges toward a local minimum of the functional. The problem of the existence and uniqueness of a solution to this PDE has not been addressed.

Changing our point of view, we have then rederived the same equations in a simpler and more natural way, i.e., without going through the trouble of turning region integrals into boundary integrals; this is achieved by applying methods of shape derivation [49, 22].

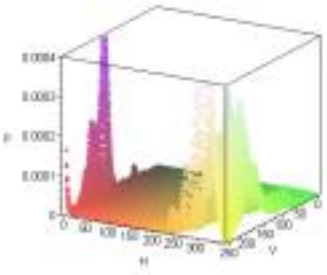
a. Initial contour



b. Initial histogram $\hat{q}(\boldsymbol{\alpha}, \Omega)$



c. Final Contour



d. Final histogram $\hat{q}(\boldsymbol{\alpha}, \Omega)$

FIG. 2. *Evolution of the region histogram* $\hat{q}(\boldsymbol{\alpha}, \Omega)$ *of the current frame during the evolution of the active contour.*

We have then turned our attention to a new class of region-based functionals by considering histograms of image features. The shape derivation tools have allowed us to easily derive the velocity field that defines the evolution of the region boundary.

The final part of the paper has been devoted to an application of the previous methods to the problem of region segmentation with a given color histogram in a sequence of images. Our experimental results show that the technique has indeed some interesting potentials.

Acknowledgments. We thank Rachid Deriche for his helpful comments on an early draft of this document. We also thank Gerardo Hermosillo for providing us with his software package for the robust estimation of image histograms.

REFERENCES

[1] G. Aubert and J.-F. Aujol, *Signed Distance Functions and Viscosity Solutions of Discontinuous Hamilton–Jacobi Equations*, manuscript, 2002.

[2] H. Brezis, *Analyse fonctionnelle. Théorie et applications*, Masson, Paris, 1983.

[3] V. Caselles, F. Catte, T. Coll, and F. Dibos, *A geometric model for active contours*, Numer. Math., 66 (1993), pp. 1–31.

[4] V. Caselles, R. Kimmel, and G. Sapiro, *Geodesic active contours*, Internat. J. Comput. Vision, 22 (1997), pp. 61–79.

[5] A. Chakraborty, L. Staib, and J. Duncan, *Deformable boundary finding in medical images by integrating gradient and region information*, IEEE Trans. Medical Imaging, 15 (1996), pp. 859–870.

[6] T. Chan and L. Vese, *An active contour model without edges*, in Scale-Space Theories in Computer Vision, Lecture Notes in Comput. Sci. 1682, Springer-Verlag, Berlin, 1999.

[7] T. Chan and L. Vese, *Active contours without edges*, IEEE Trans. Image Processing, 10 (2001), pp. 266–277.

[8] H. Chen and T. Liu, *Trust-region methods for real-time tracking*, in Proceedings of the International Conference on Computer Vision, Vol. 2, Vancouver, Canada, 2001, pp. 717–722.

[9] C. Chesnaud, P. Réfrégier, and V. Boulet, *Statistical region snake-based segmentation adapted to different physical noise models*, IEEE Trans. Pattern Analysis Machine Intelligence, 21 (1999), pp. 1145–1156.

[10] L. Cohen, *On active contour models and balloons*, Comput. Vision Graphics Image Process., 53 (1991), pp. 211–218.

[11] L. Cohen, E. Bardinet, and N. Ayache, *Surface reconstruction using active contour models*, in Proceedings of the SPIE Conference on Geometric Methods in Computer Vision, San Diego, CA, 1993.

[12] D. Comaniciu, V. Ramesh, and P. Meer, *Real-time tracking of non-rigid objects using mean shift*, in Computer Vision and Pattern Recognition, Vol. 2, IEEE Press, Piscataway, NJ, 2000, pp. 142–149.

[13] R. Dautray and J.-L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology. Vol.* 1. *Physical Origins and Classical Methods*, Springer-Verlag, Berlin, 1990.

[14] R. Dautray and J.-L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology. Vol.* 2. *Functional and Variational Methods*, Springer-Verlag, Berlin, 1988.

[15] R. Dautray and J.-L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology. Vol.* 3. *Spectral Theory and Applications*, Springer-Verlag, Berlin, 1990.

[16] R. Dautray and J.-L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology. Vol.* 4. *Integral Equations and Numerical Methods*, Springer-Verlag, Berlin, 1990.

[17] R. Dautray and J.-L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology. Vol.* 5. *Evolution Problems.* I, Springer-Verlag, Berlin, 1992.

[18] R. Dautray and J.-L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology. Vol.* 6. *Evolution Problems.* II, Springer-Verlag, Berlin, 1993.

[19] E. Debreuve, *Segmentation par contours actifs en imagerie médicale dynamique : application en cardiologie nucléaire*, Ph.D. thesis, University of Nice-Sophia Antipolis, France, 2000.

[20] E. Debreuve, M. Barlaud, G. Aubert, and J. Darcourt, *Space time segmentation using level set active contours applied to myocardial gated SPECT*, in Proceedings of the IEEE Medical Imaging Conference, Seattle USA, 1999.

[21] E. Debreuve, M. Barlaud, G. Aubert, and J. Darcourt, *Space time segmentation using level set active contours applied to myocardial gated SPECT*, IEEE Trans. Medical Imaging, 20 (2001), pp. 643–659.

[22] M. C. Delfour and J.-P. Zolésio, *Shapes and Geometries: Analysis, Differential Calculus, and Optimization*, Adv. Des. Control 4, SIAM, Philadelphia, 2001.

[23] H. Delingette and J. Montagnat, *Topology and shape constraints on parametric active contours*, Computer Vision and Image Understanding, 83 (2001), pp. 140–171.

[24] R. Duda and P. Hart, *Pattern Classification and Scene Analysis*, John Wiley, New York, 1973.

[25] L. Evans, *Partial Differential Equations*, Graduate Studies in Mathematics 19, AMS, Providence, RI, 1998.

[26] J. Gomes and O. Faugeras, *Reconciling distance functions and level sets*, J. Visual Communication and Image Representation, 11 (2000), pp. 209–223.

[27] J. Hadamard, *Mémoire sur un problème d'analyse relatif à l'équilibre des plaques élastiques encastrées*, mémoire des savants étrangers, CNRS, Paris, 1968.

[28] R. Haralick and L. Shapiro, *Computer and Robot Vision*, Vol. 1, Addison–Wesley, Reading,

MA, 1992.
[29] R. Haralick and L. Shapiro, *Computer and Robot Vision*, Vol. 2, Addison–Wesley, Reading MA, 1993.
[30] S. Jehan-Besson, M. Barlaud, and G. Aubert, *Region-based active contours for video object segmentation with camera compensation*, in Proceedings of the IEEE International Conference on Image Processing, Thessaloniki, Greece, 2001.
[31] S. Jehan-Besson, M. Barlaud, and G. Aubert, *Video object segmentation using eulerian region-based active contours*, in Proceedings of the IEEE International Conference on Computer Vision, Vancouver, Canada, 2001, pp. 353–361.
[32] S. Jehan-Besson, M. Barlaud, and G. Aubert, *A 3-step algorithm using region-based active contours for video objects detection*, EURASIP J. Appl. Signal Process., 6 (2002), pp. 572–581.
[33] S. Jehan-Besson, M. Barlaud, and G. Aubert, *DREAM^{2}S: Deformable regions driven by an eulerian accurate minimization method for image and video segmentation, application to face detection in color video sequences*, in Computer Vision - ECCV 2002, Part 3, Lecture Notes in Comput. Sci. 2352, A. Heyden, G. Sparr, M. Nielsen, and P. Johansen, eds., Springer, Berlin, 2002, pp. 365–380.
[34] M. Kass, A. Witkin, and D. Terzopoulos, *Snakes: Active contour models*, Internat. J. Comput. Vision, 1 (1988), pp. 321–332.
[35] S. Kischenassamy, A. Kumar, P. Olver, A. Tannenbaum, and A. Yezzi, *Conformal curvature flows: From phase transitions to active vision*, Arch. Rational Mech. Anal., 134 (1996), pp. 275–301.
[36] J. Montagnat, H. Delingette, and N. Ayache, *A review of deformable surfaces: Topology, geometry and deformation*, Image and Vision Computing, 19 (2001), pp. 1023–1040.
[37] S. Osher and J. Sethian, *Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton-Jacobi formulation*, J. Comput. Phys., 79 (1988), pp. 12–49.
[38] N. Paragios and R. Deriche, *Geodesic active regions for motion estimation and tracking*, in Proceedings of the IEEE International Conference on Computer Vision, Corfu Greece, 1999, pp. 688–694.
[39] N. Paragios and R. Deriche, *Coupled geodesic active regions for image segmentation: A level set approach*, in Proceedings of the European Conference in Computer Vision, Dublin, Ireland, 2000.
[40] N. Paragios and R. Deriche, *Geodesic active regions: A new paradigm to deal with frame partition problems in computer vision*, J. Visual Communication and Image Representation, 13 (2002), pp. 249–268.
[41] N. Paragios and R. Deriche, *Geodesic active regions and level set methods for supervised texture segmentation*, Internat. J. Comput. Vision, 46 (2002), p. 223.
[42] P.-A. Raviart and J.-M. Thomas, *Introduction à l'analyse numérique des équations aux dérivées partielles*, Masson, Paris, 1983.
[43] R. Ronfard, *Region-based strategies for active contour models*, Internat. J. Comput. Vision, 13 (1994), pp. 229–251.
[44] C. Samson, L. Blanc-Féraud, G. Aubert, and J. Zerubia, *A level set model for image classification*, in Scale-Space Theories in Computer Vision, Lecture Notes in Comput. Sci. 1682, Springer, Berlin, 1999.
[45] C. Samson, L. Blanc-Féraud, G. Aubert, and J. Zerubia, *A level set model for image classification*, Internat. J. Comput. Vision, 40 (2000), pp. 187–197.
[46] C. Schnörr, *Computation of discontinuous optical flow by domain decomposition and shape optimization*, Internat. J. Comput. Vision, 8 (1992), pp. 153–165.
[47] J. Sethian, *Level Set Methods*, Cambridge University Press, Cambridge, UK, 1996.
[48] J. Sethian, *Level Set Methods and Fast Marching Methods: Evolving Interfaces in Computational Geometry, Fluid Mechanics, Computer Vision, and Materials Sciences*, Cambridge Monogr. Appl. Comput. Math., Cambridge University Press, Cambridge, UK, 1999.
[49] J. Sokolowski and J.-P. Zolésio, *Introduction to shape optimization. Shape sensitivity analysis.*, Springer Ser. Comput. Math. 16, Springer-Verlag, Berlin, 1992.
[50] J. Weickert, B. ter Haar Romeny, and M. Viergever, *Efficient and reliable schemes for nonlinear diffusion filtering*, IEEE Trans. Image Process., 7 (1998), pp. 398–410.
[51] A. Yezzi, A. Tsai, and A. Willsky, *A statistical approach to snakes for bimodal and trimodal imagery*, in Proceedings of the IEEE International Conference on Image Processing, Kobe Japan, 1999.
[52] S. Zhu and A. Yuille, *Region competition: Unifying snakes, region growing, and bayes/MDL for multiband image segmentation*, IEEE Trans. Pattern Analysis Machine Intelligence, 18 (1996), pp. 884–900.

SIAM J. APPL. MATH.
Vol. 63, No. 6, pp. 2155–2175

PASSIVE LEVITATION IN ALTERNATING MAGNETIC FIELDS*

L. A. ROMERO†

Abstract. In this paper we analyze the stability of a levitated axisymmetric top carrying a system of permanent magnets in an alternating magnetic field. We show that there are stable configurations where the top is stationary, and the alternating magnetic field stabilizes the equilibrium position. We show that one mechanism for achieving stability is to periodically change the coupling between the rotational and translational degrees of freedom.

Key words. Levitron, stability, magnets

AMS subject classifications. 70E05, 70J25, 78A30

DOI. 10.1137/S003613990241031X

1. Introduction.

Earnshaw's theorem states that it is impossible to have stable levitation in a magneto-static or electro-static field. Although the theorem has a very broad scope, there are some very precise conditions that must be met in order for it to apply. In particular, the theorem fails if there are diamagnetic materials present, the levitating body is spinning, or the fields are alternating in time.

The discovery of the Levitron™ [7] demonstrated that it is possible to have stable levitation in a steady magneto-static field. This has led to a renewed interest in the subject of passive magnetic levitation using magneto-static fields [2], [4], [5], [11], [3]. Stable levitation has also been demonstrated using diamagnetic materials. A particular case of this is levitation over a superconducting disc, which can be considered as a diamagnetic material with magnetic permeability of zero.

In this paper we would like to discuss the possibility of achieving passive stable levitation in an alternating magnetic field. Although the fields are varying with time, we still refer to them as being magneto-static, since we assume that they are varying slowly enough that we can ignore the effects of the time varying terms in Maxwell's equations. We say that our system is passive because we prescribe the time variation of the magnetic field ahead of time, rather than adjusting the field in response to the position of the top.

In [10] Paul reviews the work on how small particles can be levitated in an alternating field. This work has demonstrated both theoretically and experimentally how Earnshaw's theorem can be violated in an alternating field. Our paper differs from this work in several respects. Our paper is aimed at trying to design a system on a larger scale, where we have both the flexibility and the necessity of specifying the design of both the levitating body and the supporting magnets.

The theory outlined in [10] applies to point charges in an alternating field. It is very simple and elegant, and has found profound uses in such fields as mass spectrometry. However, if one is going to make an oscillating field (and not spinning) version of the Levitron™, one needs a fuller theory. In particular, we need to take into account

*Received by the editors June 26, 2002; accepted for publication (in revised form) January 21, 2003; published electronically September 22, 2003. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin company, for the United States Department of Energy under contract DE-AC04-94AL85000. The U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. Copyright is owned by SIAM to the extent not limited by these rights.
http://www.siam.org/journals/siap/63-6/41031.html
†Sandia National Laboratories, Albuquerque, NM 87185 (laromero@cs.saudia.gov).

the full (although linearized) rigid body dynamics of the levitated body. When we do this we find that there are two quite distinct methods of stabilizing the top. The first is the mechanism outlined in [10]; the second involves periodically varying the coupling between the rotational and translational modes of the top. In a practical device, one would most likely want to bring both of these mechanisms into play.

We will now describe our proposed setup and the conclusions of the analysis. We assume that there is a system of base magnets that gives a steady magnetic field with axial symmetry about the z axis (the direction of gravity). We also assume that we have a system of axisymmetric magnetic coils that give an axisymmetric periodically varying magnetic field. We suppose that we have an axisymmetric top with an axisymmetric system of magnets attached to it. When the top's axis is pointing in the z direction, and its center of mass is at the origin, we assume that the magnetic force from the steady field exactly balances the force of gravity. We also assume that the time varying force on the top vanishes at this position. In section 7 we will show that it is possible to find configurations of magnets that achieve these conditions. Assuming these conditions are satisfied, we see that there is an equilibrium position where the top is at rest in the time varying field. We then consider the stability of this equilibrium.

The linear stability equations are identical to those used in analyzing the stability of an axisymmetric Levitron™, except the dynamical parameters are time varying and the spin rate is set to zero. In particular, the system can be described by a lateral translational spring constant A, an axial translational spring constant $-2A$, a rotational spring constant $-C$, and a term B that couples the translations to the rotations. All of these terms have both a steady and a time varying component. Due to Earnshaw's theorem, if we set the time varying fields to zero, it is not possible for the system to be stable to both axial and lateral translational perturbations.

The mechanism described in [10] is equivalent to assuming that there is no coupling term B in the equations of motion and the rotational spring constant $-C$ is positive. This gives us a rotationally stable top, but one that at each instant of time sees a negative spring constant for either axial or lateral perturbations. However, by varying the translational spring constant $2A(t)$ at the right amplitude and frequency, it is possible to achieve overall stability in both the axial and lateral directions.

A different mechanism for achieving stability is to keep the translational spring constant A and the rotational spring constant $-C$ steady in time and to periodically vary the coupling term $B(t)$. Our results show that in order to achieve stability, it is necessary that the field varies more rapidly than the natural frequency of the rotational oscillation of the top. Furthermore, there is both an upper and a lower limit to the strength of the time varying coupling term. We cannot achieve stable equilibrium if the coupling is either too big or too small.

We now give a brief outline of the rest of this paper. In section 2 we derive the equations governing the linear stability of the top. In section 3 we discuss how to apply Floquet theory to analyze these equations and give general properties of the stability equations. In section 4 we analyze the case where stabilization is achieved without any coupling between the rotations and translations. In section 5 we consider the case where this coupling is the stabilizing mechanism. In section 6 we discuss the high frequency approximation to the stability equations. In section 7 we discuss how to find configurations that give us the desired dynamical parameters. We give conclusions in section 8.

2. Basic equations. We suppose that there is an equilibrium position where the center of mass of the top is at the origin and its axis of symmetry is pointing in the z direction. We should emphasize that even though the magnetic field is varying with time, we assume that a steady state equilibrium exists. For this to be true, it is necessary that the time varying components of the force and torque vanish at the equilibrium position. Due to symmetry, when the top is placed symmetrically in the field, all components of the force and torque vanish except for the force in the z direction. We must arrange the magnets so that the time varying component of this force vanishes.

We now determine the form of the equations governing the linear stability of this equilibrium. We begin by using simple symmetry arguments to deduce the most general form of the linear stability equations.

In order to orient the axis of symmetry of the top, we assume that the axis is initially pointing in the positive z direction. We orient the body by rotating it about the x axis by θ, then about the y axis by ϕ, and then about the z axis by ψ. Since we are mainly concerned with the linear stability of the system, and we are assuming that the top is axisymmetric, the angle ψ will not appear in the equations of motion.

2.1. Symmetry considerations. For any axisymmetric top with an axisymmetric system of magnets, the energy of the top in a magnetic field can be written as

$$Energy = U(\underline{x}, \underline{d}, t),$$

where $\underline{x} = (x, y, z)$ gives the position of the center of mass of the top and $\underline{d} = (d_x, d_y, d_z)$ is a unit vector pointing in the direction of the axis of symmetry of the top. This very general form for the energy holds as long as the top is axisymmetric and carries with it an axisymmetric system of magnets.

The force and torques can be derived from the potential $U(\underline{x}, \underline{d}, t)$. The forces can be written as

$$\underline{F}(\underline{x}, \underline{d}, t) = -\nabla_x U(\underline{x}, \underline{d}, t). \tag{1}$$

Here ∇_x is the gradient with respect to $\underline{x}$. The torque $\underline{\tau}$ can be computed using the fact that if we rotate our system about the axis $\underline{a}$ by an infinitesimal amount $d\alpha$, the change in energy can be written as

$$dU = -\underline{\tau} \cdot \underline{a} d\alpha.$$

When we rotate our system in this way, the axis $\underline{d}$ changes by the amount $d\alpha \underline{a} \times \underline{d}$. It follows that the change in energy can be written as

$$dU = \nabla_d U(\underline{x}, \underline{d}, t) \cdot (d\alpha \underline{a} \times \underline{d}) = d\alpha \underline{a} \cdot (\underline{d} \times \nabla_d U).$$

Equating the two expressions for dU and requiring that they hold for all values of $\underline{a}$, we conclude that the torque is given by

$$\underline{\tau} = -\underline{d} \times \nabla_d U(\underline{x}, \underline{d}, t). \tag{2}$$

We will also use the fact that

$$\nabla_x^2 U = 0.$$

We are interested in analyzing the stability of an equilibrium position where $\underline{d} = (0, 0, 1)$. For small perturbations to this solution, both θ and ϕ are small, and we

have the approximate expression $(d_x, d_y, d_z) \approx (\phi, -\theta, 1)$. In the linear approximation, the energy U is a quadratic function of x, y, z, θ, and ϕ. In this approximation, the torques τ_x and τ_y are given by

$$\tau_x = -\frac{\partial U}{\partial \theta},$$

$$\tau_y = -\frac{\partial U}{\partial \phi}.$$

We now assume that the top is in an axisymmetric magnetic field with the z axis as the axis of symmetry. Due to the symmetry of our configuration, it is clear that if we reflect both the position and the vector $\underline{d}$ about either the x or y axes, the energy stays the same, as it does for reflections about the plane $x = y$. This implies that

$$U(x, y, z, d_x, d_y, d_z, t) = U(-x, y, z, -d_x, d_y, d_z, t), \tag{3a}$$

$$U(x, y, z, d_x, d_y, d_z, t) = U(x, -y, z, d_x, -d_y, d_z, t), \tag{3b}$$

$$U(x, y, z, d_x, d_y, d_z, t) = U(y, x, z, d_y, d_x, d_z, t). \tag{3c}$$

These symmetry properties must be satisfied by any axisymmetric system. We could write down other symmetry properties, but these suffice to specify the form of the equations of motion.

In the linear approximation these symmetry conditions can be written as

$$U(x, y, z, \theta, \phi, t) = U(-x, y, z, \theta, -\phi, t),$$

$$U(x, y, z, \theta, \phi, t) = U(x, -y, z, -\theta, \phi, t),$$

$$U(x, y, z, \theta, \phi, t) = U(y, x, z, -\phi, \theta, t).$$

These symmetries, along with the fact that the forces and torques are derivable from a potential, imply that the general form for the linearized equations of an axisymmetric top can be written as

$$m\ddot{x} - A(t)x - B(t)\phi = 0,$$

$$m\ddot{y} - A(t)y + B(t)\theta = 0,$$

$$m\ddot{z} + 2A(t)z = 0,$$

$$I_1\ddot{\theta} - C(t)\theta + B(t)y = 0,$$

$$I_1\ddot{\phi} - C(t)\phi - B(t)x = 0.$$

These linearized stability equations depend on the mass m, the moment of inertia I_1, and the dynamical constants A, B, and C. These dynamical constants are given by

$$A(t) = -\frac{1}{2}\frac{\partial F_z}{\partial z} = \frac{\partial F_x}{\partial x} = \frac{\partial F_y}{\partial y},$$

$$B(t) = -\frac{\partial \tau_x}{\partial y} = \frac{\partial \tau_y}{\partial x} = \frac{\partial F_x}{\partial \phi} = -\frac{\partial F_y}{\partial \theta},$$

$$C = \frac{\partial \tau_x}{\partial \theta} = \frac{\partial \tau_y}{\partial \phi}.$$

In these expressions the partial derivatives are to be evaluated at the equilibrium position, which is assumed to be with the top aligned with the axis of symmetry and the center of mass along the axis of symmetry. Note that the dynamical parameters A, B, and C are time dependent due to the time varying nature of the fields.

The dynamical constant A is the translational spring constant. For a steady field, if $B = 0$, then when $A > 0$ we have a stable harmonic oscillator for displacements in the z direction and (due to Earnshaw's theorem) an unstable oscillator in the x and y directions. The constant $-C$ is a torsional spring constant. Finally, the constant B gives the coupling between the rotations and translations. For example, when the top is rotated about the y axis by ϕ, it gives a force in the x direction proportional to $B\phi$, and when the top is displaced in the x direction, there is a torque about the y axis that is proportional to Bx.

Once again we emphasize that this form of the equations holds for any axisymmetric system, no matter how complicated the systems of magnets for either the base or top are. If the top has a single dipole, quadrapole, or octapole on it, and the base consists of a single coil or a single ring magnet, it is possible to find explicit expressions for the constants A, B, and C. Using linear superposition, it is then possible to compute the dynamical constants A, B, and C for arbitrarily complicated systems of magnets on the top and in the base. Alternatively, assuming we have a code that can compute the force and torque on the top when it is placed with an arbitrary position and orientation in an axisymmetric field, we can compute the constants A, B, and C by numerically taking the derivatives of the forces and torques about the equilibrium position.

Given a particular configuration of magnets, the functions $A(t)$, $B(t)$, and $C(t)$ determine the stability properties of the configuration. However, in order to determine if a configuration is in equilibrium, we also need to calculate the lift $L(t)$. In order for this to be a valid configuration, it is necessary that the lift $L(t)$ be independent of time and that it be equal and opposite to the force of gravity on the top.

2.2. Sinusoidally varying fields. In the most general case that we consider in this paper, all of the terms have both a steady and a sinusoidally varying component. We assume that all of the time varying components have the same phase. We will write

$$A(t) = A_0 + A_v \cos(\omega t),$$

$$B(t) = B_0 + B_v \cos(\omega t),$$

$$C(t) = C_0 + C_v \cos(\omega t).$$

We are assuming that the lift $L(t)$ is independent of time.

We will introduce the following dimensionless variables:

$$x = \xi\sqrt{I_1/m},$$

$$y = \chi\sqrt{I_1/m},$$

$$z = \eta\sqrt{I_1/m},$$

$$t = s/\omega.$$

In terms of these dimensionless variables, we get the equations

$$\ddot{\xi} - (\alpha_0 + \alpha_v \cos(s))\xi + (\beta_0 + \beta_v \cos(s))\phi = 0, \tag{4a}$$

$$\ddot{\phi} + (\gamma_0 + \gamma_v \cos(s))\phi + (\beta_0 + \beta_v \cos(s))\xi = 0, \tag{4b}$$

$$\ddot{\eta} + 2(\alpha_0 + \alpha_v \cos(s))\eta = 0. \tag{4c}$$

Here

$$\alpha_0 = \frac{A_0}{m\omega^2},$$

$$\gamma_0 = \frac{-C_0}{I_1\omega^2},$$

$$\beta_0 = \frac{B_0}{\sqrt{mI_1}\omega^2},$$

$$\alpha_v = \frac{A_v}{m\omega^2},$$

$$\gamma_v = \frac{-C_v}{I_1\omega^2},$$

$$\beta_v = \frac{B_v}{\sqrt{mI_1}\omega^2}.$$

Note that the equation for η decouples from those for ξ and ϕ. Furthermore, the equations for χ and θ decouple from the rest of the equations and are almost identical to the equations for ξ and ϕ (and hence we have not bothered to write them down). If the equations for ξ and ϕ are stable, then so will the ones for χ and θ. For this reason we will not carry them along in our analysis.

3. General properties concerning the stability of the equilibrium. The equations (4) can be split up into two independent systems. The equation for η is a second order system that decouples from the rest of the equations. The equations for

ξ and ϕ can be written as a fourth order system of the form

$$\dot{\underline{q}} = R(t)\underline{q}, \tag{5}$$

$$\underline{q} = \begin{pmatrix} \xi \\ \dot{\xi} \\ \phi \\ \dot{\phi} \end{pmatrix},$$

where $R(t)$ is a 4×4 matrix that satisfies $R(t) = R(t + 2\pi)$.

In order to have stability, both the equation for η and (5) must give stability. The equation for η is Mathieu's equation, whose stability properties are well known. In the rest of this section we discuss how to analyze the stability of (5). The arguments we now give are a special case of those given in [8].

In order to analyze the stability of our system of equations we consider the fundamental matrix solution $Q(t)$ satisfying

$$\dot{Q} = R(t)Q,$$

$$Q(0) = I.$$

We define the monodromy matrix Γ as

$$\Gamma = Q(2\pi).$$

The equilibrium with $\underline{q} = 0$ will be stable provided all of the eigenvalues of Γ have magnitudes less than unity. However, since our system is conservative we can at best have neutral stability. This follows since $\mathrm{tr}(R(t)) = 0$. This implies that the Wronskian $Wr(t)$ (the determinant of $Q(t)$) is constant. This follows since the Wronskian satisfies the equation

$$\frac{dWr}{dt} = \mathrm{tr}(R(t))Wr.$$

Since the Wronskian is independent of time, the determinant of $Q(2\pi)$ must be equal to one. This implies that it is not possible to have all of the eigenvalues of Γ have magnitudes less than unity. For this reason we can have only instability or neutral stability.

For given values of the parameters α_0, α_v, etc., it is straightforward to numerically integrate the system of equations and evaluate the eigenvalues of the monodromy matrix Γ. However, if one is going to track (either numerically or with perturbation theory) the curves separating regions of instability from regions of neutral stability, it is helpful to understand some more properties of Γ.

We begin by showing that due to time reversal symmetry, if λ is an eigenvalue of Γ, then so is $1/\lambda$. Suppose we integrate (5) with the initial conditions $\underline{q}(0) = \underline{q}_0$, and after integrating this equation over a single period we end up with the vector $\underline{q}(2\pi) = \underline{q}_1$. If we integrate our differential equations backwards with the initial condition $\underline{q}_1$, we will end up getting back $\underline{q}_0$ at $t = 0$. However, due to the time reversal symmetry of our system, we can get back $\underline{q}_0$ by integrating our equations

forward. In order to do this we need to multiply $\underline{q}_1$ by the matrix

$$J = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}.$$

This has the effect of keeping the values of ξ and ϕ (in the vector $\underline{q}$) the same but changing the signs of their derivatives. If we integrate (5) forward with the initial condition $\underline{q}(0) = J\underline{q}_1$, we will end up with the vector $J\underline{q}_0$ after we have integrated over a single period. This will work for any initial condition $\underline{q}_0$, and hence our monodromy matrix must satisfy

$$\Gamma J \Gamma = J. \tag{6}$$

We will now show that this equation implies that if λ is an eigenvalue of Γ, then so is $1/\lambda$. In order to show this we note that (6) can be written as

$$J\Gamma = \Gamma^{-1} J.$$

Suppose λ is an eigenvalue of Γ with eigenvector $\underline{r}$. This means that $\Gamma\underline{r} = \lambda\underline{r}$, and hence $J\Gamma\underline{r} = \lambda J\underline{r}$. However, using $J\Gamma = \Gamma^{-1}J$, this can be written as $\Gamma^{-1}J\underline{r} = \lambda J\underline{r}$. This implies that λ is an eigenvalue of Γ^{-1} with eigenvector $J\underline{r}$. This is equivalent to saying that $1/\lambda$ is an eigenvalue of Γ.

It follows that the characteristic equation for Γ is a reciprocal equation. That is, it can be written as

$$\det(\lambda I - \Gamma) = \lambda^4 + p(\Gamma)\lambda^3 + q(\Gamma)\lambda^2 + p(\Gamma)\lambda + 1 = 0. \tag{7}$$

Reciprocal equations can be solved by solving equations of half the degree and a quadratic [9]. For fourth order reciprocal equations we use the polynomials $\psi_0 = 1, \psi_1(t) = t$, $\psi_2(t) = t^2 - 2$. These polynomials satisfy

$$\psi_k(x + 1/x) = x^k + 1/x^k.$$

To solve (7) we multiply (7) by $1/\lambda^2$ to get

$$(\lambda^2 + 1/\lambda^2) + p(\Gamma)(\lambda + 1/\lambda) + q(\Gamma) = 0.$$

This can be written as

$$\psi_2(\lambda + 1/\lambda) + p(\Gamma)\psi_1(\lambda + 1/\lambda) + q(\Gamma) = 0.$$

If we make the substitution $z = \lambda + 1/\lambda$, we get the equation

$$z^2 + p(\Gamma)z + q(\Gamma) - 2 = 0, \tag{8}$$

and solving for λ in terms of z we get

$$\lambda^2 - z\lambda + 1 = 0. \tag{9}$$

In order for this last equation to have roots with $\mid \lambda \mid = 1$, it is necessary and sufficient that z be real and satisfy $z^2 < 4$. It follows that a necessary and sufficient

condition that our system is neutrally stable is that all of the roots of (8) be real and have magnitude less than two. These conditions can be written as

$$-4 < p(\Gamma) < 4,$$

$$8 + p(\Gamma)^2 - 4q(\Gamma) > 0,$$

$$2 + 2p(\Gamma) + q(\Gamma) > 0,$$

$$2 - 2p(\Gamma) + q(\Gamma) > 0.$$

It is more useful to express these equations directly in terms of the monodromy matrix. To do this we use

$$p(\Gamma) = -\mathrm{tr}(\Gamma),$$

$$q(\Gamma) = \frac{1}{2}\left(p(\Gamma)^2 - \mathrm{tr}(\Gamma^2)\right).$$

The necessary and sufficient conditions for stability can now be written as

$$Z_1(\Gamma) = 8 + 2\,\mathrm{tr}(\Gamma^2) - \mathrm{tr}(\Gamma)^2 > 0, \tag{10a}$$

$$Z_2(\Gamma) = (2 - \mathrm{tr}(\Gamma))^2 - \mathrm{tr}(\Gamma^2) > 0, \tag{10b}$$

$$Z_3(\Gamma) = (2 + \mathrm{tr}(\Gamma))^2 - \mathrm{tr}(\Gamma^2) > 0, \tag{10c}$$

$$Z_4(\Gamma) = 4 - \mathrm{tr}(\Gamma) > 0, \tag{10d}$$

$$Z_5(\Gamma) = \mathrm{tr}(\Gamma) + 4 > 0. \tag{10e}$$

On the boundary between a region of stability and instability, one of these inequalities is replaced by an equality. For example, numerically it is found that $Z_1(\Gamma) = 0$ is satisfied for the upper bound on β_v and $Z_2(\Gamma) = 0$ for the lower bound on β_v when $\alpha_v = \gamma_v = \beta_0 = 0$ and $\gamma_0 > 0$.

4. Systems with no coupling term $B(t)$. We begin by assuming that both the steady and time varying components of the coupling term $B(t)$ vanish. In this case the equations for ϕ, θ, ξ, χ, and η all decouple from each other. The equations for ϕ and θ imply that the top is rotationally stable provided $\gamma_0 > 0$ and γ_v is not too large. Assuming this is the case, the stability is determined by the condition that the equations

$$\ddot{\xi} - (\alpha_0 + \alpha_v \cos(s))\xi = 0$$

and

$$\ddot{\eta} + 2(\alpha_0 + \alpha_v \cos(s))\eta = 0$$

both give stable solutions. This theory is nearly identical to that presented in [10]. Each one of these equations is a particular case of Mathieu's equation. The theory of Mathieu's equation has been well documented in many different books on nonlinear

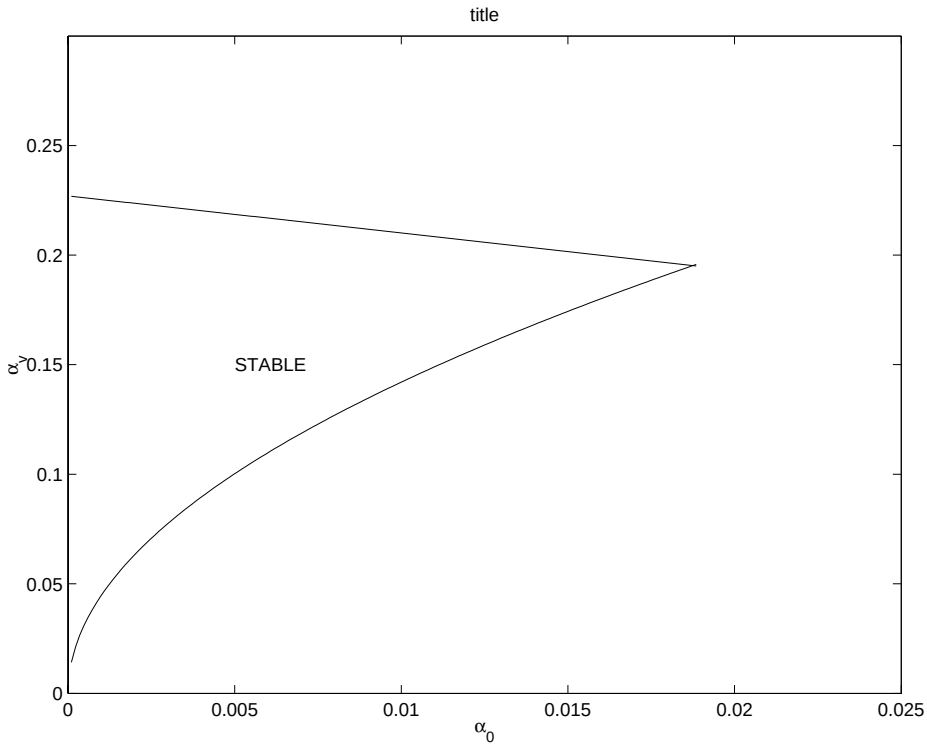


FIG. 1. *This shows the upper and lower values of α_v as a function of α_0 for the case where there is no coupling between the rotational and translational modes.*

oscillations [12], [1]. The only trick here is adjusting the parameters so that both of these equations give stable solutions at the same time. Note that at any instant in time, if we freeze the coefficients, only one of these equations would be stable. However, since the fields are changing with time, we can get overall stability even though the equations with frozen coefficients give instability.

We have numerically integrated these equations to determine the monodromy matrix and the Floquet exponents. Figure 1 shows a plot of the region of stability.

5. Systems stabilized by coupling. We now consider systems where the only time variation is in the coupling term β_v. For simplicity we assume that the steady component β_0 also vanishes. Since the equation for η is decoupled from the rest of the equations, it is clear that we must have $\alpha_0 > 0$ in order to have the top be stable to axial perturbations. Assuming that $\alpha_0 > 0$, the stability of such a system is governed by the equations

$$\ddot{\xi} - \alpha_0 \xi + \beta_v \cos(s)\phi = 0, \tag{11a}$$

$$\ddot{\phi} + \gamma_0 \phi + \beta_v \cos(s)\xi = 0. \tag{11b}$$

Before giving numerical or perturbation results, we will give some intuitive arguments that give the basic features of the numerical calculations.

Suppose that both β_v and α_0 are small. In this case it is reasonable to assume that $\xi(t)$ is changing slowly compared to the driving frequency, and hence $\xi(s)$ can be approximated as a constant in (11b). This implies that we can write

$$\phi(s) = \frac{\xi(s)\beta_v \cos(s)}{1-\gamma_0}.$$

If we substitute this back into the equation for ξ, we get

$$\ddot{\xi} - \alpha_0 \xi + \xi \frac{\beta_v^2}{1-\gamma_0} \cos^2(s) = 0.$$

This can be written as

$$\ddot{\xi} + \left(\frac{\beta_v^2}{2(1-\gamma_0)} - \alpha_0 \right) \xi + \frac{\beta_v^2}{2(1-\gamma_0)} \cos(2s)\xi = 0. \tag{12}$$

This is identical to Mathieu's equation. As long as we are away from regions of parametric resonance we expect that this will be stable provided the steady term in the spring constant is positive. By this we mean that

$$\frac{\beta_v^2}{2(1-\gamma_0)} - \alpha_0 > 0. \tag{13}$$

We see that if we make the coupling term β_v large enough, we can overcome the destabilizing effect of the parameter α_0. However, this is possible only if $\gamma_0 < 1$. However, as the coupling term increases, the steady term approaches 1 and the oscillating term grows. Eventually we are going to reach a point where the system becomes unstable due to parametric excitation.

We can numerically integrate (12) and use Floquet theory to determine regions of stability and instability. Figure 2 shows a plot of the neutral stability region for this equation. We see that the lower limit on stability is well approximated by the curve $\beta_v^2 = 2\alpha_0(1-\gamma_0)$. This lower limit on stability also agrees very well with the lower limit obtained by doing Floquet theory on the full set of equations (11).

The condition $\gamma_0 < 0$ corresponds to having the top be rotationally unstable in the absence of the coupling term β_v. In this case if β_v and α_0 are both small, our system will be unstable due to (11b). In this case it is necessary to stabilize the instability to rotations in a manner similar to the way we stabilized the translational instability. Almost identical arguments show that it is necessary to have

$$\frac{\beta_v^2}{2(1-\alpha_0)} + \gamma_0 > 0 \tag{14}$$

in order to stabilize the rotational instability. This shows that when γ_0 is negative, then even when α_0 is infinitesimally small, there is a minimum value of β_v necessary in order to achieve stability.

We now present numerical results that confirm all of our results based on heuristic reasoning. The only major modification to these results is that the results based on (12) do not do a good job of predicting the upper limit on β_v. In particular, the full numerical calculations show that the upper and lower limits converge towards each other at a finite value of α_0.

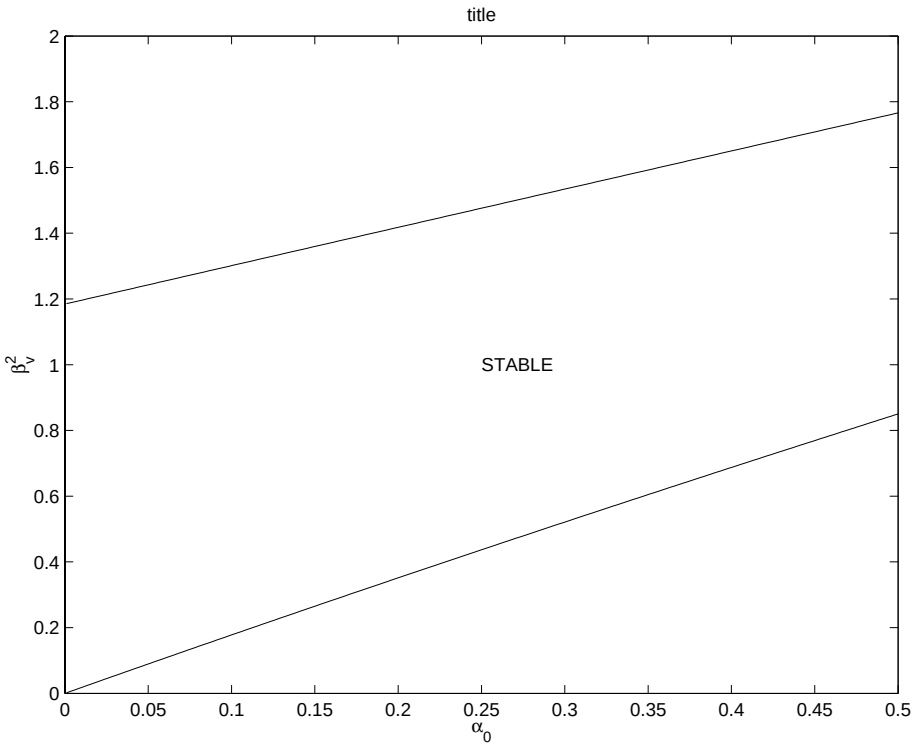


FIG. 2. *This shows the upper and lower values of β_v for stable operation predicted by* (12). *This plot is for $\gamma_0 = .1$.*

Figure 3 shows the bounds on β_v as a function of α_0 when we use the full linearized stability equations (11) to analyze the stability. We see that the upper and lower curves intersect tangentially at a finite value of α_0. As we increase γ_0, the region of stable operation shrinks until it completely vanishes just before $\gamma_0 = 1$.

Figure 4 shows similar plots for negative values of γ_0. Note that when γ_0 is negative the curve for the lower limit on β_v has a kink in it that occurs when $\alpha_0+\gamma_0=0$. The first part of the curve is very well approximated by (14), and the second part of the curve is approximated by (13). As γ_0 gets smaller, the region of stability eventually shrinks to zero.

The "exact" numerical results we present were found by numerically integrating the linearized system of equations with different initial conditions in order to obtain the monodromy matrix. For given values of our parameters this allows us to compute the quantities Z_k in (10). On a surface in parameter space where the system changes stability we must have $Z_k = 0$ for at least one of $k = 1, 2, 3, 4$. If we have a good guess for a point in parameter space satisfying $Z_k = 0$, we can use the secant method to change one parameter until we find a point that exactly satisfies this equation. Using this technique it is possible to map out the regions of stability quite efficiently.

6. The high frequency approximation. The situations we have considered in the last two sections were aimed at isolating two important mechanisms for stabilizing the equilibrium. In particular, both situations assumed that the steady coupling term β_0 was identically zero. Figure 5 shows plots of the upper and lower bounds on β_v

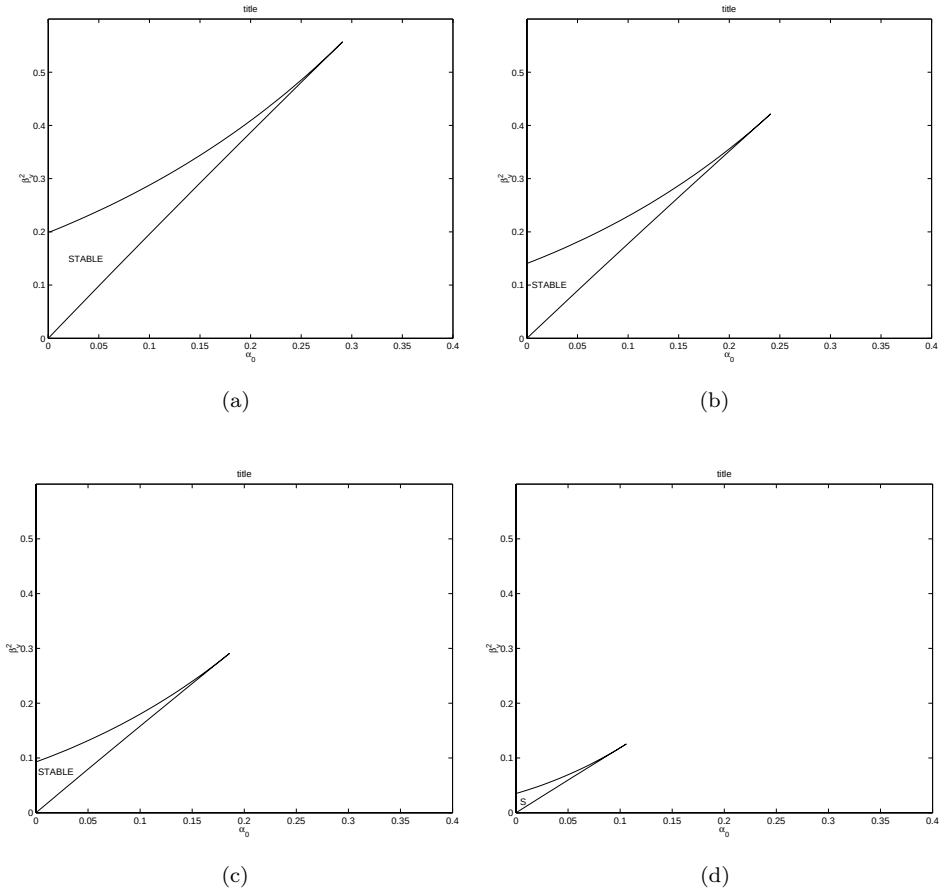


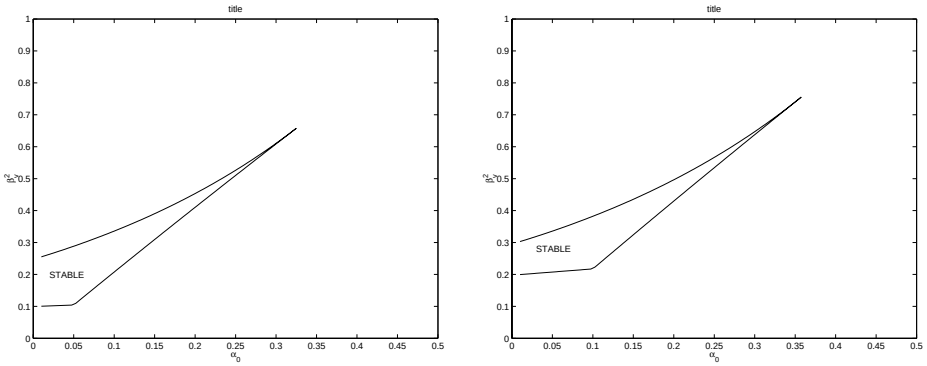
FIG. 3. *This shows the curves for the upper and lower limits to β_v as a function of α_0.* (a) $\gamma_0 = .0$, (b) $\gamma_0 = .1$, (c) $\gamma_0 = .2$, (d) $\gamma_0 = .4$. *Note that as γ_0 gets close to* 1, *the region of stability shrinks to zero.*

as a function of α_0 when we let the parameter β_0 be nonzero. We see that the lower limit on stability is raised as we increase β_0, and for $\alpha_0 = 0$ the amount it is raised is nearly proportional to $\sqrt{\beta_0}$.

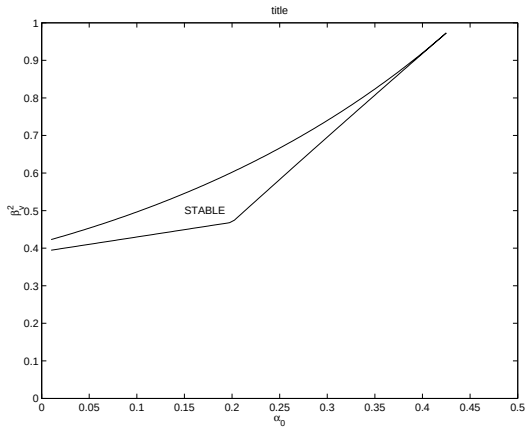
The dependence of our stability curves on β_0 is just one aspect of the stability boundaries that were not explored in the last two sections. In order to understand fully the stability of the equilibrium, we have a six parameter space to explore. The analysis in this section is aimed at attempting to understand the stability boundaries of this space. In order to do this we consider the stability of the equilibrium when all of the dimensionless parameters α_0, β_0, γ_0, α_v, β_v, and γ_v are small. Physically this can be achieved by making the driving frequency ω large.

We can derive the high frequency approximation using the method of averaging. We briefly outline the method of averaging outlined in [6] when applied to linear systems with periodic coefficients. In this special case, we are concerned with systems of the form

$$\dot{\underline{z}} = \epsilon M(t)\underline{z}, \tag{15}$$



(a) (b)



(c)

FIG. 4. *This shows the curves for the upper and lower limits to β_v as a function of α_0.* (a) $\gamma_0 = -.05$, (b) $\gamma_0 = -.1$, (c) $\gamma_0 = -.2$. *Note that the lower bound on β_v follows one branch until $\alpha_0 = -\gamma_0$, then it follows a different branch. As γ_0 decreases, the lower bound on β_v for $\alpha_0 = 0$ is continually raised.*

where ϵ is small and M is a 2π periodic matrix. The method of averaging shows that we can make a change of variable of the form

$$\underline{z} = \underline{\xi} + \left(\epsilon N_1(t) + \epsilon^2 N_2(t) + \cdots + \epsilon^k N_k(t)\right) \underline{\xi},$$

where each $N_i(t)$ is a 2π periodic matrix with zero time average, such that the equation for $\underline{\xi}$ can be written as

$$\dot{\underline{\xi}} = M_{av}\underline{\xi} + O(\epsilon^{k+1}),$$

where M_{av} can be written as

$$M_{av} = \epsilon M_1 + \epsilon^2 M_2 + \cdots + \epsilon^k M_k.$$

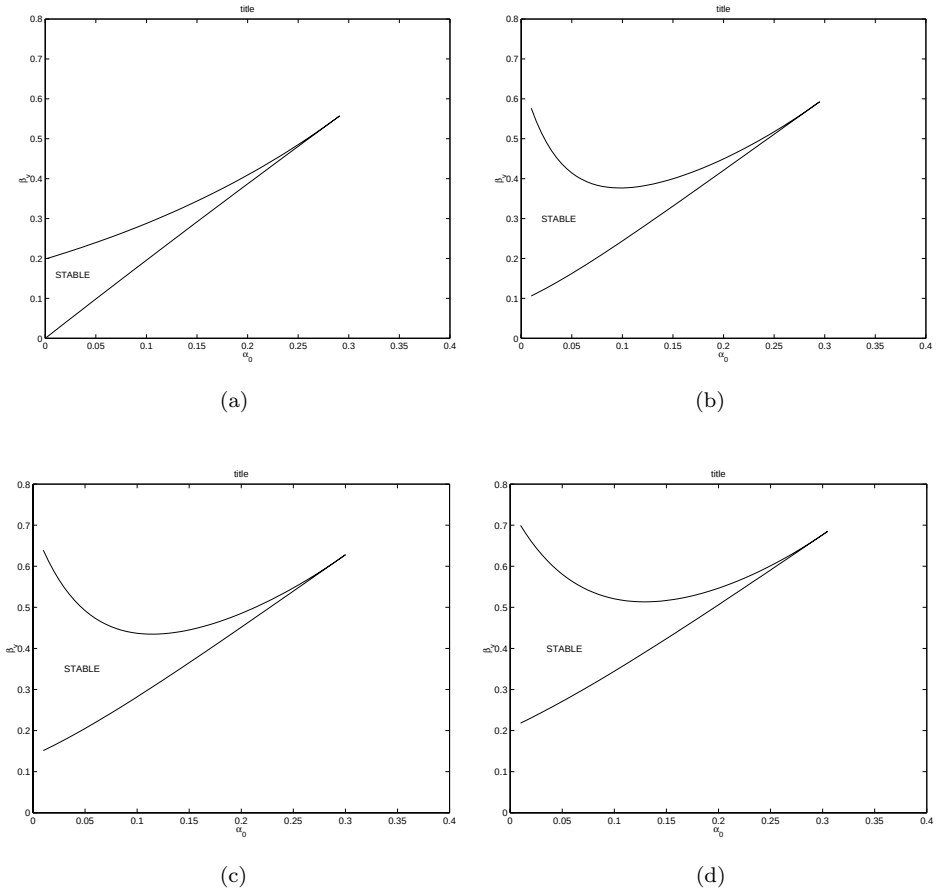


FIG. 5. *This shows the curves for the upper and lower limits to β_v as a function of α_0 with $\gamma_0 = 0$ and different values of β_0.* (a) $\beta_0 = 0$, (b) $\beta_0 = .005$, (c) $\beta_0 = .002$, (d) $\beta_0 = .01$. *Note that increasing β_0 moves the region of neutral stability upwards.*

Here each of the matrices M_j is independent of time. Simple arguments show that the eigenvalues of the matrix M_{av} must agree with the Floquet exponents of the original system (15) to order ϵ^k.

If we carry out this method of averaging, we get

$$M_1 = \langle M(t) \rangle,$$

where

$$\langle f(t) \rangle = \frac{1}{2\pi} \int_0^{2\pi} f(t) dt.$$

The matrix $N_1(t)$ is the unique matrix with $\langle N_1 \rangle = 0$ satisfying

$$\frac{dN_1}{dt} = M(t) - M_1.$$

The matrix M_2 is given by

$$M_2 = \langle M(t)N_1(t) - N_1(t)M_1 \rangle.$$

The matrix $N_2(t)$ is the unique matrix with $\langle N_2 \rangle = 0$ satisfying

$$\frac{dN_2}{dt} = M(t)N_1(t) - N_1(t)M_1 - M_2.$$

The matrix M_3 is given by

$$M_3 = \langle M(t)N_2(t) - N_1(t)M_2 - N_2(t)M_1 \rangle.$$

Assuming that the dimensionless parameters $\alpha(t)$, $\beta(t)$, and $\gamma(t)$ are small, it is possible to apply the method of averaging to our linearized stability equations. We have, in fact, done this. However, we will now present a simpler heuristic analysis that leads to the same results as carrying out this averaging process to third order.

We suppose that the steady parameters α, β, and γ are small compared to the time varying parameters α_v, β_v, and γ_v; we assume that the time varying parameters are themselves small. We will write the solution $\xi(s)$ and $\phi(s)$ as

$$\xi(s) = \xi_0(s) + \xi_1(s),$$

$$\phi(s) = \phi_0(s) + \phi_1(s),$$

where ξ_0 and ϕ_0 are slowly varying and ξ_1 and ϕ_1 are rapidly varying.

Under these assumptions the first two equations in equations (4) imply that the fast components satisfy

$$\ddot{\xi}_1 - \alpha_v \cos(s)\xi_0 + \beta_v \cos(s)\phi_0 = 0, \tag{16}$$

$$\ddot{\phi}_1 + \gamma_v \cos(s)\phi_0 + \beta_v \cos(s)\xi_0 = 0. \tag{17}$$

We can solve these equations to get

$$\xi_1 = \cos(s)\left(-\alpha_v \xi_0 + \beta_v \phi_0\right),$$

$$\phi_1 = \cos(s)\left(\gamma_v \phi_0 + \beta_v \xi_0\right).$$

If we now substitute these expressions into the first two of equations (4) and take the time average, ignoring the time variation of ξ_0 and ϕ_0, we get the averaged equations

$$\ddot{\xi}_0 - \alpha_{eff}\xi_0 + \beta_{eff}\phi_0 = 0,$$

$$\ddot{\phi}_0 + \gamma_{eff} + \beta_{eff}\xi_0 = 0,$$

where

$$\alpha_{eff} = \alpha - \frac{1}{2}\alpha_v^2 - \frac{1}{2}\beta_v^2, \tag{18a}$$

$$\beta_{eff} = \beta - \frac{1}{2}\alpha_v\beta_v + \frac{1}{2}\beta_v\gamma_v, \tag{18b}$$

$$\gamma_{eff} = \gamma + \frac{1}{2}\gamma_v^2 + \frac{1}{2}\beta_v^2. \tag{18c}$$

A necessary and sufficient condition for these averaged equations to have neutrally stable solutions is that the eigenvalues of the matrix

$$K_{eff} = \begin{pmatrix} \alpha_{eff} & -\beta_{eff} \\ -\beta_{eff} & -\gamma_{eff} \end{pmatrix}$$

all be negative. A necessary and sufficient condition for this to be true is that

$$\alpha_{eff} - \gamma_{eff} < 0, \tag{19}$$

$$\alpha_{eff}\gamma_{eff} + \beta_{eff}^2 < 0. \tag{20}$$

These are the necessary and sufficient conditions that the averaged equations yield a neutrally stable equilibrium. One can also obtain these results by carrying out a regular perturbation expansion to find the quantities $Z_k(\Gamma)$ (defined in (10)) assuming that all of the dynamical parameters are small. If we do this, we find that all of the quantities $Z_k(\Gamma)$ are guaranteed to be positive for small values of the dynamical parameters, except for $Z_2(\Gamma)$ and $Z_4(\Gamma)$. The condition $\alpha_{eff} - \gamma_{eff} < 0$ is related to the condition $Z_4(\Gamma) > 0$. In particular, we have

$$Z_4(\Gamma) = 2\pi^2\left(\alpha_v^2 + 2\beta_v^2 + \gamma_v^2 - 2\alpha_0 + 2\gamma_0\right) + \cdots. \tag{21}$$

The requirement that $\alpha_{eff}\gamma_{eff} + \beta_{eff}^2 < 0$ is related to the condition that $Z_2(\Gamma) > 0$. In particular, we have

$$\begin{aligned}\frac{1}{8\pi^4}Z_2(\Gamma) = {} & -4\beta_0^2 + (\beta_v^2 + 2\gamma_0)(\beta_v^2 - 2\alpha_0) + 2\alpha_v^2\gamma_0 \\ & + 4\alpha_v\beta_0\beta_v + \alpha_v^2\gamma_v^2 - 4\beta_0\beta_v\gamma_v - 2\alpha_0\gamma_v^2 + 2\alpha_v\gamma_v\beta_v^2.\end{aligned} \tag{22}$$

The expression for $Z_2(\Gamma)$ is quite complicated, but it contains a considerable amount of information. If we set $\beta_0 = \gamma_v = \alpha_v = 0$, our asymptotic stability conditions imply that

$$Z_4 = 2\pi^2\left(2\beta_v^2 - 2\alpha_0 + 2\gamma_0\right) > 0$$

and

$$Z_2 = 8\pi^4\left(\beta_v^2 + 2\gamma_0\right)\left(\beta_v^2 - 2\alpha_0\right) > 0.$$

In order for the second of these conditions to be satisfied, we must have $\beta_v^2 > 2\alpha_0$. If this is so, and $\gamma_0 > 0$, then the first condition will also be satisfied. In this case, we get the curve for the lower bound on β_v in Figure 3. Note that our high frequency approximation assumes that all of the parameters are small and does not predict the upper limit for β_v. If $\gamma_0 < 0$, then the second condition does not necessarily imply the first and we end up getting a kink in our curve for the lower limit on stability, as in Figure 4. This shows that our asymptotic results agree with the results we previously derived using intuitive arguments.

Suppose we still set $\alpha_v = \gamma_v = \alpha_0 = 0$, but we now let β_0 be nonzero. In this case the condition that $Z_2 = 0$ can be written as

$$-4\beta_0^2 + (\beta_v^2 + 2\gamma_0)\beta_v^2 = 0. \tag{23}$$

If γ_0 is small, this shows that the lower bound on β_v is proportional to the square root of β_0, a result we observed numerically in Figure 5, but had not yet derived. In general, (23) gives a quadratic equation for β_v. Figure 6 shows a plot of $\beta_v(\beta_0)$ for a fixed value of γ_0. It shows both the exact expression obtained by numerically computing the monodromy matrix and the analytical expression obtained by solving (23). We see that there is excellent agreement between these results until β_v gets to be nearly .5, where the high frequency approximation should be expected to break down. Figure 7 shows a similar plot for $\beta_v(\gamma_0)$ and β_0 fixed.

These examples show that although (22) is quite complicated, it contains a considerable amount of information. Furthermore, if we restrict ourselves to simple slices through parameter space, this equation is not so complicated.

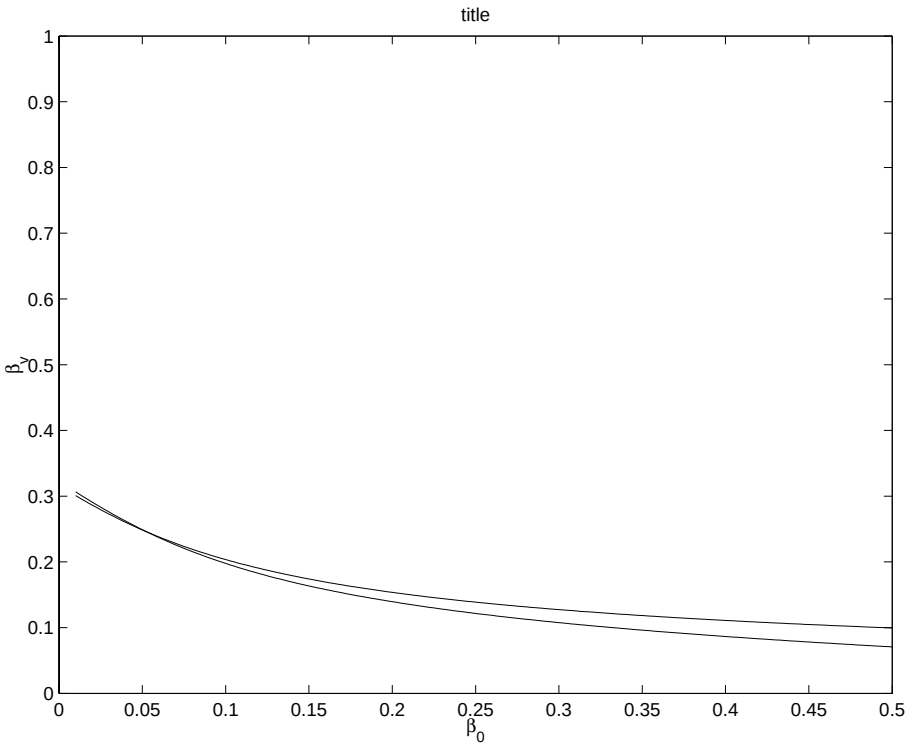


FIG. 6. *This shows the curve for $\beta_v(\beta_0)$ determined by the condition $Z_2 = 0$ ($\alpha_0 = \alpha_v = \gamma_v = 0$, $\gamma_0 = .1$). This gives the lower limit on the value of β_v. The upper curve is found by numerically solving the full set of linearized differential equations. The lower curve is found by using the high frequency approximation. Note that by the time we see significant disagreement between the numerical and asymptotic results, $\beta_v > .5$, which is clearly out of the range of applicability of the high frequency approximation.*

7. Finding real configurations. We will not propose any specific configurations of magnets in this paper. However, we would at least like to explain how one can theoretically find configurations of magnets that yield stable equilibria. We will illustrate the procedure for the case where the only time varying term is B_v and the steady coupling term B_0 is set to zero.

We assume that the top has several systems of axisymmetric multipoles that are, in general, displaced from the center of mass of the top. We assume that each system of magnets on the top can be approximated as a linear combination of dipoles, quadrapoles, and octapoles. It should be pointed out that a ring of finite thickness can be well approximated as a dipole plus an axisymmetric octopole.

We assume that the steady fields are produced by one or more axisymmetric rings of axially magnetized material. These rings in the base can either be approximated as infinitesimally thin rings that have a prescribed magnetization per unit length P_0, or they can be considered as rings of finite thickness that have a prescribed magnetization per unit volume. If we approximate them as being infinitesimally thin, then to actually build such a ring, we prescribe a magnetization per unit volume M_0 and a cross sectional area S for the ring such that SM_0 is equal to P_0.

We assume that the alternating fields are produced by one or more loops of wires that are carrying sinusoidally oscillating currents.

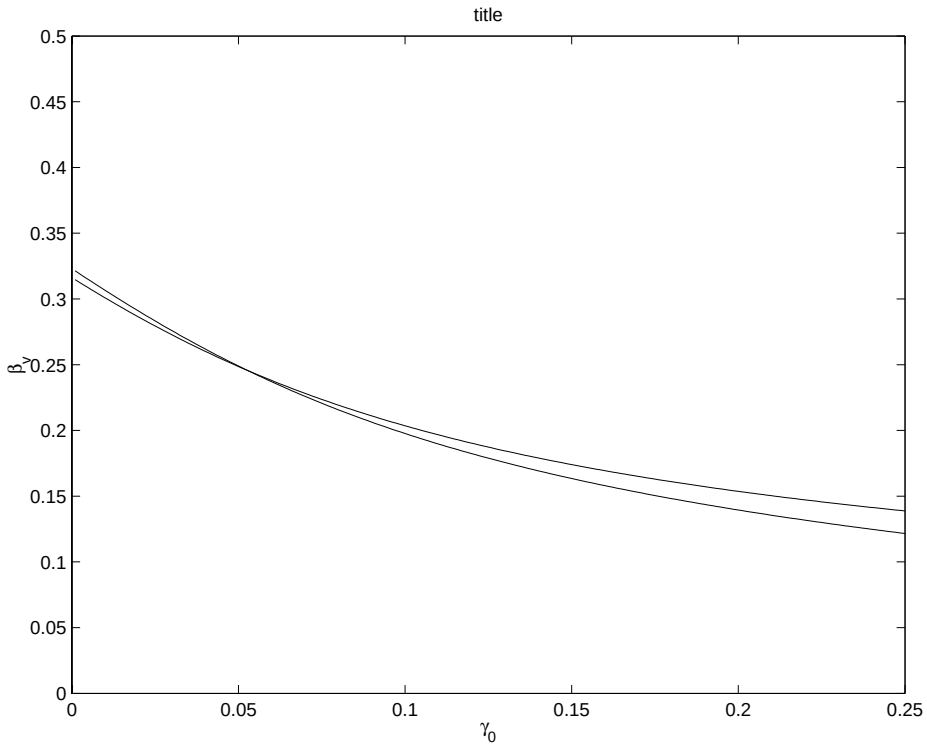


FIG. 7. *This shows the curve for $\beta_v(\gamma_0)$ (for $\alpha_0 = \alpha_v = \gamma_v = 0$ and $\beta_v = .05$) determined by the condition $Z_2 = 0$. This gives the lower limit on the value of β_v. The upper curve (for large values of γ_0) is found by numerically solving the full set of linearized differential equations. The lower curve is found by using the high frequency approximation. Note that the agreement between the numerical and asymptotic results is excellent provided $\gamma_0 < .1$. Beyond this point, the agreement is still quite good, even though we are pushing the limits of the high frequency approximation.*

For a given configuration of magnets it is a straightforward (though tedious) exercise to compute the parameters A, B, C, and L (the lift). We have, in fact, carried out this procedure, but we will just briefly outline it here.

All of the dynamical parameters can be computed if we know the first few terms of the magnetic potential about the equilibrium position of any magnet on the top. In order to find the multidimensional Taylor series, it is only necessary to have the Taylor series on the axis of symmetry. This is because for an axisymmetric field the Taylor series on the axis is sufficient to generate all of the terms in the spherical harmonic expansion of the field. For axisymmetric rings and coils, the Taylor series of the potential on the axis of symmetry is easily computed. This is a very brief outline of how one can compute the dynamical coefficients for quite complicated systems of magnets on the base and the top.

This means that the forward problem of determining whether a given configuration is stable or not can be solved. For a given configuration we can compute both the steady and periodically varying dynamical constants and the lift. If the time varying component of the lift is not zero, or the steady lift term does not balance the force of gravity, then we do not have a good configuration. Assuming these hold, we can compute both the steady and time varying components of the dimensionless

parameters α, β, and γ and use Floquet theory to determine if the configuration is, in fact, stable.

If we could use arbitrarily large values of the magnetization per unit volume M_0, we could easily find configurations of magnets that give stable equilibrium. In order to do this, we could specify the magnets on the top and the positions of four rings in the base. For the given mass and moments of inertia of the top, we determine values of A_0, B_0, C_0, L, and B_v that will give us a stable equilibrium when the center of mass is at the origin. We now determine magnetizations per unit length P_i that give us these values of the dynamical parameters. In order to do this we note that if A_i, B_i, C_i, and L_i are the dynamical parameters coming from the ith ring when it has unit magnetization per unit length, then the total value of A_0 is given by

$$A_0 = \sum P_i A_i.$$

Similar expressions hold for B_0, C_0, and L_0. We see that if we have four rings in the base, we have four linear equations in four unknowns for determining P_i. If the magnetization M_0 is large enough, we are guaranteed of being able to choose a reasonable cross sectional area S_i so that we could actually build the ith ring.

We can carry out a similar procedure for the time varying parameters. Once again, if the current in the coils can be as large as we need, and we have at least four coils, we can find systems of coils such that we get any specified values of the time varying dynamical parameters.

The procedure we have outlined is not quite the one we would want to actually carry out in practice. In practice we would want to wisely position both the magnets on the top and the magnets in the base. Furthermore, we would want to look at several different configurations and choose the one that was the most robust to imperfections in the placement or strengths of the magnets.

8. Conclusions. We believe we have given a comprehensive analysis of the stability of an axisymmetric body in a time periodic magnetic field. We have analyzed the most general configuration possible, taking into account the coupling between the rotational and translational modes of the top. Our analysis applies to any configuration of magnets, as long as they are axisymmetric, and the time varying parts of the field are sinusoidal, and all in phase.

Acknowledgments. The author would like to acknowledge Todd Christenson, Gene Aronson, and Khalid Bou-Rabee for many valuable discussions on this subject.

REFERENCES

[1] A. Andronov, A. Vitt, and S. Khaikin, *Theory of Oscillators*, Dover, New York, 1987.
[2] M. V. Berry, *The levitron: An adiabatic trap for spins*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., 452 (1996), pp. 1207–1220.
[3] H. Dullin and R. Easton, *Stability of levitrons*, Phys. D, 126 (1999), pp. 1–17.
[4] R. F. Gans, T. B. Jones, and M. Washizu, *Dynamics of the levitron*, Phys. D, 31 (1998), pp. 671–679.
[5] G. Genta, C. Delprete, and D. Rondando, *Gyroscopic stabilization of passive magnetic levitation*, Meccanica, 34 (1999), pp. 411–424.
[6] J. Guckenheimer and P. Holmes, *Non-Linear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer-Verlag, New York, 1983.
[7] E. Hones and W. Hones, *Magnetic Levitation and Method*, U.S. patent 5,404,062, 1995.
[8] J. Howard and R. MacKay, *Linear stability of symplectic maps*, J. Math. Phys., 28 (1987), pp. 1036–1051.

[9] A. Mostowski and M. Stark, *An Introduction to Higher Algebra*, Pergamon, New York, 1964.
[10] W. Paul, *Electromagnetic traps for charged and neutral particles*, Rev. Modern Phys., 62 (1990), pp. 531–540.
[11] M. D. Simon, L. O. Heflinger, and S. L. Ridgway, *Spin stabilized magnetic levitation*, Amer. J. Phys., 65 (1997), pp. 286–292.
[12] J. Stoker, *Non-Linear Vibrations*, Wiley-Interscience, New York, 1950.

SIAM J. APPL. MATH.
Vol. 63, No. 6, pp. 2176–2194

SPIN STABILIZED MAGNETIC LEVITATION OF HORIZONTAL ROTORS*

L. A. ROMERO†

Abstract. In this paper we present an analysis of a new configuration for achieving spin stabilized magnetic levitation. In the classical configuration, the rotor spins about a vertical axis; the spin stabilizes the lateral instability of the top in the magnetic field. In this new configuration the rotor spins about a horizontal axis; the spin stabilizes the axial instability of the top in the magnetic field.

Key words. Levitron, stability, magnets

AMS subject classifications. 70E05, 70J25, 78A30

DOI. 10.1137/S0036139902406899

1. Introduction. Earnshaw's theorem [9] implies that it is impossible to achieve stable static magnetic levitation in a static magnetic field. However, the discovery of the Levitron™ [7] has shown that it is in fact possible for a spinning top to be in stable equilibrium in a static magnetic field. We refer to this as spin stabilized magnetic levitation. There have been numerous papers analyzing spin stabilized magnetic levitation [1], [2], [6], [10], [5]. In this paper we extend these results by considering the case of a rotor that spins about a horizontal axis. Although no such device has yet been built, a program is currently under way to build one. A sketch of what such a device might look like is given in Figures 1 and 2. As with the classical Levitron™, we anticipate that there will be a high degree of sensitivity in such a device, so that it may take an adept experimentalist to build one. For this reason we believe that it is worth presenting the theory even though there is as yet no experimental justification.

Classically, spin stabilized magnetic levitation devices are axisymmetric. In principle we could achieve a horizontally spinning device using systems of magnets that have no symmetry properties at all. However, we choose to consider systems that have enough symmetry so that equilibrium of forces and torques is guaranteed in all directions except for the vertical. One such situation (depicted in Figure 1) is the following:

- The base magnets have reflectional symmetry about the planes $y = 0$ and $x = 0$.
- The rotor is axisymmetric and has reflectional symmetry about its midplane.
- The rotor is placed with its center of mass at $(0, 0, z_0)$ and its axis of symmetry pointing in the direction $(1, 0, 0)$.

We will show that due to the symmetry of this configuration, there are no forces in the y and x direction when the rotor is placed symmetrically in the field. Similarly,

*Received by the editors May 2, 2002; accepted for publication (in revised form) February 2, 2003; published electronically September 22, 2003. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under contract DE-AC04-94AL85000.
http://www.siam.org/journals/siap/63-6/40689.html
†Sandia National Laboratories, Applied Numerical Mathematics Division, MS 1110, Albuquerque, NM 87185 (lromero@sandia.gov).

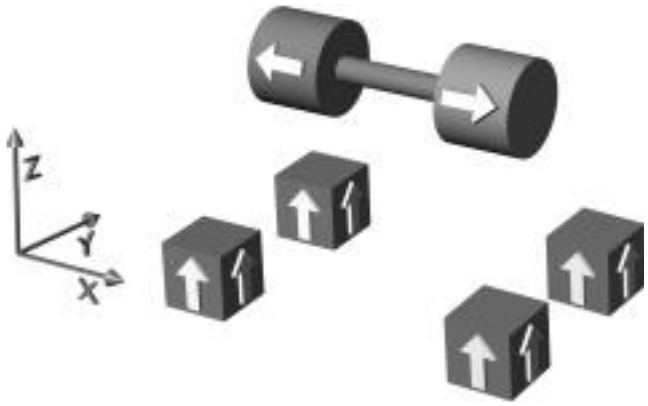


FIG. 1. *This shows a sketch of what a horizontal spin stabilized magnetic levitation device might look like. This particular configuration has an axisymmetric rotor with a system of magnets that has reflectional symmetry about its midplane. The magnets in the base have reflectional symmetry about the planes* $y = 0$ *and* $x = 0$.

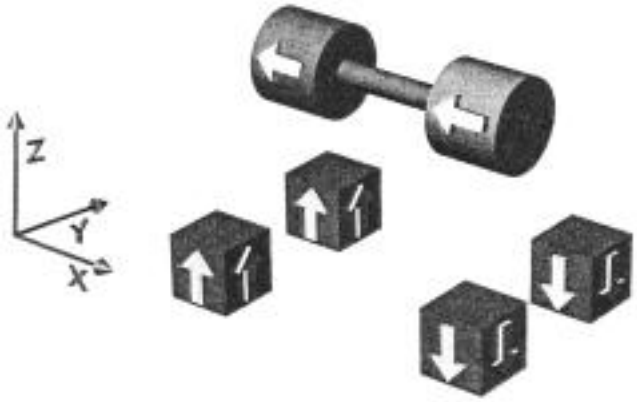


FIG. 2. *This is a schematic of a second type of symmetry for achieving horizontal spin stabilized magnetic levitation. This particular configuration has an axisymmetric rotor with a system of magnets that is antisymmetric about its midplane. The magnets in the base have reflectional symmetry about the plane* $y = 0$ *and are antisymmetric about the plane* $x = 0$.

there are no torques in any direction. Equilibrium in the z direction can be obtained by adjusting the height or weight of the rotor.

A similar situation (depicted in Figure 2) exists when the base magnets are antisymmetric about the plane $x = 0$, and the magnets on the rotor are antisymmetric with respect to reflections about the midplane of the rotor.

Earnshaw's theorem implies that this equilibrium position must be unstable if the rotor is not spinning. When we analyze the stability of a spinning rotor in such a configuration, we find that the equations for perturbations in the y and z directions decouple from the perturbations in the axial (x) direction and from the angular perturbations. This implies that it is not possible for spinning to stabilize the perturbations in the y and z directions. If we are going to stabilize this configuration by spinning the rotor, the rotor must be unstable to perturbations in the axial direction (in the absence of spin). In certain situations we can stabilize the perturbations in

the axial direction by spinning the rotor. As with the vertically spinning systems, there is an upper and lower spin rate for stable equilibrium.

We would like to emphasize that for spin stabilized magnetic levitation of a vertically spinning rotor in an axisymmetric field it is not possible to stabilize the axial direction by spinning. This means that in the absence of spin, the system is stable axially, and unstable laterally. This is exactly the opposite of horizontal spin stabilized systems that we discuss in this paper.

We now give an outline of the rest of this paper. In section 2 we discuss the symmetry properties of these configurations. In section 3 we show that these properties imply that when the rotor is placed symmetrically in the field, all of the forces and torques vanish except for the force in the vertical direction. In section 4 we derive the equations governing the linear stability of the equilibrium. In section 5 we give simple necessary conditions for stability, a simple stability condition similar to the adiabatic approximation made in [1], and a quartic equation that can be solved to determine the upper and lower spin rates. In section 6 we discuss how to compute the dynamical parameters in the linear stability equations for a given configuration of magnets. In section 7 we discuss how to find configurations of magnets that have the desired stability properties. We give our conclusions in section 8.

2. Symmetry properties. We assume that the rotor and its magnets are axisymmetric, that in equilibrium it is aligned with its axis of symmetry in the x direction, and that it spins about the x axis. In equilibrium its center of mass is at $\underline{x} = (0, 0, z_0)$. We consider two different situations:

- Systems where the supporting magnets produce a potential that is antisymmetric with respect to a reflection about the plane $x = 0$ and symmetric with respect to a reflection about the plane $y = 0$. In this case we assume that the magnets on the rotor are antisymmetric with respect to reflections about the midplane.
- Systems where the supporting magnets produce a potential that is symmetric with respect to reflections about the planes $x = 0$ and $y = 0$. In this case we assume that the magnets on the rotor are symmetric with respect to a reflection about the midplane.

We will show that in both of these situations when the center of mass of the rotor is at $y = 0$, $x = 0$ and the axis of symmetry of the rotor is aligned in the x direction, we are guaranteed of having no forces in the y or x direction and no torques on the rotor. By suitably adjusting the weight of the rotor, or the strengths of the magnets, we can make it so that the force in the z direction balances the force of gravity, which we assume points in the z direction.

The first of these symmetries can be constructed by building a rotor with two dipoles on the axis of symmetry, symmetrically located about the midplane, and both pointing in the same direction along the axis of symmetry. In this case a system of supporting magnets having the proper symmetry could consist of magnets in a plane $z =$ constant all pointing in the z direction. In this case any supporting magnet at (x_0, y_0, z_0) would have companion magnets at $(\pm x_0, \pm y_0, z_0)$. The dipole at $(x_0, -y_0, z_0)$ would be in the same direction as the first dipole, and the dipoles at $(-x_0, \pm y_0, z_0)$ would be in the opposite direction. This is just one example of how to achieve this symmetry. More generally we could have the magnets in the base have the dipoles pointing in arbitrary directions as long as their companion magnets have been appropriately reflected.

The second of these symmetries can be constructed by building a rotor with two dipoles on the axis of symmetry, symmetrically placed about the midplane, and pointing in opposite directions along the axis of symmetry. In this case a system of supporting magnets having the proper symmetry could consist of magnets in a plane $z =$ constant all pointing in the z direction. In this case any supporting magnet at (x_0, y_0, z_0) would have companion magnets at $(\pm x_0, \pm y_0, z_0)$. All of the magnets would have their dipoles pointing in the same direction. Once again, this is just one way of achieving systems with this symmetry.

Since the rotor is axisymmetric, the energy of the rotor in an arbitrary magnetic field can be written as

$$Energy = U(\underline{x}, \underline{d}),$$

where $\underline{x} = (x, y, z)$ is the center of mass of the rotor and $\underline{d} = (d_x, d_y, d_z)$ is a unit vector pointing in the direction of the axis of symmetry. The energy satisfies

$$\nabla_x^2 U = 0,$$

where ∇_x^2 is the Laplacian with respect to the variable $\underline{x}$.

The energy of systems where the potential is antisymmetric with respect to reflections about the x axis satisfy the following symmetry properties:

$$U(x, y, z, d_x, d_y, d_z) = U(-x, y, z, d_x, -d_y, -d_z), \tag{1a}$$

$$U(x, y, z, d_x, d_y, d_z) = U(x, -y, z, d_x, -d_y, d_z), \tag{1b}$$

$$U(x, y, z, d_x, d_y, d_z) = -U(x, y, z, -d_x, -d_y, -d_z). \tag{1c}$$

Systems where the potential is symmetric with respect to reflections about the x axis satisfy the identical symmetry properties.

2.1. Examples illustrating the symmetry properties. These symmetry properties become clearer if we consider special cases of such systems. Suppose we have a rotor that has two equal dipoles on the axis of symmetry, each pointing in the direction of the axis of symmetry. We suppose that the magnets are placed symmetrically a distance $\delta/2$ from the center of mass. When the rotor gets displaced and rotated, one of the dipoles will be located at $\underline{x}_+ = \underline{x} + \delta/2\underline{d}$ and the other one at $\underline{x}_- = \underline{x} - \delta/2\underline{d}$. The dipole moment of the magnet at $\underline{x}_+$ will be $\underline{m}_+ = m_0\underline{d}$, and the moment at $\underline{x}_-$ will be $\underline{m}_- = m_0\underline{d}$. The total magnetic energy of the rotor will be

$$U(\underline{x}, \underline{d}) = m_0 \left(\underline{d} \cdot \nabla\phi(\underline{x}_+) + \underline{d} \cdot \nabla\phi(\underline{x}_-) \right).$$

It can be verified that assuming that $\phi(x, y, z)$ is symmetric in y and antisymmetric in x, the energy $U(\underline{x}, \underline{d})$ satisfies the symmetry properties stated in (1). Note that these symmetry properties would hold for more complicated systems, such as rotors having more than one pair of symmetrically placed dipoles or symmetrically placed rings.

An example illustrating the second sort of symmetry comes from a rotor that once again has symmetrically placed dipoles, but in this case the dipoles are equal and opposite to each other. In this case the energy can be written as

$$U(\underline{x},\underline{d}) = m_0\left(\underline{d}\cdot\nabla\phi(\underline{x}_+) - \underline{d}\nabla\phi(\underline{x}_-)\right).$$

Once again it can be verified that if $\phi(x,y,z)$ is symmetric in x and y, then the energy satisfies the symmetry properties (1).

3. Equilibrium. We will now show that assuming our system of magnets and the rotor satisfy the symmetry properties of the last section, we can easily find equilibrium configurations. In particular, we will show that if we place the rotor so that its center of mass is at $(0,0,z_0)$ and its axis of symmetry is pointing in the direction $(1,0,0)$, then there is no torque on the rotor and the only component of force is in the z direction. By appropriately adjusting the weight or the strengths of the magnets, we can make it so that the force of gravity balances this magnetic force.

The force and torque on the rotor can be computed using

$$\underline{F} = -\nabla_x U(\underline{x},\underline{d}),$$

$$\underline{\tau} = -\underline{d}\times\nabla_d U(\underline{x},\underline{d}).$$

Here, ∇_x is the gradient with respect to $\underline{x}$, and ∇_d is the gradient with respect to $\underline{d}$.

We can derive these formulas using generalizations of the derivations for the force and torque on a point dipole [3]. The principle of virtual work tells us that the change in energy when we move the center of mass without rotating it is given by

$$\delta U = -\underline{F}\cdot\delta\underline{r},$$

where $\underline{F}$ is the force on the top and $\delta\underline{r}$ is the change in the center of mass of the top. Since we can write $\delta U = \nabla_x U\cdot\delta\underline{r}$, we see that

$$\nabla_x U\cdot\delta\underline{r} = -\underline{F}\cdot\delta\underline{r}.$$

Since this must hold for all values of $\delta\underline{r}$, we see that

$$\underline{F} = -\nabla_x U.$$

On the other hand, if we rotate the body about the axis $\underline{e}$ by an angle $\delta\theta$, then the principle of virtual work requires that the change in energy is given by

$$\delta U = -\tau\cdot\underline{e}\delta\theta.$$

When we rotate the body about $\underline{e}$ by $\delta\theta$, the change in the unit vector $\underline{d}$ is given by $\delta\underline{d} = \underline{e}\times\underline{d}\delta\theta$. We see that

$$\delta U = \nabla_d U\cdot\delta\underline{d} = \nabla_d U\cdot(\underline{e}\times\underline{d})\delta\theta = (\underline{d}\times\nabla_d U)\cdot\underline{e}\delta\theta.$$

When we equate this expression to the expression from the principle of virtual work, and require that it hold for all values of $\underline{e}$ and θ, we get

$$\underline{\tau} = -\underline{d}\times\nabla_d U.$$

The symmetry properties of the energy show that for both the antisymmetric and symmetric cases we have

$$U(x,0,z,1,0,0) = U(-x,0,z,1,0,0),$$

$$U(0,y,z,1,0,0) = U(0,-y,z,1,0,0).$$

When the rotor is placed symmetrically in the field, the forces F_x and F_y in the x and y directions satisfy

$$F_x(0,0,z,1,0,0) = -\frac{\partial U(0,0,z,1,0,0)}{\partial x} = 0,$$

$$F_y(0,0,z,1,0,0) = -\frac{\partial U(0,0,z,1,0,0)}{\partial y} = 0.$$

To show that the torques vanish, we substitute $x = 0$, $y = 0$ into the symmetry property $U(x,y,z,d_x,d_y,d_z) = U(-x,y,z,d_x,-d_y,-d_z)$ to get

$$U(0,0,z,1,d_y,d_z) = U(0,0,z,1,-d_y,-d_z).$$

This shows that the energy at $x = y = 0$ is an even function of d_y and d_z, and hence the derivatives with respect to d_y and d_z must vanish. Using the fact that $\underline{\tau} = -\underline{d} \times \nabla_d U$ we see that

$$\underline{\tau}(0,0,z,1,0,0) = 0.$$

We see that based on the symmetry of our problem, if we put the rotor so that its center of mass is at $x = y = 0$, so that its axis of symmetry is pointing in the x direction, there will be no forces in the x or y directions and no torques at all.

4. The linearized equations of motion. We describe the kinematics of the rotor in a manner similar to [6]. In our discussion the coordinates (x,y,z) refer to coordinates fixed in space. We assume that the body is axisymmetric with a moment of inertia of I_3 about the axis of symmetry and I_1 about the other two principal axes.

We will orient the body by rotating about the z axis by θ, the y axis by ϕ, and then the x axis by ψ. If the rotor is spinning about the x axis with angular velocity ω_0, then a small perturbation to this state gives approximate angular momenta L_y and L_z of

$$L_y = I_1\dot{\phi} + I_3\omega_0\theta,$$

$$L_z = I_1\dot{\theta} - I_3\omega_0\phi.$$

These formulas can be derived rigorously by expressing the angular momenta in terms of the angular variables and their derivatives and then assuming that θ and ϕ are small. They also have a simple intuitive interpretation. The expression for L_y consists of two terms. The first term is the angular momentum we would get if ω_0 were zero and the body were spinning about the y axis. The second term is the angular momentum we would get if the body kept spinning about the axis of symmetry with angular velocity ω_0 but was slowly tilted by an amount θ about the

x axis. As a result of this tilting some of the angular momentum that was initially in the x direction gets projected onto the y axis. A similar interpretation can be given for the angular momentum in the z direction.

The linearized equations of motion can be written

$$m\ddot{x} = F_x(x, y, z, \theta, \phi),$$

$$m\ddot{y} = F_y(x, y, z, \theta, \phi),$$

$$m\ddot{z} = F_z(x, y, z, \theta, \phi),$$

$$I_1\ddot{\theta} - I_3\omega_0\dot{\phi} = \tau_z(x, y, z, \theta, \phi),$$

$$I_1\ddot{\phi} + I_3\omega_0\dot{\theta} = \tau_y(x, y, z, \theta, \phi).$$

In the linear approximation, the forces and torques are linear functions of (x, y, z, θ, ϕ). In the linear approximation, we have

$$\underline{d} = (d_x, d_y, d_z) = (1, \theta, -\phi).$$

Also, in the linear approximation the forces and torques are derivable from a quadratic potential. The symmetry properties show that many of the terms in the quadratic potential must be missing. For example, the fact that $U(x, y, z, d_x, d_y, d_z) = U(x, -y, z, d_x, -d_y, d_z)$ implies that the Taylor series expansion of the energy cannot have any terms of the form xy, yz, $y\phi$, $x\theta$, $z\theta$, or $\theta\phi$. The fact that $U(x, y, z, d_x, d_y, d_z) = U(-x, y, z, d_x, -d_y, -d_z)$ implies that we cannot have any terms of the form xy, xz, $y\phi$, $y\theta$, $z\phi$, or $z\theta$. Using these symmetry properties we conclude that the linearized equations of motion are of the form

$$m\ddot{y} + A_1 y = 0,$$

$$m\ddot{z} + A_2 z = 0,$$

$$m\ddot{x} - Ax - B\phi = 0,$$

$$I_1\ddot{\theta} - I_3\omega_0\dot{\phi} - C_1\theta = 0,$$

$$I_1\ddot{\phi} + I_3\omega_0\dot{\theta} - C_2\phi - Bx = 0.$$

Note that the equations for y and z decouple from the other equations. This means that in order to have stability we must have A_1 and A_2 both be bigger than zero. In other words, the system would have to be stable to lateral perturbations if the rotor were not spinning. The fact that $\nabla_x^2 U = 0$ (or Earnshaw's theorem) implies that $A_1 + A_2 = A$, and hence the system must be unstable to axial perturbations if the rotor is not spinning.

4.1. The dimensionless equations of motion. We now introduce the dimensionless variables

$$x = \sqrt{I_1/mA}\hat{x},$$

$$t = \sqrt{m/A}\hat{t}.$$

In terms of these dimensionless variables, we get the dimensionless equations (after dropping the hats for notational convenience)

$$\ddot{x} - x - \sqrt{\Lambda}\phi = 0, \tag{2a}$$

$$\ddot{\theta} - \Omega\dot{\phi} - \Gamma_1\theta = 0, \tag{2b}$$

$$\ddot{\phi} + \Omega\dot{\theta} - \Gamma_2\phi - \sqrt{\Lambda}x = 0. \tag{2c}$$

Here we have introduced the dimensionless parameters

$$\Gamma_1 = \frac{mC_1}{I_1 A}, \tag{3}$$

$$\Gamma_2 = \frac{mC_2}{I_1 A}, \tag{4}$$

$$\Lambda = \frac{mB^2}{I_1 A^2}, \tag{5}$$

$$\Omega^2 = \frac{I_3^2\omega_0^2 m}{I_1^2 A}. \tag{6}$$

5. The stability of the equilibrium. We now analyze the stability of the system of equations (2). In the first subsection we compute the characteristic equation governing the stability and give some necessary conditions for stability. In the next subsection we carry out an analysis assuming that Γ_1, Γ_2, and Λ are all large. This analysis gives very simple criteria for stability, and we believe it is similar to making the adiabatic assumption as in [1] (see the discussion in Appendix B). In the next subsection we use results from the theory of polynomials that allow us to predict the exact upper and lower spin rates by solving a quartic equation. This is similar to the procedure carried out in [2] in the analysis of the vertically spinning Levitron™.

5.1. The characteristic equation and its properties. We now assume solutions of the form $e^{i\sigma t}$ in the linearized dynamical equations. This leads to the characteristic polynomial

$$(\sigma^2 + 1)\left((\sigma^2 + \Gamma_1)(\sigma^2 + \Gamma_2) - \Omega^2\sigma^2\right) - \Lambda(\sigma^2 + \Gamma_1) = 0. \tag{7}$$

Expanding this we get

$$G(q,\Omega) = q^3 + q^2\left(1+\Gamma_1+\Gamma_2-\Omega^2\right) + q\left(\Gamma_1+\Gamma_2+\Gamma_1\Gamma_2-\Lambda-\Omega^2\right) + \Gamma_1\Gamma_2 - \Lambda\Gamma_1 = 0, \tag{8}$$

where

$$q = \sigma^2.$$

In order for our system to be stable, all of the roots of (8) must be real and positive. Descarte's theorem [8] implies that for an equation of the form $z^3 + px^2 + qz + r = 0$ to have all real and positive roots, it is necessary that $p < 0$, $q > 0$, and $r < 0$. Furthermore, if all of the roots are real, then these conditions are both necessary and sufficient conditions for all of the roots to be positive. This, along with the condition that $A > 0$ gives us several necessary conditions for stability:

$$\Omega^2 > 1 + \Gamma_1 + \Gamma_2, \tag{9a}$$

$$\Gamma_1 + \Gamma_2 + \Gamma_1\Gamma_2 - \Lambda > \Omega^2, \tag{9b}$$

$$\Lambda\Gamma_1 > \Gamma_1\Gamma_2, \tag{9c}$$

$$\Lambda > 0. \tag{9d}$$

The last of these conditions is the requirement that $A > 0$ in order to have lateral stability. As with the vertically spinning spin stabilized magnetic levitation, we see that there is both an upper and a lower value of Ω for stability.

5.2. Asymptotic stability analysis. We can gain considerable insight into these equations by analyzing their behavior when Γ_1, Γ_2, and Λ are all large. We claim that this is similar to making the adiabatic approximation as in [1]. We elaborate on the connection between our asymptotic stability criterion and the adiabatic approximation in Appendix B.

To be precise, we assume that

$$\Lambda = \lambda/\epsilon^2,$$

$$\Gamma_1 = \gamma_1/\epsilon^2,$$

$$\Gamma_2 = \gamma_2/\epsilon^2,$$

$$\Omega = \omega/\epsilon.$$

If we substitute these expression into (7), multiply by ϵ^4, and set $\epsilon = 0$, we get the equation

$$\sigma^2\gamma_2 = \lambda - \gamma_2.$$

This gives us two roots of our 6th order polynomial. We can have only positive solutions to σ^2 if

$$\gamma_2 > 0$$

and

$$\lambda - \gamma_2 > 0.$$

We get the four other roots by assuming that $\sigma = \hat{\sigma}/\epsilon$. This gives us the equation

$$\hat{\sigma}^2 \left((\hat{\sigma}^2 + \gamma_1)(\hat{\sigma}^2 + \gamma_2) - \Omega^2 \hat{\sigma}^2\right) = 0.$$

After factoring out $\hat{\sigma}^2$ this is the characteristic equation for a spinning rotor in a harmonic potential:

$$(\hat{\sigma}^2 + \gamma_1)(\hat{\sigma}^2 + \gamma_2) - \Omega^2 \hat{\sigma}^2 = 0.$$

A simple application of the quadratic equation shows that in order for this to have all real roots we must have $\Gamma_1 \Gamma_2 > 0$, which along with our previous stability criterion requires that both Γ_1 and Γ_2 be positive. We must also have $\Gamma_1 + \Gamma_2 - \Omega^2 < 0$ and $(\Gamma_1 + \Gamma_2 - \Omega^2)^2 - 4\Gamma_1\Gamma_2 > 0$. By choosing Ω large enough we can satisfy all of these criterion.

We can give a simple interpretation of these stability conditions. If Γ_1, Γ_2, and Λ are large, and the system is not responding too quickly, (2b) implies that

$$\Gamma_2 \phi + \sqrt{\Lambda} x = 0.$$

This is equivalent to saying that as the rotor moves around, it orients itself so that there is no torque on it. This gives us the expression $\phi = -\sqrt{\Lambda} x/\Gamma_2$. When we substitute this into (2a) we get

$$\ddot{x} + x(\Lambda/\Gamma_2 - 1) = 0.$$

We see that this will be a stable harmonic oscillator provided $\Lambda > \Gamma_2$. This is the first of our asymptotic stability conditions. In order to satisfy this condition we must have $\Gamma_2 > 0$, which implies that the rotor would want to flip over in the absence of spin.

Our second criterion is the condition that we are spinning the rotor fast enough that it will not flip over. To analyze this mode we have assumed that σ is of order $1/\epsilon$. In this case, (2a) implies that x is small compared to ϕ. This means that we can solve (2b) and (2c) ignoring x. This is equivalent to considering a rotor spinning in a potential where we ignore the translational energy. This leads to our second stability condition.

The asymptotic analysis we just presented does not predict the existence of an upper spin rate. In order to predict the upper spin rate we once again assume that Γ_1, Γ_2, and Λ are large. We will see that if Ω is too large, the eigenvalues that are of order one will eventually become unstable.

Assuming that σ is order unity and that all of our parameters are large, our eigensystem can be approximated by

$$(\sigma^2 + 1)x + \sqrt{\Lambda}\phi = 0,$$

$$-i\Omega\sigma\phi - \Gamma_1\theta = 0,$$

$$i\sigma\Omega\theta - \Gamma_2\phi - \sqrt{\Lambda}x = 0.$$

These equations are obtained by ignoring the second derivatives of θ and ϕ in (2). They are an extension of the results we have already presented where we ignore all derivatives of these quantities.

These equations imply that

$$(\sigma^2 + 1)\left(\Gamma_1\Gamma_2 - \Omega^2\sigma^2\right) - \Gamma_1\Lambda = 0.$$

This is a quadratic equation for σ^2. We need this equation to have positive real roots. In order for this to be so we must have $\Gamma_1\Gamma_2 > \Omega^2$, $\Lambda > \Gamma_2$, and

$$(z - \Gamma_2)^2 - 4z(\Lambda - \Gamma_2) > 0,$$

where

$$z = \frac{\Omega^2}{\Gamma_1}.$$

This gives a quadratic equation in z whose roots are

$$z_\pm = 2\Lambda - \Gamma_2 \pm \sqrt{(2\Lambda - \Gamma_2)^2 - \Gamma_2^2}. \tag{10}$$

In order to have real roots we must have $z < z_-$ or $z > z_+$. However, if $z > z_+$, we cannot satisfy the other inequalities necessary to have positive real roots. It follows that we must have $\frac{\Omega^2}{\Gamma_1} < z_-$. This is the asymptotic prediction for the upper spin rate. Note that assuming that Γ_1, Γ_2, and Λ are of order $1/\epsilon^2$, this upper limit on the spin rate is also of order $1/\epsilon^2$. On the other hand, the lower spin rate is on the order of $1/\epsilon$. It follows that as we make ϵ smaller, the ratio of the upper and lower spin rate can be made very large.

We will now collect all of our results from the asymptotic stability analysis. Assuming that $\Gamma_1 = \gamma_1/\epsilon$, $\Gamma_2 = \gamma_2/\epsilon$, $\Lambda = \lambda/\epsilon^2$, we see that necessary and sufficient conditions for stability are

$$\Gamma_1 > 0, \tag{11a}$$

$$\Gamma_2 > 0, \tag{11b}$$

$$\Lambda > \Gamma_2, \tag{11c}$$

$$\Omega^2 > \Gamma_1 + \Gamma_2 + 2\sqrt{\Gamma_1\Gamma_2}, \tag{11d}$$

$$\Omega^2 < \Gamma_1 z_-. \tag{11e}$$

Once again we emphasize that if Γ_1, Γ_2, and Λ are order $1/\epsilon^2$, then the lower spin rate is of order $1/\epsilon$ and the upper spin rate is of order $1/\epsilon^2$. This shows that as we keep the ratios of Γ_1, Γ_2, and Λ fixed but let the quantities get large, the ratio of the upper and lower spin rates also gets large.

In the next section we will show that by finding the roots of a fourth order polynomial we can find exact expressions (that must be computed numerically) for the upper and lower spin rates. Figure 3 shows that our asymptotic estimates for the upper and lower spin rates are in fact quite accurate even for moderate values of Γ_1, Γ_2, and Λ.

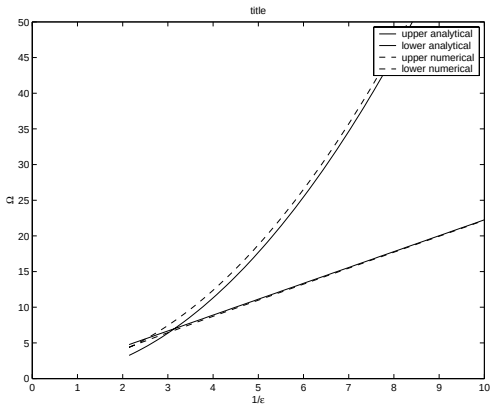


FIG. 3. *This shows the curves for the upper and lower limits on the spin rate* Ω *as a function of* ϵ*, where* $\Gamma_1 = \hat{\Gamma_1}/\epsilon^2$*,* $\Gamma_2 = \hat{\Gamma_2}/\epsilon^2$*, and* $\Lambda = \hat{\Lambda}/\epsilon^2$ *(*$\hat{\Gamma}_1 = 1.$*,* $\hat{\Gamma}_2 = 1.5$*,* $\hat{\Lambda} = 2$*). This figure compares the numerical bounds on the upper and lower spin rate to the asymptotic approximation to these bounds given in* (11).

5.3. Upper and lower bounds on the spin rate. We will now find an exact expression for determining the upper and lower spin rates. In order to do this we first note that in a region of stability we must have $\Lambda\Gamma_1 > \Gamma_1\Gamma_2$. This is both one of our asymptotic stability criteria and one of the conclusions in (9) from Descarte's theorem. This implies that we can never have roots of our characteristic equation $G(q, \Omega) = 0$ (defined in (8)) with $q = 0$. It follows that if Ω_0 is at a boundary of a stability region, then $G(q, \Omega_0)$ must have all real roots, but a small perturbation of Ω will yield complex roots. This implies that on a boundary of a region of stability there must be a root q_0 such that both $G(q_0, \Omega)$ and $G'(q_0, \Omega) = \frac{dG}{dq}$ vanish.

We will write

$$G(q, \Omega) = q^3 + (D - \Omega^2)q^2 + (E - \Omega^2)q + F,$$

where

$$D = 1 + \Gamma_1 + \Gamma_2,$$

$$E = \Gamma_1 + \Gamma_2 + \Gamma_1\Gamma_2 - \Lambda,$$

$$F = \Gamma_1\Gamma_2 - \Lambda\Gamma_1.$$

On the boundary of stability G and G' must have a common root or, equivalently, G must have a multiple root. A necessary and sufficient condition that a polynomial have multiple roots is that the discriminant vanishes. This is equivalent to saying that the resultant of G and G' vanishes. Suppose we have two polynomials

$$g(x) = g_0x^3 + g_1x^2 + g_2x + g_3$$

and

$$h(x) = h_0x^2 + h_1x + h_2.$$

A necessary and sufficient condition that these two polynomials have roots in common is that the resultant vanish. The resultant is the determinant of the following matrix:

$$R = \begin{pmatrix} g_3 & g_2 & g_1 & g_0 & 0 \\ 0 & g_3 & g_2 & g_1 & g_0 \\ h_2 & h_1 & h_0 & 0 & 0 \\ 0 & h_2 & h_1 & h_0 & 0 \\ 0 & 0 & h_2 & h_1 & h_0 \end{pmatrix}.$$

When we substitute the polynomials G and G' into this expression we find (with the help of Mathematica) that the resultant can be written as

$$\psi(\Omega) = \Omega^8 + K_6\Omega^6 + K_4\Omega^4 + K_2\Omega^2 + K_0, \tag{12}$$

$$K_6 = (32 - 16D - 16E + 32F)/8,$$

$$K_4 = (8D^2 - 96E + 32DE + 8E^2 + 144F - 96DF)/8,$$

$$K_2 = (-16D^2E + 96E^2 - 16DE^2 - 144DF + 96D^2F - 144EF)/8,$$

$$K_0 = (8D^2E^2 - 32E^3 - 32D^3F + 144DEF - 216F^2)/8.$$

This is a quartic polynomial in Ω^2. We have shown that on the boundary of stability $\psi(\Omega)$ must vanish, but we have not shown that any root of this equation will yield a value of Ω that is on the boundary of stability. In Appendix A we apply the theory of Hankel matrices [4] to show that the polynomial $G(q, \Omega)$ will have all real roots if and only if $\psi(\Omega) < 0$. We will further show that $G(q, \Omega)$ will have all positive real roots if and only if $\psi(\Omega) < 0$, and all of the inequalities in (9) are satisfied.

If we compute the roots of the polynomial $\psi(\Omega)$, we find that there are roots that do not satisfy the conditions in (9). If we limit ourselves to roots that satisfy the conditions in (9), we find that the roots of $\psi(\Omega)$ do in fact give the upper and lower limits on the spin rates. Figure 3 shows the numerically computed upper and lower spin rates and compares them to the previously derived asymptotic estimates.

6. Computing the dynamical constants. In this section we will explain how, for a given configuration of magnets on the rotor and in the base, one can compute the dynamical constants A_1, A_2, A, B, C_1, C_2, and B that are needed in order to compute the stability of the equilibrium. We also show how to compute the lift L.

For simplicity we will assume that the magnets on the rotor can be approximated by dipoles. We could extend this analysis so that the magnets on the rotor were approximated as a combination of axisymmetric dipoles, quadrapoles, and octopoles. However, that would make some of our results quite tedious. We will begin by analyzing the case where the rotor is in an antisymmetric potential. That is, we assume that the potential $f(x, y, z)$ satisfies

$$f(x, y, z) = f(x, -y, z),$$

$$f(x, y, z) = -f(-x, y, z).$$

We will assume that when the rotor is oriented in its equilibrium position, it has dipoles at $(\pm\delta/2, 0, z_0)$, both of magnitude M_R and both pointing in the direction $(1, 0, 0)$. We will compute the dynamical constants when we have just a single pair of dipoles on the rotor. If we have more than one pair, then the constants can be computed by summing over all the different pairs.

In order to compute the force and torques on the rotor as it gets displaced from its equilibrium, we need to compute the Taylor series (up to the cubic terms) of the magnetic potential about the points $(\pm\delta/2, 0, z_0)$:

$$f(x+\delta/2, y, z_0+z) = \alpha_0 x+\alpha_1 z+\beta_0 xz+\frac{1}{2}\beta_1(2x^2-y^2-z^2)+\frac{1}{2}\beta_2(y^2-z^2)+\Gamma_+(x,y,z),$$

$$\begin{aligned}\Gamma_+(x, y, z_0 + z) = {} & \gamma_0\left(x^3/3 - xy^2/2 - xz^2/2\right) + \gamma_1\left(xy^2/2 - xz^2/2\right) \\ & + \gamma_2(z^3/6 - x^2z/2) + \gamma_3(z^3/6 - y^2z/2) + \cdots.\end{aligned}$$

Around the point $(-\delta/2, 0, z_0)$ we have the Taylor series expansion

$$f(x-\delta/2, y, z_0+z) = \alpha_0 x-\alpha_1 z+\beta_0 xz-\frac{1}{2}\beta_1(2x^2-y^2-z^2)-\frac{1}{2}\beta_2(y^2-z^2)+\Gamma_+(x,y,z),$$

$$\begin{aligned}\Gamma_-(x, y, z_0 + z) = {} & \gamma_0\left(x^3/3 - xy^2/2 - xz^2/2\right) + \gamma_1\left(xy^2/2 - xz^2/2\right) \\ & - \gamma_2(z^3/6 - x^2z/2) - \gamma_3(z^3/6 - y^2z/2) + \cdots.\end{aligned}$$

This is the most general form for the Taylor series (up to cubic terms) of a function $f(x, y, z)$ that is antisymmetric in x and symmetric in y.

The dynamical constants can be computed with the following procedure, which is easily implemented in Mathematica.

- Compute the orientation of the dipole which for small angles is approximated by $\underline{d} = (1 - \theta^2/2 - \phi^2/2, \theta, -\phi)$.
- Set the position of the right dipole to $\underline{x}_+ = \underline{x}_{cm} + \underline{d}\delta/2$ and the dipole moment to $\underline{m}_+ = M_R\underline{d}$.
- Set the position of the left dipole to $\underline{x}_- = \underline{x}_{cm} - \underline{d}\delta/2$ and the dipole moment to $\underline{m}_- = M_R\underline{d}$.
- Compute the magnetic energy $U = \underline{m}_+ \cdot \nabla f(\underline{x}_+) + \underline{m}_- \cdot \nabla f(\underline{x}_-)$.
- Calculate the force $\underline{F} = -\nabla U$.
- Compute the torques, which in the linear approximation can be written as $\tau_z = -\frac{\partial U}{\partial \theta}$ and $\tau_y = -\frac{\partial U}{\partial \phi}$.
- Set $\underline{x}_{cm} = \epsilon(\hat{x}, \hat{y}, \hat{z})$, $\theta = \epsilon\hat{\theta}$, and $\phi = \epsilon\hat{\phi}$.
- Expand the forces and torques up to order ϵ.
- Set the lift L equal to the zeroth order term in the force F_z.
- Set $-A_1$ to the term in F_y that is linearly proportional to y, $-A_2$ to the term in F_z that is linearly proportional to z, and A equal to the term in F_x that is linearly proportional to x. Set B equal to the term in F_x that is linearly proportional to ϕ.
- Set C_1 and C_2 to the terms in τ_z and τ_y that are linearly proportional to θ and ϕ, respectively.

After carrying out this procedure, we arrive at the following expressions for the dynamical constants:

$$L = -2m_0\beta_0, \tag{13}$$

$$A_1 = 2m_0(\gamma_1 - \gamma_0), \tag{14}$$

$$A_2 = -2m_0(\gamma_0 + \gamma_1), \tag{15}$$

$$A = -4m_0\gamma_0, \tag{16}$$

$$B = m_0(2\beta_0 - \gamma_2 d), \tag{17}$$

$$C_1 = 2m_0\alpha_0 + m_0 d\,(4\beta_1 - 2\beta_2) + m_0 d^2(\gamma_0 - \gamma_1)/2, \tag{18}$$

$$C_2 = 2m_0\alpha_0 + m_0 d\,(4\beta_1 + 2\beta_2) + m_0 d^2(\gamma_1 + \gamma_0)/2. \tag{19}$$

If we have several systems of dipoles on the rotor, the dynamical constants are the sum of the dynamical constants for each system of magnets.

It should be pointed out that we get the exact same formula for systems with potentials that are symmetric with respect to reflections about the x axis and whose rotor magnets are also symmetric with respect to reflections about the midplane. In this case we get the same expansion of the field about the point $(\delta/2, 0, z_0)$, but the expansion about $(-\delta/2, 0, 0)$ is exactly opposite that given for the antisymmetric case. If we define our fields using the Taylor expansion about $(\delta/2, 0, 0)$, the dynamical constants have the exact same values as those given for the antisymmetric case.

7. Finding realizable configurations. So far we have discussed how to compute the dynamical constants assuming that we have a given configuration of magnets. We now discuss how one could in fact find a given configuration of magnets that gives the desired dynamical constants. We will present at least one way of going about this for systems that have potentials with reflectional symmetry about the x axis.

We will suppose that the base magnets consist of $4N$ dipoles all pointing in the z direction. The positions of the dipoles are given by

$$\underline{p}_i = (\pm a_i, \pm b_i, c_i), \qquad i = 1, N,$$

and the magnetizations are given by

$$M_i = (0, 0, d_i), \qquad i = 1, N.$$

For each value of i (four symmetrically placed magnets in the base), we can compute the dynamical constants $A_1(i)$, $A_2(i)$, $A(i)$, $B(i)$, $C_1(i)$, $C_2(i)$, and $L(i)$ for $d_i = 1$. The values of the dynamical parameters for the whole system can be obtained by summing over the different sets of magnets multiplied by the strengths of the dipoles. For example,

$$L = \sum_{i=1}^{N} d_i L(i).$$

If we have 6 or more systems of magnets, we can choose the strengths d_i so that we get any desired values of the parameters that we want. This means that in theory we can specify the desired values of A_1, A_2, L, Γ_1, Γ_2, and Λ that we would like, and

thus the values of A, B, C_1, and C_2 that we would like. Once these are known we can determine the dipole strengths of the magnets that give these parameters.

The procedure we have outlined is meant to show that these configurations can be realized in theory. It does not address how to actually find a good configuration. For example, it is possible that the configurations could be very sensitive to small variations in the positions of the magnets or to their strengths. We have carried out some more elaborate forms of this procedure in order to find possible configurations. We do not feel that it is appropriate to give any specific examples until we have analyzed them for their robustness.

8. Conclusions. We have theoretically demonstrated the existence of what is a distinctly different form of spin stabilized magnetic levitation. As with the traditional set up for spin stabilized magnetic levitation, we expect that most configurations will have a high degree of sensitivity to the placement of the magnets. For this reason we believe that it is necessary to come up with some measure of the robustness of a configuration, and to search over a large class of configurations trying to find robust configurations.

Although nobody has ever used spin stabilized magnetic levitation for anything other than a scientific toy, it is possible that this principle could in fact have practical applications. It is hoped that by showing that the classical vertical configuration is not the only possibility, this paper may contribute to the eventual practical use of this principle.

Appendix A. We have shown that on the boundary of a region of stability, we must have $\psi(\Omega) = 0$. In this appendix we will show that the condition $\psi < 0$ is a necessary and sufficient condition for $G(q, \Omega)$ to have all real roots. To do this we apply the method of Hankel matrices presented in [4]. In [4], this method is explained for arbitrary polynomials; to simplify the notation, we will limit ourselves to cubic polynomials. Suppose we have a cubic polynomial of the form

$$a_0 x^3 - a_1 x^2 + a_2 x - a_3.$$

The theory we present allows us to determine the number of real roots of this polynomial.

Suppose (x_0, x_1, x_2) are the roots to this polynomial (which of course we do not know). We begin by computing the Newton polynomials

$$\sigma_k = x_0^k + x_1^k + x_2^k.$$

Even though we do not know the roots to the polynomial, we can compute the Newton polynomials. This follows from the fact that the σ_k's are symmetric polynomials in the variables x_i and hence can be written as polynomials in the coefficients a_j of our polynomial. The theory of how to do this is explained in [8]. We will need to know σ_k up to $k = 4$. We can compute these recursively using

$$\sigma_0 = 3,$$

$$\sigma_1 = a_1,$$

$$\sigma_2 = a_1 \sigma_1 - 2a_2,$$

$$\sigma_3 = a_1\sigma_2 - a_2\sigma_1 + 3a_3,$$

$$\sigma_4 = a_1\sigma_3 - a_2\sigma_2 + a_3\sigma_1.$$

We now form the Hankel matrix:

$$H = \begin{pmatrix} \sigma_0 & \sigma_1 & \sigma_2 \\ \sigma_1 & \sigma_2 & \sigma_3 \\ \sigma_2 & \sigma_3 & \sigma_4 \end{pmatrix}.$$

The number of real roots is equal to $3-2V$, where V is the number of sign changes in the sequence D_0, D_1, D_2, where $D_0 = \sigma_0$,

$$D_1 = \det \begin{pmatrix} \sigma_0 & \sigma_1 \\ \sigma_1 & \sigma_2 \end{pmatrix},$$

and

$$D_2 = \det(H).$$

In order to have all real roots all of the determinants D_0, D_1, and D_2 must be positive. However, for a cubic polynomial, it is not possible to have D_1 be negative while D_2 is positive. This can be shown algebraically, or by noting that if this were the case, then our formula for the number of real roots would yield a negative number of real roots, which is impossible. It follows that a necessary and sufficient condition for our cubic polynomial to have all real roots is that the determinant D_2 is positive.

When we substitute the coefficients from the polynomial $G(q, \Omega)$ into the general expression for D_2, this yields the polynomial $-\psi(\Omega)$. It follows that a necessary and sufficient condition for $G(q, \Omega)$ to have all real roots is that $\psi(\Omega) < 0$.

If a polynomial has all real roots, then a necessary and sufficient condition that all of its roots are positive is that its coefficients alternate in sign. This implies that our system will be stable if and only if both $\psi(\Omega) < 0$ and all of the inequalities in (9) are satisfied.

Appendix B. In this appendix we will discuss the relation between the asymptotic stability analysis made in section 5.2 (assuming Γ_1, Γ_2, and Ω are large) and the adiabatic approximation presented for the vertically spinning Levitron™ in [1]. We will show that these two approaches give the same results, and we will show that the conditions that the dimensionless parameters Γ_1, Γ_2, and Ω be large are equivalent to the conditions stated in [1] for the adiabatic approximation to hold.

Since the adiabatic approximation in [1] is worked out for a point dipole, we will now restrict our analysis to that case. That is, we will assume that our rotor only has a single dipole pointing in the direction of the axis of symmetry. This is an example of one of our two symmetries that we discussed in section 2.

We begin by applying the adiabatic approximation to our problem. Following [1], we argue that assuming that the top is fast we can make the approximation

$$\underline{L} = I_3\omega_0\underline{d}.$$

Here ω_0 is the initial spin of the top, and $\underline{d}$ is the unit vector in the direction of the axis of symmetry. This assumes that we can ignore all components of the angular momentum except for the component about the axis of symmetry of the top. As

pointed out in [1] the fast top approximation holds as long as the spin of the top is large compared to the precession rate of the top.

Under this fast top approximation, the equation for the change in angular momentum can be written as

$$\omega_0 I_3 \dot{\underline{d}} = -m_0 \underline{d} \times \underline{B}.$$

Here m_0 is the dipole moment of the dipole on the rotor. In the adiabatic approximation this equation implies that the quantity

$$\mu_{ad} = \underline{d} \cdot \underline{B} / \mid \underline{B} \mid$$

stays constant. In order for this approximation to hold it is necessary that the rate of change of the vector $\underline{d}$ be large compared to the rate of change of the quantity $\underline{d} \cdot \underline{B} / \mid \underline{B} \mid$.

In the adiabatic approximation, the magnetic energy of the rotor can be written as

$$U_{mag} = -\mu_{ad} \mid \underline{B} \mid .$$

This is equivalent to saying that the top is moving in an effective potential that depends only on the center of mass of the top, not on its orientation. This effective potential is computed by using the magnetic energy $U(\underline{x}, \underline{d}) = -m_0 \underline{d} \cdot \underline{B}$ of the top but using the fact that $\underline{d}$ is always pointing in the direction of the magnetic field. This is clearly equivalent to the approximation made in section 5.2 where we assumed that the rotor always orients itself so that there is no torque on it and then used this to get an effective simple harmonic oscillator for the x component.

We would now like to show that the criteria that our parameters Γ_1, Γ_2, and Ω be large are equivalent to the criteria given in [1] for the adiabatic approximation to hold. We will discuss the scaling properties using the dimensionless linearized equations of motion 2. The precession frequency of the top is given by

$$\Omega_{prec} = \sqrt{\frac{\Gamma_1 \Gamma_2}{\Omega^2}}. \tag{20}$$

This precession frequency is obtained by ignoring the second derivatives of θ and ϕ and the term $\sqrt{\Lambda}$ in (2). The fast top assumption assumes that this precession rate is small compared to the spin rate of the top. This can be written as

$$\Omega \gg \Omega_{prec}. \tag{21}$$

Another condition stated in [1] for the adiabatic approximation to hold is that the bobbing frequency of the top be much less than the precession rate of the top. Physically this means that as the top moves around it can quickly orient itself so that it is aligned with the direction of the magnetic field. In our case the bobbing frequency of the top is obtained by ignoring the term $\sqrt{\Lambda}\phi$ in (2). Since we have made our equations dimensionless by this bobbing frequency, our bobbing frequency is unity. The condition that the precession rate is fast compared to the bobbing frequency can be written as

$$\Omega_{prec} \gg 1. \tag{22}$$

In order to satisfy both of the conditions in (21) and (22), it is clearly necessary that $\Omega \gg 1$. The condition that $\Omega_{prec} \gg 1$ implies that

$$\Gamma_1 \Gamma_2 \gg \Omega^4.$$

Since Ω is large, this implies that the product $\Gamma_1 \Gamma_2$ must be large. In the case of an axisymmetric top considered in [1], this would imply that $\Gamma_1 = \Gamma_2 \gg 1$.

REFERENCES

[1] M. V. Berry, *The levitron: An adiabatic trap for spins*, Proc. Roy. Soc. London Ser. A, 452 (1996), pp. 1207–1220.
[2] H. Dullin and R. Easton, *Stability of levitrons*, Phys. D, 126 (1999), pp. 1–17.
[3] R. Feynman, R. Leighton, and M. Sands, *The Feynman Lectures on Physics*, Vol. II, Addison–Wesley, Reading, MA, 1964.
[4] F. P. Gantmacher, *Matrix Theory*, Chelsea, New York, 1959.
[5] R. F. Gans, T. B. Jones, and M. Washizu, *Dynamics of the levitron*, Phys. D, 31 (1998), pp. 671–679.
[6] G. Genta, C. Delprete, and D. Rondano, *Gyroscopic stabilization of passive magnetic levitation*, Meccanica, 34 (1999), pp. 411–424.
[7] E. Hones and W. Hones, *Magnetic Levitation and Method*, U.S. Patent 5, 404, 062, USPTO, Washington, DC, 1995.
[8] A. Mostowski and M. Stark, *An Introduction to Higher Algebra*, Pergamon Press, New York, 1964.
[9] L. Page and N. Adams, *Principles of Electricity*, Van Nostrand, New York, 1968.
[10] M. D. Simon, L. O. Heflinger, and S. L. Ridgway, *Spin stabilized magnetic levitation*, Amer. J. Phys., 65 (1997), pp. 286–292.